Waiting period from diagnosis for mortgage insurance issued to cancer survivors

Antoine Soetewey

Thesis submitted in partial fulfillment of the requirements for the Degree of Doctor of Sciences

March, 2024

Institute of Statistics, Biostatistics and Actuarial Sciences (ISBA)
Louvain Institute of Data Analysis and Modeling in Economics and Statistics (LIDAM)
UCLouvain, Louvain-la-Neuve, Belgium

Thesis Committee:

Pr. Christian **Hafner**	Chairman	UCLouvain/ISBA/LIDAM, Belgium
Pr. Catherine **Legrand**	Advisor	UCLouvain/ISBA/LIDAM, Belgium
Pr. Michel **Denuit**	Advisor	UCLouvain/ISBA/LIDAM, Belgium
Dr. Geert **Silversmit**	Advisor	Belgian Cancer Registry, Belgium
Pr. Donatien **Hainaut**	Member	UCLouvain/ISBA/LIDAM, Belgium
Pr. Roch **Giorgi**	Member	Aix-Marseille Université, France
Pr. Philippe-Jean **Bousquet**	Member	Collecteur Analyseur de Données, France

Waiting period from diagnosis for mortgage insurance issued to cancer survivors

by Antoine Soetewey

ISBA/LIDAM
UCLouvain
Voie du Roman Pays, 20
1348 Louvain-la-Neuve
Belgium

"You can't connect the dots looking forward; you can only connect them looking backwards. So you have to trust that the dots will somehow connect in your future."
(STEVE JOBS)

Acknowledgments

This thesis is first and foremost the result of a number of encounters and collaborations. I owe this PhD thesis to the help and support of many kind people around me, to only some of whom it is possible to give particular mention here.

Above all, this thesis would not have been possible without the support of my principal supervisors, Pr. Catherine Legrand and Pr. Michel Denuit from UCLouvain. I thank them for their trust, their patience, their rigor, and for giving me the opportunity to pursue a PhD on this very topical subject. Furthermore, I thank them for their time and availability during the past 7 years. There are many great PhD thesis supervisors at UCLouvain, but I believe that the pair Pr. Catherine Legrand and Pr. Michel Denuit is probably the best pair a researcher could hope for. I am confident that my successors are in great hands with them, and that their synergy will continue to be of exceptional help for anyone under their supervision.

I express my gratitude to my co-supervisor, Dr. Geert Silversmit from the Belgian Cancer Registry for his research assistance. Thanks to his thoughtful and valuable comments, he considerably improved the quality of our papers and this thesis. I also thank the staff of the Belgian Cancer Registry and all physicians, pathologists and data managers involved in Cancer Registration in Belgium for their dedicated data collection. I hope that the data will continue to be shared with PhD students and researchers so that cancer research can continue to evolve.

I would like to thank Pr. Donatien Hainaut, Pr. Roch Giorgi, Pr. Philippe-Jean Bousquet and Pr. Christian Hafner for accepting to be part of my doctoral committee. I also thank them in advance for their feedback and helpful suggestions which will undoubtedly improve the quality of the final version of this thesis.

I owe thanks to the anonymous referees of the papers for their comments and questions that have been very helpful in revising previous versions of our papers and the present work.

I am also grateful for the financial support from UCLouvain and funding from the FWO and F.R.S.-FNRS under the Excellence of Science (EOS) program, project EOS 40007517.

I am grateful to my colleagues and to all the members of the ISBA and SMCS I had the chance to work with; teaching assistants and researchers (Alexandre Jacquemain, Aurèle Bartolomeo, Benjamin Deketelaere, Charles-Guy Njike Leunga, Charlotte Jamotton, Edouard Motte, Emmanuel Niyigena, Fanny Hoogstoel, Florian Pechon, Gabriel Bailly, Gilles Mordant, Hélène Morsomme, Hortense Doms, Hugo Brunet, Hugues Annoye, Jean-Loup Dupret, John-John Ketelbuters, Lara Wautier, Lise Léonard, Madeline Vast, Manon Martin, Mathilde Foulon, Mickaël De Backer, Morine Delhelle, Nathalie Lucas, Nathan Uyttendaele (with whom I hope to continue collaborating in the future), Oswaldo Gressani (with whom I hope to continue seeing in conferences), Oussama Belhouari, Patricia Ortega Jiménez, Quentin Le Coënt, Rebecca Marion, Rémi Gengler, Sophie Mathieu, Stefka Kirilova Asenova and Stéphane Lhaut), professors (Pr. Anouar El Ghouch, Pr. Bernadette Govaerts, Pr. Cédric Heuchenne, Pr. Christian Hafner, Pr. Eugen Pircalabelu, Pr. Ingrid Van Keilegom, Pr. Johan Segers, Pr. Laura Symul, Pr. Marie-Paul Kestemont, Pr. Philippe Lambert and Pr. Rainer von Sachs),

statistical consultants (Alain Guillet, Aurélie Bertrand, Benjamin Colling, Catherine Rasse, Céline Bugli, Christian Ritter, Nathalie Lefèvre, Lieven Desmet and Séverine Guisset) and the administrative team (Marguerite-Marie Hanon, Nadja Peiffer, Nancy Guillaume, Sophie Malali and Tatiana Regout).

I would like to express my special thanks to Pr. Dominique Deprins for the 6 years we have worked together on the LESPO2102 course at UCLouvain; thanks to the confidence she has shown in me, I have been able to develop my teaching skills tremendously.

I thank Pr. Anna Kiriliouk for giving me the opportunity to teach the course "Advanced Quantitative Methods" at UNamur as visiting lecturer during the last semester of my PhD journey. A special mention also to Pr. Niko Speybroeck for our collaborations during the COVID-19 and for giving me the chance to be co-author for one of his paper. I would like to thank my PhD supervisors, Pr. Catherine Legrand and Pr. Michel Denuit, once again for agreeing to let me take part in these two projects, which were in addition to my thesis and my teaching tasks at UCLouvain. Despite the fact that our research already took a large part of my time, they trusted me and I cannot thank them enough for these additional opportunities to flourish as a researcher and teacher. I also thank them for their invaluable advice and guidance regarding my post-doctorate career choices.

I thank my friend Vincent Bremhorst for informing me about the position that was opening up at UCLouvain beginning in September 2017. I probably would not have started a PhD if he had not told me about this position. He has been a turning point in my career.

All persons mentioned here helped in providing a very enriching, challenging and stimulating work environment. I would not have learned as much without them. I am sure our paths will cross again in the near future. I would also like to thank all the students and professionals I met. Thanks to them, I discovered a passion for teaching statistics and R which will undoubtedly be useful for my professional career.

I thank my family, in particular my parents; Sabine Fontaine and Christian Soetewey, their partner; Philippe Deboutez and Sophie Norro, my parents-in-law; Isabelle Lecocq and Daniel Vancaster, my godparents; Bénédicte Fontaine and Dario Pinchetti, my siblings; François, Caroline and Arthur, and my cousins, aunts and uncles (who are too numerous to be all mentioned here). All this, and the rest, would not have been possible without them.

I also thank my friends; Cédric Hamiet, Cédric Vranckx, Christophe Metten, Fanny Tacheny, Floriane Dierckx, Frédéric Metten, Frédéric Szikora, Dorian Delobbe, Guillaume Woelfle, Maximilien de le Hoye, Quentin Vanderlinden and Stéphan Henry for their entertainment and for helping me to stay healthy (both physically and mentally).

Amongst my fellow undergraduate colleagues at UCLouvain, I thank Bastien Castiaux, Cyprien Georges, Dylan Déom and Matthieu Depreter whose wisdom helped to shape this thesis in many large and small ways.

I would like to express my special thanks to the members of my family and my friends who agreed to read this thesis and share their feedback. Another special thank you goes to Quentin Vanderlinden for his detailed and constructive comments on some mathematical notations.

After thanking all people who have contributed in one way or another to my research journey at UCLouvain and who supported me over the past years, I express my deep gratitude to my wife, Dr. Elsa Vancaster. I thank her for her daily support since the very first day and for giving me the inspiration and the motivation to always give my best. I also thank her for constantly putting me back on the right track when I forget what is important or when I come up with new ideas (most of which are not so good).

I hope to return all these favors someday.

For any errors or inadequacies that may remain in this work, of course, the responsibility is entirely my own.

Antoine Soetewey
UCLouvain, Belgium
March, 2024

To Elsa.

Contents

Acknowledgments **i**

Bibliographic notes **xi**

List of acronyms **xiii**

List of figures **xv**

List of tables **xvii**

List of symbols **xix**

1 Introduction **1**
1.1 The right to be forgotten . . . 2
1.2 Summary of the chapters . . . 5
1.3 Prerequisites . . . 7
1.3.1 Survival analysis . . . 7
1.3.2 Censoring and truncation . . . 8
1.3.3 Common functions in survival analysis . . . 9
1.3.3.1 Survival function . . . 9
1.3.3.2 Hazard function . . . 10
1.3.3.3 Cumulative hazard function . . . 11
1.3.4 Estimation of the survival function . . . 11
1.3.5 Relative and net survival . . . 12
1.3.6 Cure models and time-to-cure . . . 13
1.3.7 Number of years of life lost . . . 15
1.3.8 Multi-state models . . . 16
1.3.9 Expected present value . . . 17
1.3.9.1 Common life and health insurance products . . . 20

2 Waiting period from diagnosis for mortgage insurance issued to cancer survivors **23**
2.1 Introduction . . . 23
2.2 Data sources . . . 26
2.2.1 Belgian Cancer Registry (BCR) . . . 26
2.2.2 General population . . . 26
2.3 Survival of cancer patients . . . 27
2.3.1 Overall survival . . . 27
2.3.2 Relative survival . . . 29
2.3.3 Proportional excess hazards . . . 32
2.3.4 Flexible parametric model . . . 33
2.3.5 Cure models . . . 35
2.4 Time-to-cure . . . 36

2.5 Application to mortgage insurance . . . 37
2.6 Discussion . . . 40

3 Semi-Markov modeling for cancer insurance 43
3.1 Introduction and motivation . . . 43
3.2 Semi-Markov 3-state model . . . 46
3.2.1 State space and transitions . . . 46
3.2.2 Transition intensities . . . 46
3.2.3 Data . . . 47
3.2.3.1 Belgian Cancer Registry (BCR) . . . 47
3.2.3.2 General population . . . 47
3.2.4 Estimation . . . 47
3.2.5 Transition probabilities . . . 51
3.3 Cancer insurance products . . . 51
3.3.1 Notation and specific policy conditions . . . 51
3.3.2 Stand-alone covers . . . 54
3.3.2.1 Lump sum . . . 54
3.3.2.2 Temporary life annuities . . . 54
3.3.3 Combined products . . . 57
3.3.3.1 Premium exemption . . . 57
3.3.3.2 Term-life insurance with cancer acceleration benefit 57
3.3.4 Cover option . . . 60
3.4 Discussion . . . 61

4 Health indices for disease incidence risk and duration in the Semi-Markov setting 65
4.1 Introduction and motivation . . . 66
4.2 Data . . . 68
4.3 MSM and YLL for cancer patients . . . 70
4.4 Derived health indices - case of three specific cancer types . . . 75
4.4.1 Incidence risk . . . 76
4.4.2 Years of life lost from diagnosis . . . 76
4.5 Discussion . . . 78
4.6 Additional notes . . . 80

5 Right to be forgotten for mortgage insurance issued to cancer survivors: Critical assessment and new proposal 83
5.1 Introduction and motivation . . . 83
5.2 Mortage insurance . . . 85
5.3 Data . . . 87
5.4 Critical assessment . . . 88
5.4.1 Impact of extrapolation in case of limited follow-up . . . 88
5.4.2 Conditional relative net survival . . . 90
5.5 Proposed approach for limited follow-up . . . 92
5.6 Impact of the stage of the tumor . . . 98
5.7 Discussion . . . 101

6 Conclusion 103
6.1 General discussion . . . 103
6.2 Future research . . . 106

A Appendix **111**
A.1 Development of $e_{11}^{\tau}(t;z)$ 111
A.2 Conditional one-year observed survival probability 112
A.2.1 Observed survival and hazard rate 112
A.2.2 Flexible parametric model for the hazard rate 112
A.2.3 Predicted observed survival 112
A.2.4 Predicted conditional one-year observed survival 113

Bibliographic notes

Journal publications

- A. Soetewey, C. Legrand, M. Denuit, and G. Silversmit. Waiting period from diagnosis for mortgage insurance issued to cancer survivors. *European Actuarial Journal*, 11(1):135–160, 2021. https://doi.org/10.1007/s13385-020-00254-x
- A. Soetewey, C. Legrand, M. Denuit, and G. Silversmit. Semi-markov modeling for cancer insurance. *European Actuarial Journal*, 12(2):813–837, 2022. https://doi.org/10.1007/s13385-022-00308-2
- H. C. Truong, T. Van Phan, H. T. Nguyen, K. H. Truong, V. C. Do, N. N. M. Pham, T. V. Ho, T. T. Q. Phan, T. A. Hoang, A. Soetewey, T. N. L. Ho, Q. D. Pham, Q. C. Luong, D. T. T. Vo, T. V. Nguyen, and N. Speybroeck. Childhood Bacterial Meningitis Surveillance in Southern Vietnam: Trends and Vaccination Implications From 2012 to 2021. *Open Forum Infectious Diseases*, 10(7):ofad229, 2023. https://doi.org/10.1093/ofid/ofad229
- A. Soetewey, C. Legrand, M. Denuit, and G. Silversmit. Right to be forgotten for mortgage insurance issued to cancer survivors: Critical assessment and new proposal. *Under review*, 2023. http://hdl.handle.net/2078.1/281061
- A. Soetewey, C. Legrand, M. Denuit, and G. Silversmit. Health indices for disease incidence risk and duration in the Semi-Markov setting. *Under review*, 2024. https://doi.org/10.21203/rs.3.rs-3953605/v1

Talks & posters

- Joint PhD seminar in statistics, actuarial and financial mathematics (2018). University of Oldenburg, Germany [slides]
- Workshop: “Insurability of people who have had cancer” (2018). UCLouvain, Belgium [slides]
- Young Researchers Day (2018). UCLouvain, Belgium [slides]
- 26th Annual meeting of the Royal Statistical Society of Belgium (RSSB) (2018). Domaine des Hautes Fagnes, Belgium [poster]
- Workshop: “How can we predict the evolution of Covid-19 in Belgium?” (2020). UCLouvain, Belgium [slides]
- Scientific exchange day: “Cure models: estimating the recovery time to improve access to insurance” (2020). French National Cancer Institute (INCa), France [slides]

- 15th Clinical epidemiology conference (EPICLIN) (2021). Marseille, France [poster]
- Young Researchers Day (2021). UCLouvain, Belgium [slides]
- Hackathon of the 28th annual meeting of the Royal Statistical Society of Belgium (RSSB) (2021). ULiège, Belgium [slides]
- Workshop on behalf of the 30th anniversary of the UCLouvain's Institute of Statistics: "Statistics and its role in societal challenges" (2022). UCLouvain, Belgium [poster]
- Hackathon of the 30th annual meeting of the Royal Statistical Society of Belgium (RSSB) (2023). UCLouvain, Belgium [slides]

Miscellaneous contributions

- How can we predict the evolution of COVID 19 in Belgium? [LIDAM communication]
- The right to be forgotten - Are we all equal when it comes to insurance? [LIDAM news]
- Research: a profession, expertise and passion [UCLouvain communication]

List of acronyms

AERAS	S'Assurer et Emprunter avec un Risque Aggravé de Santé
BCR	Belgian Cancer Registry
BH	Baseline hazard
BRCA1/BRCA2	Breast cancer 1/Breast cancer 2
CI	Confidence interval
CII	Critical illness insurance
CR	Cancer Registry
DD	Dread disease
EEA	European Economic Area
EPV	Expected present value
EU	European Union
FRANCIM	French network of cancer registries
FSMA	Financial Services and Markets Authority
HIV	Human Immunodeficiency Virus
ICD-10	International Classification of Diseases, Tenth Revision
IP	Income protection
KCE	Belgian Health Care Knowledge Centre
KM	Kaplan-Meier
LNPH	Linear and non-proportional hazard
LPH	Linear and proportional hazard
LTCI	Long-term care insurance
MSM	Multi-state model
NBB	National Bank of Belgium
NIS	National Institute of Statistics
NISS/INSZ	Numéro d'Identification à la Sécurité Sociale/IdentificatieNummer Sociale Zekerheid
NLNPH	Non-linear and non-proportional hazard
NLPH	Non-linear and proportional hazard
RTBF	Right to be forgotten
SD	Standard deviation
SE	Standard error
TTC	Time-to-cure
UK	United Kingdom
YLL	Years of life lost

List of figures

1.1 Status of the right to be forgotten in the EU/EEA as of February 2024. *Source: European Initiative on Ending Discrimination against Cancer Survivors* . . . 4
1.2 Example of a survival function $S(t)$ . . . 10
1.3 Visual representation of the 'illness-death model' without recovery for cancer patients . . . 17

2.1 Estimated overall survival probability by gender and site using the nonparametric Kaplan-Meier (1958) estimator. . . . 28
2.2 Net survival by gender and site using the nonparametric Perme et al. (2012) estimator. . . . 30
2.3 Net survival by age group and site using the nonparametric Perme et al. (2012) estimator. . . . 31
2.4 Excess hazard by age and cancer site estimated with a non-linear and non-proportional hazard model. . . . 34
2.5 Cure proportion versus time-to-cure (in years, with $\epsilon = 0.05$) by gender and age group at diagnosis in patients diagnosed with melanoma and thyroid cancer. . . . 38
2.6 Expected present value (EPV) of a life insurance contracted by a 30 and 50-year-old cancer patient for a period of 20 years with interest of 1 percent and benefit of 100,000. Horizontal lines correspond to EPV calculated with XK, NBB and Statbel life tables. . . . 39

3.1 Semi-Markov 3-state model for cancer insurance. . . . 46
3.2 Estimated transition intensities μ_y^{ad} as functions of attained age y. General population (Statbel, continuous line) and insurance regulatory life tables XR (broken line) and XK (dotted line). . . . 50
3.3 Estimated transition intensities μ_y^{ai} as functions of attained age y, for different cancer types. . . . 50
3.4 Mortality intensities according to age at diagnosis and sojourn time in the cancer state. . . . 52
3.5 Values of $\overline{A}^{a;a\to i}_{x;n\rceil}$ as function of age $x \in \{20, 21, \ldots, 40\}$ for different cancer types with $n = 20$ and yearly interest rate 1%. . . . 55
3.6 Values of $\overline{a}^{\text{ai}}_{x;n\rceil}$ as function of age x for different cancer types with $n = 20$ and yearly interest rate 1%. . . . 57
3.7 Values of $\overline{A}^{(\alpha)}_{x;n\rceil}$ as function of age x for different cancer types with $n = 20$, yearly interest rate 1% and $\alpha = 50\%$. . . . 59
3.8 Values of premiums $\Pi^{i}_{x+t+k;k}$ and Π^{XK}_{x+t+k} for $x + t = 30$ and 40, and time k since diagnosis in $\{0, 1, \ldots, 10\}$, with $n = 20$, yearly interest rate 1% and the reference outstanding balance cover. . . . 62

3.9 Differences $\Pi^{i}_{x+t+2;2} - \Pi^{XK}_{x+t+2}$ according to age $x + t \in \{20, 21, \ldots, 50\}$ in the upper panel and $\text{EPV}_{\text{wc}}(x, n, 2)$ for $x \in \{20, 21, \ldots, 40\}$, with $n = 20$, yearly interest rate 1% in the lower panel. 63

4.1 Visual representation of the 'illness-death model' without recovery for cancer patients 68
4.2 Representation of MSM to estimate YLL^{i} from diagnosis 74
4.3 Probabilities of being diagnosed with breast, melanoma and thyroid cancer over the next $n = 20$ years for a healthy individual as function of age $t \in \{20, 21, \ldots, 40\}$ 77
4.4 Number of life years lost at the individual level before the age of 70 years due to cancer, estimated from $z = 0, 5$ and 10 years after diagnosis, as a function of age at diagnosis 77

5.1 Expected present value (EPV) of a life insurance contracted by a 30-year-old cancer patient for a period of 20 years with interest of 1 percent and benefit of 100 000. Horizontal dashed lines correspond to EPV calculated according to XK life table. 90
5.2 Expected present value (EPV) of a life insurance contracted by a 50-year-old cancer patient for a period of 20 years with interest of 1 percent and benefit of 100 000. Horizontal dashed lines correspond to EPV calculated according to XK life table. 91
5.3 Survival probabilities (with 95% confidence interval) by cancer site and age at diagnosis. Mexhaz method corresponds to the probabilities obtained via a flexible parametric model (dashed line), whereas KM-based method corresponds to the ones obtained based on the nonparametric Kaplan-Meier estimator (solid line). 94
5.4 Conditional one-year survival probabilities (with 95% confidence interval) by cancer site and age at diagnosis. Mexhaz method corresponds to the probabilities obtained via a flexible parametric model, $p^{\text{CR}}_{x,w}$, whereas KM-based method corresponds to the ones obtained based on the nonparametric Kaplan-Meier estimator, $p^{\text{KM}}_{x,w}$. 95
5.5 Ratio of conditional one-year survival probability (with 95% confidence interval) by cancer site and age at diagnosis, together with the additive correction $\exp(-\gamma)$ (horizontal dashed lines). Mexhaz method corresponds to $p^{\text{CR}}_{x,w}/p^{\text{NIS}}_{y}$, whereas KM-based method corresponds to $p^{\text{KM}}_{x,w}/p^{\text{NIS}}_{y}$. 97
5.6 Ratio of conditional one-year survival probability obtained via the proposed approach (i.e., $p^{\text{CR}}_{x,w}/p^{\text{NIS}}_{y}$) by cancer site, age and stage at diagnosis. Horizontal dashed lines correspond to the additive correction $\exp(-\gamma)$. 100

List of tables

2.1 Numbers of melanoma and thyroid cancer cases diagnosed in Belgium between 2004 and 2016 (BCR data) by gender, site and age group, with percentage of lost to follow-up. 27

2.2 Estimated overall survival probabilities by gender and site using the nonparametric Kaplan-Meier estimator at 5 and 10 years after diagnosis, $\widehat{S}(t = 5)$ and $\widehat{S}(t = 10)$, with their confidence interval (CI) at 95% level. 28

2.3 Net survival probabilities by gender, site and age group using the nonparametric Perme et al. (2012) estimator at 5 and 10 years after diagnosis, $\widehat{S}_n(t = 5)$ and $\widehat{S}_n(t = 10)$, with their confidence intervals at 95% level. 31

2.4 Results of model (2.4) fitted to melanoma cancer data. 33

2.5 Results of model (2.4) fitted to thyroid cancer data. 33

2.6 Estimated cured fractions (in %) and mean survival time (in year) for the fatal cases by gender, site and age group, with their 95% confidence intervals. *Note: Est. mean T = Estimated mean survival time of fatal cases. NAs for young age groups are due to an insufficient number of cases.* 36

2.7 Estimated value of time-to-cure ($\widehat{\mathrm{TTC}}_B$ in years, with $\epsilon = 0.05$ and $\epsilon = 0.01$) together with 95% confidence intervals and cure proportion (in %) by cancer site, sex and age group. 37

3.1 Number of persons diagnosed with melanoma, thyroid and female breast cancer in Belgium between 2004 and 2018 (BCR data) by gender, site and age group, together with the percentage of lost to follow-up and the number of deaths. 48

4.1 Number of persons diagnosed with melanoma, thyroid and female breast cancer in Belgium between 2004 and 2020 (BCR data) by sex, site and age group, together with the percentage of lost to follow-up and the number of deaths. 69

5.1 Number of persons diagnosed with melanoma, thyroid and female breast cancer in Belgium between 2004 and 2020 (BCR data) by sex, site and age group, together with the percentage of lost to follow-up and the number of deaths. 88

5.2 Waiting periods by cancer site and age at diagnosis. The star indicates that EPV does not stay below XK level but start to increase a few years after diagnosis. 91

5.3 Waiting periods by cancer site and age at diagnosis computed via our approach and via the KM-based method. 96

5.4 Number of included cases, number of observed deaths, one-year and 5-year observed survival probabilities (with 95% confidence interval) by cancer site and stage of the tumor. Survival probabilities are obtained with the nonparametric Kaplan-Meier estimator. 99
5.5 Comparison of waiting periods by cancer site and age at diagnosis resulting from our approach and from Soetewey et al. (2021). 100

List of symbols

$c(\cdot)$	Amount of death benefit
$CI_{95\%}$	95% confidence interval
$e(\cdot)$	Remaining life expectancy
E^i	Exposure to risk in state i
f	Deferred period
$f(\cdot)$	Density function
$F(\cdot)$	Cumulative distribution function
$I(\cdot)$	Indicator function, equal to 1 if the argument is true and 0 otherwise
l_x	Number of individuals living at the beginning of age x
m	Maximum payment duration
n_j	Remaining number of individuals at risk for each distinct event time t_j
N^{ij}	Number of transitions from state i to state j
O_j	Number of events observed for each distinct event time t_j
p_x	One-year survival probability at age x
${}_tp_x$	t-year survival probability at age x
p^{ij}	Transition probability from state i to state j
p^{ii}	Sojourn probability in state i
$P(\cdot)$	Probability
q_x	One-year death probability at age x
${}_tq_x$	t-year death probability at age x
r	Annual interest rate
$r(\cdot)$	Relative survival function
$s(\cdot)$	Smooth function
$S(\cdot)$	Survival function
$\widehat{S}(\cdot)$	Estimated survival function
$S_n(\cdot)$	Net survival function
$\widehat{S}_n(\cdot)$	Estimated net survival function
$S_P(\cdot)$	Survival function of the general population
$S_u(\cdot)$	Survival function of the uncured observations
t_j	j ordered distinct event times
T	Non-negative continuous random variable representing the real time to the event of interest
TTC_B	Time-to-cure as defined by Boussari et al. (2018)
$\widehat{TTC}_B$	Estimated value of time-to-cure as defined by Boussari et al. (2018)
TTC_C	Time-to-cure as defined by Chauvenet et al. (2009)
TTC_D	Time-to-cure as defined by Dal Maso et al. (2014)
$v(\cdot,\cdot)$	Discount factor
w	Waiting period
YLL^c	Number of years of life lost by the entire cohort
YLL^i	Number of years of life lost per individual
X_t	Random variable giving the state occupied at time t
$\boldsymbol{Z}$	Vector of covariates, with $\boldsymbol{Z}^t = (Z_1, \ldots, Z_p)$
$\boldsymbol{z}_i$	Observed vector of covariates for the i^{th} individual, with $\boldsymbol{z}_i^t = (x_{i1}, \ldots, x_{ip})$
Z_t	Random variable defining the time spent in the state occupied at time t

∞	Infinity
$\widehat{\beta}_i$	Coefficient estimate of the i^{th} parameter
δ	Instantaneous force of interest
κ	Amount of capital
$\lambda(\cdot)$	Hazard function
$\lambda_E(\cdot)$	Excess hazard
$\lambda_O(\cdot)$	Overall hazard rate
$\lambda_P(\cdot)$	Population hazard
$\Lambda(\cdot)$	Cumulative hazard function
μ_x	Force of mortality at age x
μ^{ij}	Transition intensity from state i to state j
$\mu^{i\bullet}$	Exit intensity from state i
π	Proportion of cured observations
π_0	Net single premium
τ	Time horizon

Introduction 1

Property loans are often accompanied with mortgage insurance that pays the balance of the loan if the mortgagor dies. Coverage is usually awarded in the form of term insurance with decreasing sum insured, with the amount of death benefit diminishing as the debt decreases. This is common practice in Belgium, with about 170,000 new mortgage loans per year, mainly contracted by young adults acquiring their first family house (statistics from the Belgian Central Credit Register indicate that 36% of new mortgage loans in 2017 were contracted by borrowers younger than 35 years and about 68% were granted to borrowers younger than 45 years).

Based on answers to a health questionnaire, insurers evaluate applicant's health status. As with any other term life insurance product, applicants with poor health conditions may be denied insurance or charged increased amounts of premium compared to standard conditions. In extreme cases, this may prevent them from accessing property (in case of a house loan) or develop their business project (in case of a professional loan). It is normal that, like any other applicants, clients who have had cancer in the past must fill in a health questionnaire at the time of the application for a loan. However, the problem comes from the fact that, although people who have survived cancer are not particularly in poor health at the time of application, they are still often penalized because of their past medical history. Moreover, filling such health questionnaires may create frustration for patients having survived cancer and which was diagnosed many years ago. Having repeatedly to answer questions related to this disease has psychological consequences and being charged higher premiums or denied coverage, based solely on their medical history rather than their actual state of health, generates a feeling of discrimination (Massart, 2018). This is often felt as a double penalty by cancer survivors. For this reason, several EU countries passed laws to ease access to mortgage insurance for long-term disease survivors. This materializes into the "right to be forgotten" adopted in several EU member states, granting access to insurance after a waiting period of at most 10 years starting at the end of the successful therapeutic protocol.

Combining concepts from biostatistics and actuarial sciences, there are three topics that dominate the thesis. The first one is related to the right to be forgotten in insurance. In particular, the aim is to develop a method to adequately estimate the threshold after which cancer patients can be considered as cured. For some types of cancer, survivors actually have a chance of survival comparable to that of the general population only after a few years after diagnosis, or pose a moderately increased risk and could therefore be covered in the event of death. This involves measuring and quantifying the potential excess mortality so that the premiums claimed reflect the risk in terms of financial services. This topic has been studied in different contexts in the literature. Our contribution is that the right to be forgotten starting from the date of diagnosis, instead of the date of the end of the therapeutic treatment, is presented. Indeed, to avoid disputes in case of death due to potentially unclear definitions of a successful treatment, a waiting period starting from the diagnosis is promoted. The second dominant topic of the thesis is to find a proper way to adapt the

actuarial pricing of life insurance products to each category of risk, disease, person, etc. This problem is tackled by concentrating on pricing insurance covers on a market where such a right has been implemented. The third topic focuses on computing the incidence risk (i.e., the risk of developing cancer for a healthy individual), and the number of years of life lost due to cancer at different ages at diagnosis and given that the patient survived some years after diagnosis. This subject has been documented extensively in the biostatistical and epidemiological literature. Nonetheless, most studies refer to the number of years of life lost (due to a specific disease or condition) at the time of diagnosis, without taking the time survived since diagnosis into account. This is a major difference, given that the time since diagnosis is known to have an influence on survival for cancer patients.

To better clarify the three problems in the focus of the thesis, an introduction of the right be forgotten in Belgium and other European countries is needed. This is the goal of the next section. Next, a section is dedicated to each of the chapters, summarizing the main results and contributions to the literature. Finally, the introduction ends with an overview of the key concepts covered in this thesis. This is done in sufficient details and generality at the same time to grasp a sound understanding of the tools used in the present thesis while avoiding too advanced methodological developments.

1.1 The right to be forgotten

The first initiative dates back to 2007, when France launched the AERAS Convention (AERAS is the acronym for "*s'Assurer et Emprunter avec un Risque Aggravé de Santé*" in French, which could be translated as "insuring and borrowing under poor health conditions"). This agreement, signed by the public authorities, banking and insurance sectors, and patients' and consumers' associations purposed to allow people who survived a given amount of years after their cancer diagnosis or people suffering certain chronic diseases to access insurance in case of death or disability, as well as to guaranteed income insurance. Faced with a similar situation, this lead Belgian authorities to create the *Bureau du suivi de la tarification assurance solde restant dû* (`www.bureaudusuivi.be`) – *Opvolgingsbureau voor de tarifering schuldsaldoverzekering* (`www.opvolgingsbureau.be`) (which could be translated literally as the "Outstanding balance insurance pricing monitoring office") in 2014, in application of the law on insurance. This body reviews health questionnaires used by insurance companies selling mortgage insurance in Belgium and checks whether the proposed premium surcharges or cover denials are justified for impaired lives.

Considering long-term cancer survivors, France established in 2016 a "*droit à l'oubli*" (translated literally as "right to be forgotten" (RTBF) in the remainder of this text), that is, the right for an insurance applicant not to declare a previous cancer after a period of 10 years starting at the end of the therapeutic protocol. This 10-year waiting period is reduced to 5 years if the applicant suffered cancer before the age of 18. These periods of 10 and 5 years start from the date of the end of the therapeutic treatment, in absence of relapse within this period. The 10-year length of the waiting period has been further shortened for several types of cancer (and other non cancer related pathologies), as detailed in the AERAS Convention (see `www.aeras-infos.fr/cms/sites/aeras/accueil.html`) with reduced duration after which survivors have access to the right to be forgotten. Once this period after a successful treatment has elapsed, cancer survivors are not obliged to declare their pathology to insurers. If they choose to disclose their health history, insurance companies cannot take the condition into account in risk assessment and cannot refuse the insurance, nor impose a premium surcharge because of the pathology. In June 2022, France has led the way by reducing the RTBF period to 5 years for insurance contracts occurring before the borrower's 71^{st} birthday. The same year

in October, France went further in protecting cancer survivors against financial discrimination, by abandoning medical questionnaires for any loan with a maximum amount of 200,000€ per person and which ends before the age of 60, by imposing no extra premium or exclusion of cover for HIV under certain criteria, and by reducing the waiting period for hepatitis C.

In Belgium, the RTBF entered the insurance law in April 2019, making Belgium the second European country to adopt it after France. At that time, the waiting period was set to 10 years after successful treatment for all cancers, and 5 years if the diagnosis occurred before the age of 18. Based to a large extent on the reference tables published in the AERAS Convention, a Royal Decree dated May 26, 2019 lists certain types of cancer for which, depending upon entry criteria (such as cancer stage or age), the standard waiting period of 10 years from the end of active treatment is reduced. The RTBF has recently been adapted in Belgium, again following similar changes in France. As from November 2022, the standard waiting period opening the RTBF has been shortened, for all cancers, from 10 years to 8 years after successful treatment. This waiting period is reduced to 5 years if the insurance applicant was diagnosed of cancer before the age of 21. Since November 2022, this also applies to guaranteed income insurance for all workers, regardless of employment status. As from January 2025, the RTBF will be reduced to 5 years after the end of the therapeutic protocol for all cancer survivors, and reference grids will also be implemented for chronic diseases in addition to cancer. Shorter waiting periods for specific pathologies are still being discussed. Additionally, every two years, the Belgian Healthcare Knowledge Centre (KCE) evaluates the reference grid in the light of medical progress and available scientific data on the pathologies it covers. The KCE then proposes an adaptation of the reference grid, which it communicates to the *Bureau du suivi de la tarification*. The latter forwards the proposal, together with its opinion, to the Minister of Finance and the Minister of Social Affairs and Public Health. The reference grid may then be modified if necessary. The reader is referred to the KCE's report (2022) for a more comprehensive overview of the evolution of the RTBF in Belgium and the mandate of the KCE in the Belgian legislation with regards to the RTBF.

The RTBF has now also been installed through an agreement in Luxembourg, and through a legal framework in The Netherlands, Portugal, Romania, and more recently in Italy, Spain and Cyprus. Established in 2020 in Luxembourg, the agreement applies exclusively to the outstanding balance of an insurance policy for the acquisition of a principal residence or business premises, up to a maximum amount of 1,000,000€. It does not apply to the acquisition of a second home or to rental investments. Established in 2021 in The Netherlands, the law applies to life insurance for applicants under the age of 71. The decree establishes that it is no longer allowed to ask whether someone has already had cancer when, in the opinion of the health care provider who treated the applicant, there has been a complete remission and no recurrence has been diagnosed for an uninterrupted period of 10 years, starting from the moment when complete remission was established. The decree also states that if the candidate is aged under 21 at the time of cancer diagnosis, the period is reduced to 5 years. Established in 2022 in Portugal, the law gives people who have overcome serious illnesses such as cancer, HIV and diabetes the right to be forgotten when taking out a mortgage, consumer credit or insurance (whether compulsory or optional) linked to these loans, provided that 10 years have elapsed without interruption since the end of the therapeutic protocol, and 5 years have elapsed since the end of the therapeutic protocol in the case the illness was diagnosed before the age of 21. Also established in 2022, in Romania the right to be forgotten applies to adults 5 years after the end of treatment in case the cancer diagnosis occurred before the age of 18 years, and after 7 years following the end of their treatment and without any evidence of relapse or recurrence in case the cancer diagnosis occurred after the age

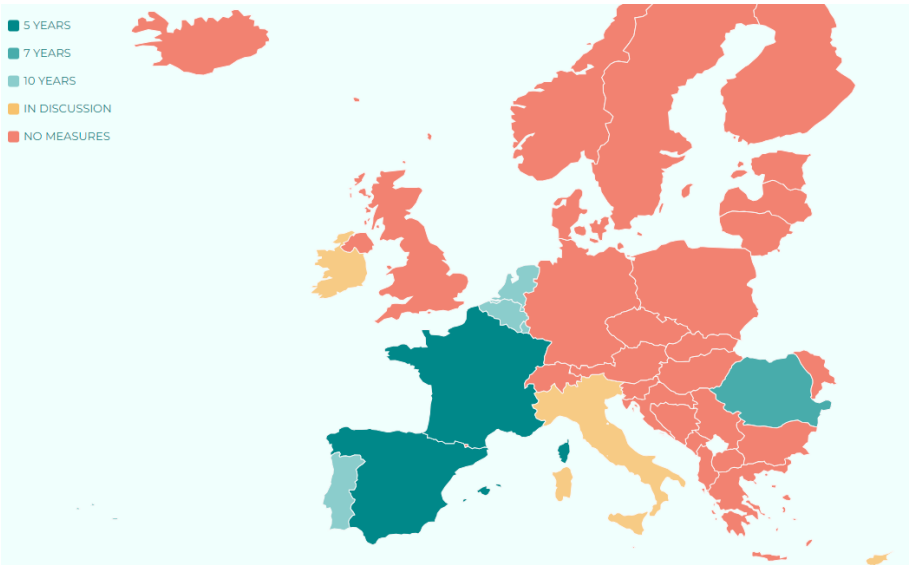


Figure 1.1: Status of the right to be forgotten in the EU/EEA as of February 2024. *Source: European Initiative on Ending Discrimination against Cancer Survivors*

of 18 years. Established in 2023 in Spain, the right to be forgotten in the field of insurance and banking products for cancer patients becomes effective once 5 years have elapsed since the end of treatment without relapse, regardless of age at diagnosis. The same year, Cyprus has followed by approving a law that states that no insurance company will be able to reject an application made by a cancer survivor if 10 years or more have passed since the completion of their treatment, or 5 years in the case of cancer diagnosed before the age of 21. Late 2023, Italy passed a law allowing cancer survivors not to share information about their previous condition with financial institutions, or adoption authorities, provided that 10 years had passed since the successful end of their treatment, or 5 years in the case of cancer diagnosed before the age of 21. Finally, Ireland, Denmark, Greece and Finland have adopted non-legislative frameworks which take the form of code of conduct and self-regulatory practices. Figure 1.1 shows the status of the RTBF in the EU/EEA as of February 2024.

Besides the implementation of the RTBF in these countries, it is being debated and advocated for at the European level to expand to the other EU countries as well. There are some ongoing discussions between Insurance Europe, the European Commission and the European Parliament on a possible EU-wide RTBF for cancer survivors. See for instance Scocca and Meunier (2020, 2022), as well as to "Survivorship challenge 3.4: Lack of knowledge of the stigma associated with cancer" listed in Lawler et al. (2021) purposing to take advantage of the existing legal framework in four EU member countries (France, Belgium, Luxembourg and The Netherlands) to investigate a pan-European legal framework on access to financial services for cancer survivors. Moreover, several European initiatives such as the Europe's Beating Cancer Plan, Horizon Mission on Cancer and The Consumer Credit Directive have led to the creation of recommendations aimed at giving consumers who survived cancer equal access to financial services.

1.2 Summary of the chapters

The manuscript is divided into several chapters. The main findings and contributions of each of the chapters are summarized below.

Chapter 2: Waiting period from diagnosis for mortgage insurance issued to cancer survivors This chapter, which mirrors our first paper (Soetewey et al., 2021), focuses on the application of several tools from biostatistics to assess the length of the waiting period opening the right to be forgotten. Although the establishment of such a right to be forgotten in several European countries and possibly at the European level is clearly an improvement for cancer survivors, it is shown, using data from the Belgian Cancer Registry, that there is room for further reducing the waiting period for some cancer types. In particular, the chapter aims to show that for some types of cancer (with melanoma and thyroid as examples), survivors actually have a survival comparable to that of the general population, that is, excess mortality is negligible. It is also demonstrated that patients having survived long enough to some types of cancer (still with melanoma and thyroid as examples) can access life insurance market at standard insurance rates, contrarily to the common belief within the actuarial community. Moreover, there remains some ambiguity about what is considered as treatment and thus what marks the end of the therapeutic protocol. Therefore, a waiting period starting at diagnosis rather than at the end of the therapeutic treatment protocol is promoted in order to avoid disputes in case of death. Results appear to be particularly encouraging as they suggest a considerable shortening of the 10-year waiting period for some types of cancer.

Chapter 3: Semi-Markov modeling for cancer insurance This chapter, identical to our second paper (Soetewey et al., 2022), focuses on insurance covers on a market where a right to be forgotten has been implemented. More precisely, the products considered here are specifically related to the waiting period opening the right to be forgotten, with temporary covers restricted to that period to fill the gap in coverage on a market where such a right has been implemented. First, stand-alone products are studied, including cancer insurance with lump sum payment at diagnosis, or temporary life annuity starting at diagnosis. In the latter case, periodic payments may correspond to insurance premiums of another product, or even to loan reimbursement. Then, riders included in a package are discussed. Term insurance with accelerated death benefit paid as a lump sum at diagnosis or as a temporary life annuity starting at diagnosis are considered. Finally, products granting access to some specific insurance cover (such as mortgage insurance) during the waiting period opening the right to be forgotten are discussed. This is especially important at young age, to guarantee access to property and home ownership (in case of house loan) and to entrepreneurship (in case of professional loan) to cancer patients whose health status has improved but who cannot benefit from the right to be forgotten because the waiting period is not exhausted. The 3-state (healthy–ill–dead) Semi-Markov hierarchical model developed in Denuit et al. (2019) for long-term care insurance is adopted here for actuarial calculations. Semi-Markov transition intensities are estimated from cancer cases recorded by the Belgian Cancer Registry. Our proposals are illustrated through three cancers with clear differences in terms of incidence, survival after diagnosis, and waiting periods defined by Royal Decree: (i) melanoma, (ii) thyroid and (iii) female breast cancers. The obtained results suggest that a new offer could develop, targeting the particular needs of cancer patients.

Chapter 4: Health indices for disease incidence risk and duration in the Semi-Markov setting This chapter, identical to our third paper (Soetewey et al.,

2024), focuses on illustrating how common health indices (with disease incidence risk and years of life lost as examples) can be estimated based on a Semi-Markov 3-state illness-death model using cancer registry data. The main advantage of computing these quantities in a Semi-Markov context is that it allows to take into account the number of years a patient survived after the diagnosis. To the best of our knowledge, most studies refer to the number of years of life lost at the time of diagnosis, without taking the time survived since diagnosis into consideration. This is a major difference, given that time survived since diagnosis is known to have an influence on survival for cancer patients. Based on 161,007 melanoma, thyroid and female breast cancer cases recorded by the Belgian Cancer Registry, it appears that the probabilities of being diagnosed with cancer over the next 20 years for a healthy individual remain rather low for melanoma and thyroid cancers for both sexes, but considerably increases with age for female breast cancer. Results also suggest that, for female breast cancer, the number of years of life lost before the age of 70 years due to cancer is highest when diagnosed at young ages and then decreases with age at diagnosis, whereas for melanoma and thyroid cancers, it peaks when diagnosed at later ages (between 35 and 55 years depending on the cancer and sex). Whether the decrease with age at diagnosis is linked to a real decrease of the number of years of life lost or simply due to the fact that the younger the patient at diagnosis, the more years he or she can still loose before the age of 70 years will be discussed too. It also turns out that the number of years of life lost before the age of 70 due to cancer is larger for men than for women for both melanoma and thyroid cancers. Last, it is found that, for melanoma and thyroid cancer patients diagnosed between the age of 20 and 70 years, once they have survived their cancer for 10 years, the number of years of life lost before the age of 70 due to cancer remains below one year. This indicates that, up to the age of 70 years, these patients lose a limited number of years of life due to cancer compared to the general population.

Chapter 5: Right to be forgotten for mortgage insurance issued to cancer survivors: Critical assessment and new proposal This chapter, identical to our fourth paper (Soetewey et al., 2023), is a follow-up of Chapter 2. In Chapter 2, it has been proposed to determine the waiting period opening the right to be forgotten as the time after diagnosis needed for the premium to revert back to some acceptable level, expressed by means of regulatory life tables. However, this approach requires data up to 30 years after diagnosis (10 years of standard right to be forgotten plus the typical duration of the loan), or extrapolating the results up to that time horizon. In this chapter, it is shown that when survival statistics are only available over a shorter duration, it turns out that the length of the resulting waiting period opening the right to be forgotten may strongly depend on the extrapolation method. This problem, not arising from the method proposed in Chapter 2 but coming from the limited follow-up period for patients in some cancer registries (including the Belgian one), is not acceptable in the context of the right to be forgotten. This is why an alternative method is proposed here, based on a constraint imposed to the premium. This constraint is then transposed into a target on the conditional observed survival probabilities. The length of the waiting period opening the right to be forgotten can then be derived from the comparison of the conditional one-year survival probabilities of cancer patients with the corresponding probabilities at general population level. The main advantage is that the time from which the right to be forgotten can be exercised can be estimated from the available data only, without the need to extrapolate mortality rates beyond 10 years. For the sake of robustness, results obtained with the proposed approach are compared to results obtained with Kaplan-Meier estimate taken as a nonparametric reference. Furthermore, while cancer stage at diagnosis has not been taken into account in Chapter 2, the impact of the stage of the tumor at diagnosis on

waiting periods is investigated in this chapter.

Chapter 6: Conclusion In the last chapter, contributions of this manuscript are briefly summarized. Furthermore, some problems that could not be solved during the thesis and which possibly already appeared in the previous chapters are listed. Some general questions that arose while working on the topics from the previous chapters, and whose solution might either have important practical or conceptual implications for both biostatisticians and actuaries, are then proposed.

1.3 Prerequisites

Before delving into the different chapters of this PhD thesis, it is recommended for the reader to possess a foundational understanding of several key concepts and methodologies in survival analysis and actuarial sciences. These prerequisites are carefully curated to ensure a seamless comprehension of the intricate topics covered in the thesis.

For more comprehensive guides related to survival analysis, the reader is referred to Therneau (1997); Klein et al. (2003); Kalbfleisch and Prentice (2011); Andersen et al. (2012); Cox (2018); Legrand (2021) and Collett (2023). For a more applied outlook, the reader is redirected to Kleinbaum and Klein (1996); Allison (2010) and Moore (2016).

For more resources about health insurance and actuarial sciences in general, the reader is redirected to Bowers (1997); Gerber (2013); Pitacco (2014); Promislow (2014) and Dickson et al. (2019). For a more applied perspective, see e.g., Ruckman and Francis (2005) and Charpentier (2014).

1.3.1 Survival analysis

Survival analysis, also known as time-to-event analysis or duration analysis, is a branch of statistics aiming at analyzing the duration of time from a well-defined time origin until the occurrence of some particular event or end-point. In other words, we are interested in a certain event and we would like to analyze the time until the event happens (referred to as survival time).

Although survival analysis is rooted in medical and public health applications (with, for example, the time to death), it is now used in many domains. For example, one may also be interested in the time until:

- an unemployed person finds a job,
- a citizen is being arrested again after having been released from jail,
- a woman becomes pregnant for the first time,
- a machine breaks down,
- a company goes bankrupt,
- a customer buys a new product or stops its current subscription,
- a letter is delivered,
- a taxi picks you up after having called the taxi company,
- an employee leaves the company,
- etc.

As it can be seen, the event of interest does not necessarily have to be the death or the occurrence of a disease, but in all situations we are interested in analyzing the time until a specific event occurs. Note that several of these examples are clearly example of survival time with a cure fraction. This notion of cure will be defined later on, in Section 1.3.6.

Survival data requires a special set of statistical methods for three main reasons:

1. Survival times are always positive. The time until an event of interest occurs cannot be less than 0. Moreover, the distribution of survival times is generally not symmetric and tend to be positively skewed.

2. Different measures are of interest depending on the research question and the context. For instance, one could be interested in knowing (i) the probability that a cancer patient survives longer than 5 years after diagnosis, (ii) the typical waiting time for a cab to arrive after having called the taxi company, or, (iii) out of 100 unemployed people, how many are expected to have a job again after 2 months of unemployment.

3. Censoring is almost always an issue. When the event occurred before the end of the study, the survival time is known. However, for some individuals, the event is not yet observed at the end of the study. This results in (right) censored data, which can be seen, in some sense, as a type of missing data because the exact survival time for these individuals is missing.

For completeness, the different types of censoring and the concept of truncation are presented in the next subsection.

1.3.2 Censoring and truncation

We distinguish between three types of censoring; right, left and interval-censoring.

Right-censoring When the event is not yet observed at the end of the study (i.e., the real survival time is greater than the observed duration), this is referred to as right-censoring. There are three common types of right-censoring (Geskus, 2015); (i) administrative censoring, (ii) lost to follow-up and (iii) competing risk, and two less common; (i) type I and (ii) type II. The three most common types of right-censoring are detailed below, while we refer the interested reader to Legrand (2021) for a discussion on the two less common types of right-censoring.

- *Administrative censoring* Suppose that we study the time from diagnosis until death of a cancer cohort. Luckily, some patients will not die before the end of the study. Patients still alive at the end of follow-up for the analysis are referred to as administrative censoring.

- *Lost to follow-up* Individuals are considered as lost to follow-up when they stopped being under observation before the end of follow-up for the analysis and before the event occurred. This can happen, for instance, when a patient withdraws from the study or moves to another country.

- *Competing risk* Sometimes, another event occurs before the event of interest which prevents it from ever happening, or at least modify the risk of the event of interest to occur. The most straightforward example is when a cancer patient dies from a car accident.

Left-censoring Left-censoring occurs if a participant is entered into the study when the event of interest occurred prior to study entry but it is not known when exactly.

Interval-censoring Interval-censoring implies that the event occurred within a time interval (between two known dates, two visits, etc.); the exact moment of occurrence is not known.

In all situations above, survival time is not fully observed for all individuals. The goal of survival analysis is of course to analyze all available data, including information about censored patients. Note that an important underlying assumption for most standard statistical methods in survival analysis is that censoring is independent of the occurrence of events. This is usually referred to as independent or non-informative censoring, and means that censored subjects have the same survival prospects as subjects who are not censored and who continue to be followed. Most of the time, independent censoring is assumed, as it is the case throughout the present thesis. Nonetheless, it may happen that survival and censoring times are dependent on each other, for instance when the event of interest is death and a patient leaves the study because his or her health has deteriorated or, on the contrary, has greatly improved. In these situations, survival and censoring times are likely to be positively and negatively correlated, respectively (Delhelle and Van Keilegom, 2023). Several approaches have been proposed to deal with dependent censoring in different contexts, see for example, amongst others, Emoto and Matthews (1990); Rivest and Wells (2001); Deresa and Van Keilegom (2020); Czado and Van Keilegom (2023) and Deresa and Keilegom (2023).

Another common characteristic of survival data is truncation. In a nutshell, truncation arises from the fact that some observations are actually absent from the data. For an individual to be included in a study and monitored for the event of interest, he or she must initially fulfill a specific condition. In the context of cancer research for example, if an individual who has cancer dies without having been diagnosed with cancer, he or she was obviously not included in the study. Truncation can thus be seen more as a sampling selection problem. There are two types of truncation: (i) left and (ii) right truncation. As it is beyond the scope of this introduction, we refer the reader to Klein et al. (2003) for a discussion on truncation.

Due to the data being asymmetric, and due to the presence of censoring and truncation, classical analysis tools for continuous variables cannot be used. The aim of survival analysis is to model and describe survival data in an appropriate way, taking these particularities into account.

1.3.3 Common functions in survival analysis

Without going into too much detail, we lay the foundations with the most common function in survival analysis; the survival function. Two frequent functions which characterize the distribution of survival times are also presented; the hazard function and the cumulative hazard function.

1.3.3.1 Survival function

Let T be a non-negative continuous random variable, representing the real time until the event of interest. The survival function $S(t)$ is the probability that a randomly chosen individual is still at risk at time t, where $0 \leq t \leq +\infty$. For each t, it is given by

$$S(t) = P(T > t)$$

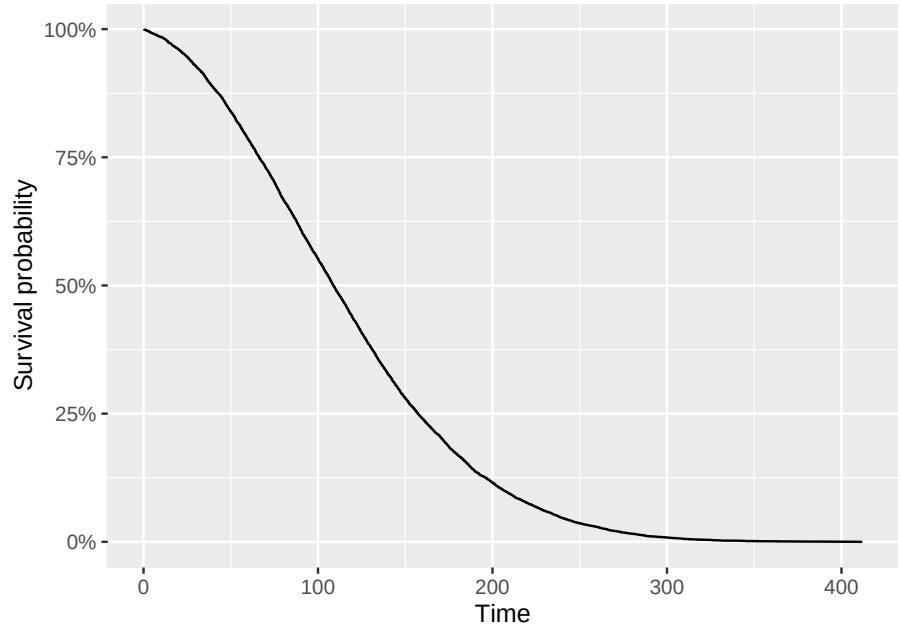


Figure 1.2: Example of a survival function $S(t)$

$$
\begin{aligned}
&= 1 - P(T \leq t) \\
&= 1 - F(t) \\
&= 1 - \int_0^t f(u)\mathrm{d}u,
\end{aligned}
$$

where $f(\cdot)$ and $F(\cdot)$ are the density and the cumulative distribution functions of T, respectively.

The survival function $S(t)$ is a decreasing function equal to 1 at $t = 0$ (i.e., $S(0) = 1$) and 0 at $t = \infty$ (i.e., $S(\infty) = 0$). Since it is a probability, it takes values in $[0, 1]$. Figure 1.2 shows an example of a survival function. The survival curve gives the proportion of individuals (or experimental units) who, as time goes on, have not experienced the event of interest. As time progresses, events occur, so the proportion who have not experienced the event decreases. In the context of this thesis where death is the event of interest, $S(t)$ gives the probability that a randomly selected cancer patient will survive beyond time t, or the proportion of cancer patients still alive after time t.

1.3.3.2 Hazard function

The hazard function $\lambda(t)$, or hazard rate, defines the instantaneous event rate at time t for an individual still at risk at that time. It can be obtained by

$$
\begin{aligned}
\lambda(t) &= \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \\
&= \frac{-d}{dt} \log\left(S(t)\right) \\
&= \frac{f(t)}{S(t)}.
\end{aligned}
$$

The hazard function is a positive function, not necessarily monotone. Since it is a snapshot of the data at each time point (measuring the proportion of individuals experiencing the event of interest at that specific moment among those who remain at risk of the event at that time), it can have many different shapes and is therefore a useful tool to summarize survival data. In the context of cancer research when death is the event of interest, $\lambda(t)$ measures the instantaneous risk of dying right after time t given the patient is alive at time t.

To link the hazard rate with the survival function; the survival curve represents the hazard rates. A steeper slope indicates a higher hazard rate because events happen more frequently, reducing the proportion of individuals who have not experienced the event at a faster rate. On the contrary, a gradual and flatter slope indicates a lower hazard rate because events occur less frequently, reducing the proportion of individuals who have not experiences the event at a slower rate. More formally:

$$S(t) = \exp\left(-\int_0^t \lambda(u)\mathrm{d}u\right).$$

1.3.3.3 Cumulative hazard function

The cumulative hazard function $\Lambda(t)$, which corresponds to the total amount of risk experienced up to time t, is defined as:

$$\begin{aligned}\Lambda(t) &= -\log\left(S(t)\right)\\ &= \int_0^t \lambda(u)\mathrm{d}u,\end{aligned}$$

Since the cumulative hazard function is a cumulative measure (based on the proportion of individuals who have encountered the event up to a given time, relative to the total number of individuals who were at risk at the beginning of the study), it is an increasing function. Moreover, it is defined in $[0, +\infty]$. If the event of interest is death, then $\Lambda(t)$ summarizes the risk of death up to time t, given that death has not occurred before t (Collett, 2023). Note that the survival function can be expressed in terms of the cumulative hazard function as follows

$$S(t) = exp(-\Lambda(t)).$$

1.3.4 Estimation of the survival function

To estimate the survival function, which provides information about the overall survival, an estimator that is able to deal with this type of data is needed. Amongst the different methods for estimating the survival function, the most common one is the nonparametric Kaplan-Meier (1958) estimator:

$$\widehat{S}(t) = \prod_{j:t_j \leq t}\left(1 - \frac{O_j}{n_j}\right)$$

for each t, with $0 < t_1 < t_2 < \ldots < t_j$ corresponding to the j ordered distinct event times, O_j corresponding to the number of events observed for each distinct event time t_j, and n_j corresponding to the remaining number of individuals at risk for each distinct event time t_j.

The advantages of this estimator are that:

- it is simple and straightforward to use and interpret,

- it is a nonparametric estimator, so it estimates a survival curve from the data and no assumptions is made about the shape of the underlying distribution, and
- it gives a graphical representation of the survival function(s), useful for illustrative and descriptive purposes.

It is a decreasing step function with a downward jump at each event time. It starts at 1 and reaches 0 if the largest observed survival time corresponds to an event, whereas it starts at 1 but does not reach 0 if the largest observed survival time is censored. The principle behind this estimator is that surviving beyond time t_i implies surviving beyond time t_{i-1} and surviving at time t_i. Note that an important assumption for the estimation to hold is the independent censoring, also known as non-informative censoring. Strictly speaking, the assumptions of independent and non-informative censoring are not identical. This is however beyond the scope of this introduction. The interested reader is referred to Lagakos (1979) for more information about the difference between the two concepts.

In order to estimate a survival function with the Kaplan-Meier estimator, the following two pieces of information are required:

1. the time until the event of interest or the time until the censoring, and
2. the event status (whether the event happened or not, so whether the subject is censored or not).

The survival curve resulting from the Kaplan-Meier estimator can be seen as a descriptive statistic for survival data. This estimator is used to estimate the overall survival function $S(t)$, without distinguishing according to causes of death (Belot et al., 2019). Several quantities can be obtained from this estimator, such as the median survival time (i.e., the time beyond which 50% of the individuals in the population under study have experienced the event of interest) or the probability that a cancer patient will survive, say, more than 1 or 5 years after diagnosis (known as the 1 and 5-year survival probability). Other important concepts in survival analysis is the relative and net survival. They are detailed in the next subsection.

1.3.5 Relative and net survival

Information on cause of death is often unavailable or unreliable and not all deaths of cancer patients can be easily classified as a death due to the cancer of interest or due to another cause (Percy et al., 1981). Relative survival, which does not require information on the cause of death, provides a measure of the excess mortality experienced by cancer patients by comparing the mortality in the cancer population with the mortality of a comparable group in the general population. This led the relative survival to become the standard measure of patient survival for population-based cancer registries, as shown by its prominence around the world in studies related to cancer survival (Ries et al., 2002; Coleman et al., 1999; Berrino et al., 1999; Perme et al., 2012, 2016; Pavlič and Pohar Perme, 2019).

Relative survival models are divided into two types: (i) additive and (ii) multiplicative models. See Pohar and Stare (2006) for a more detailed discussion. Despite the wide acceptance of multiplicative specifications within the actuarial community, it turns out that additive models are biologically more plausible in cancer studies and provide a better fit to the data (Dickman et al., 2004; Buckley, 1984; Hakulinen and Tenkanen, 1987; Esteve et al., 1990; Bolard et al., 2001). The additive specification is thus favored here. In the additive framework, the hazard rate at time t since diagnosis for cancer patients is decomposed into two additive components: (i) the population

hazard denoted as $\lambda_P(t)$, and (ii) the excess hazard specific for the cancer of interest denoted as $\lambda_E(t)$:

$$\lambda_O(t) = \lambda_P(t) + \lambda_E(t),$$

where $\lambda_O(t)$ corresponds to the overall hazard rate. The population hazard $\lambda_P(t)$ is usually estimated on the basis of external data such as population life tables, which are usually stratified according to the main factors affecting patient survival such as age, gender, and calendar year.

The relative survival function $r(t)$ corresponds to the ratio of the survival function of the studied group $S(t)$ to the survival function of a comparable group (i.e., with the same characteristics) from the general population $S_P(t)$ (Dickman et al., 2004):

$$r(t) = \frac{S(t)}{S_P(t)}.$$

Net survival is a measure of patient survival corrected for the effect of other causes of death (Dickman et al., 2004). It represents the (hypothetical) survival that would be observed if the only possible cause of death was the disease of interest (Berkson and Gage, 1950; Schaffar et al., 2017). This allows to compare, among others, treatment success in different countries without being affected by the differences in the general population mortality (e.g., different diagnosis years, different countries, etc.). If we are in a situation where we have information on the cause of death, then estimation methods for net survival can be estimated in a competing risks framework (a situation where several risks "compete" to become the actual cause of death). On the other hand, net survival can also be estimated when the cause of death is unknown by making use of information from the general population. In this different framework, net survival can then be estimated using the relative survival method. Since the Belgian Cancer Registry (BCR) does not collect information on the cause of death, we are in this latter situation. The net survival function, denoted $S_n(t)$, is thus derived from the excess mortality hazard:

$$S_n(t) = \exp\left(-\int_0^t \lambda_E(u)\mathrm{d}u\right).$$

Chapter 2 presents a more detailed explanation, together with illustrations, of the concepts of relative and net survival.

1.3.6 Cure models and time-to-cure

In survival analysis, a fundamental premise is that if the follow-up period is sufficiently long, all subjects under examination will experience the event of interest. Essentially, this assumption posits that all subjects are susceptible to the event. However, in various fields, this assumption may not hold true. An example in the medical field is when considering the time to recurrence after treatment for a curable disease; some patients may be cured and thus not experience a recurrence of the disease. Similarly, it can be the case in other fields such as economics (e.g., an unemployed person may never find a new job), engineering (e.g., a machine or device may never fail), demography (e.g., a woman may never have a child), etc. In such scenarios, the population includes a subset of individuals who are said to be non-susceptible or who have been cured for the event of interest. The presence of a fraction of individuals who will never develop the event of interest (i.e., known as the cure fraction or cure rate) led to the need for a specific class of survival models which would be able to accommodate this assumption (Maller and Zhou, 1996). We highlight the four main motivations behind the development of this new type of model, referred to as cure model in the statistical literature (Legrand, 2021).

First, as t goes to ∞, the survival function does not reach 0 and the cumulative hazard function is bounded from above, that is, $\lim_{t\to\infty} S(t) > 0$ and $\lim_{t\to\infty} \Lambda(t) < \infty$. In this specific case, the survival function is said to be improper and the cumulative hazard function is constrained from above, implying that the accumulated instantaneous risk of experiencing the event will not reach infinity as the follow-up time goes to infinity. Visually, the estimated survival function reaches what is called a long and stable "plateau" (including a large number of right-censored observations at the level of the sample) after a sufficiently long follow-up period. This happens because a fraction of the individuals who entered the study will never experience the event of interest. Second, a new parameter of interest emerges. In the context of cancer research for instance, when such a cure can indeed be achieved, the benefits of a new treatment is not only limited to how the treatment helps to postpone death or a relapse of the disease, but also what proportion of cancer patients can be considered statistically cured (Maetani and Gamel, 2013; Legrand and Bertrand, 2019). Third, the existence of a cure fraction might result in a violation of the proportional hazards assumption (an important assumption for many standard survival analysis tools). Fourth, a challenge in survival data analysis arises when there are, at the same time, a fraction of cured and some censored observations. Without censoring, individuals are assumed to belong to either a cured or uncured sub-population, but right-censoring complicates this distinction by making the cure status partially latent. While the cure status is fully observed for those who experience the event of interest (they are obviously uncured), the cure status of the right-censored cases cannot be distinguished as it is not observed for these subjects. Furthermore, standard survival analysis techniques assume that censored individuals have the same survival pattern after censoring as non-censored ones. This becomes problematic if censored cases include both cured and uncured subjects.

Hence, cure models have been formulated to take into account these situations, and more broadly, to investigate not only the time until the event of interest (the goal of most classical survival analysis techniques) but also to estimate the share of the population that will not develop the event (and thus the probability to be cured). Cure models are particularly suitable for some cancer sites when the event of interest is death or recurrence. Indeed, for situations where recurrence is the endpoint, if the treatment is successful, the patient will never suffer a relapse of the cancer. For situations where death is the endpoint, although it is evident that no one can be cured of death, cure models are also especially appropriate for less aggressive cancers such as, among others, the three considered in this thesis because a considerable percentage of subjects exhibits long-term survival (who are sometimes referred to as long-term survivors and who can be considered as statistically cured (Lambert, 2007; Othus et al., 2012; Yilmaz et al., 2013)).

There are two main types of cure models in the literature: (i) mixture cure models (the most common type, based on the seminal work by Boag (1949) and Berkson and Gage (1952)) and (ii) non-mixture cure models (Andrei et al., 1996; Chen et al., 1999; Tsodikov et al., 2003); see Amico and Van Keilegom (2018) for an overview. In the present thesis, we consider only the family of mixture cure models. In this approach, the population is considered as a mix of two types of individuals, that is, (i) the cured subjects who will never develop the event of interest (referred to as long-term survivors) and (ii) the uncured subjects who, if not censored, will develop the event of interest (Lambert et al., 2006). We refer the reader to Legrand (2021) for a detailed discussion on the differences and the link between the two types of cure models.

The time-to-cure (TTC) is generally referred to as the time after which subjects can be considered as long-term survivors. Different approaches to define the TTC have emerged in the literature. In this thesis, we use the TTC as introduced by Boussari

et al. (2018), and defined as the shortest time from which the conditional probability of being cured at a given time t after diagnosis knowing that the patient was alive up to time t is close to 1, that is TTC is the smallest value of t such that, for some given (small) value of ϵ

$$\frac{\pi}{S(t)} = \frac{\pi}{\pi + (1-\pi)S_u(t)} \geq 1 - \epsilon, \tag{1.1}$$

where π is the proportion of cured patients and $S_u(t)$ is the survival function of the uncured population. The main advantage of TTC is that, although the results may depend on the chosen definition, it is a simple and straightforward indicator to set the time after which a patient who had cancer should not be penalized anymore when applying for mortgage insurance.

Chapter 2 presents a more detailed explanation, together with illustrations, of the concepts of cure models and time-to-cure.

1.3.7 Number of years of life lost

Over the last decade, the number of years of life lost (YLL) became a popular tool in biostatistics and epidemiology to measure discrepancies in life expectancy or mortality. The idea behind YLL is to quantify the number of years of life a specific cohort of patients has lost due, for example, to a given disease, compared to the general population. This measure, as defined by Andersen (2013) and Andersen et al. (2013), has the advantage that it is measured on a time metric (usually in years) making its interpretation easy for policy-makers and meaningful for gauging public health outcomes (Latouche et al., 2019).

It was first introduced to measure the reduction in life expectancy for a group of individuals compared to a hypothetical cohort where no one dies before a given age (Andersen, 2013). However, in most situations, it may seem more natural to measure the reduction in life expectancy for a group of individuals compared to a reference population (where some years of life are lost because of some standard or background mortality rates). In this sense, YLL can be used to estimate the number of years a specific cohort of patients (cancer patients, for instance) are expected to lose compared to the general population (i.e., the reference population to which the cancer cohort is compared). The difference between the life expectancy of the general population and the one of the considered cohort of patients corresponds to YLL. This measure is sometimes referred to as excess YLL because it is the number of years of life patients lose in excess of that seen in the general population. The larger this measure, the more important the societal burden of the disease or condition.

There are two distinct metrics of YLL within the epidemiological literature. First, the cohort-based YLL, which measures the total number of years of life lost by an entire cohort due to a specific disease or condition. This metric is useful for estimating the overall impact of a given disease or condition on a population, informing resource allocation, public health priorities, etc. Second, the individual-based YLL, which quantifies the average number of years of life lost per individual, for those suffering from a particular disease or condition. This measure helps assessing how a diagnosis affects the life expectancy of an individual and is useful for evaluating the health impacts of diseases on a person-by-person basis. While the individual-based YLL provides an average per person, the cohort-based YLL represents the sum of years of life lost for a cohort of patients or a group of individuals within a population. Note that individuals may not necessarily lose some years of life compared to the general population; they could potentially gain years, as seen with elite athletes for which survival may be better than that of the general population (Antero-Jacquemin et al., 2018). The main advantages of YLL are that (i) it is measured on a time metric (usually in years), facilitating its interpretation and communication (Baade et al., 2015; Licher

et al., 2019), (ii) information on the cause of death is not required to estimate it, making it a practical measure for population-based studies in which the cause of death is often unavailable or unreliable (Percy et al., 1981), and (iii) it can be computed for any time horizon and for a comprehensive list of causes of death (see for instance Aragon et al. (2008) who ranked leading causes of premature death and Chu et al. (2008) who measured the health impact of several common cancers, both based on YLL). When applied to cancer patients, on the one hand, the cohort-based YLL represents the total number of years of life lost by the cancer cohort. This is useful to compare, for instance, the societal burden of cancer with other diseases or between different countries. On the other hand, the individual-based YLL can be interpreted as the average number of years of life lost that a cancer patient experiences from the time of diagnosis in comparison to an healthy individual of the same age (and possibly sex, year and other covariates such as ethnicity or socio-economic factors). In this thesis, it is the individual-based YLL which is chosen and illustrated as it resonates more in the patient-clinician communication. Formally, the individual-based YLL in a certain time interval is the sum of life years lost due to (i) population mortality (governed by mortality rates in that reference population) and due to (ii) the cancer of interest. This quantity can be computed based on the difference between the estimated survival observed in the general population and the one observed in the cohort of cancer patients.

Chapter 4 presents a more detailed discussion, together with illustrations, of the concept of YLL.

1.3.8 Multi-state models

In research where the outcome of interest is survival from the time origin to death, the occurrence of other non-fatal incidents throughout the follow-up period could offer additional insights into the underlying mortality process. The succession of intermediary events defines an event history, which can be taken into account to increase knowledge of a biological process resulting to death. Such data and mechanisms can be analyzed using multi-state models (MSM) (Collett, 2023). More generally, MSM are a powerful statistical approach to study the evolution of individuals between several "states" (see Andersen et al. (2012) and Hougaard (1999) for a general review). MSM can be seen as an extension of classical survival analysis, in which there are only two states (i.e., alive and dead) and only the transition from being alive to being dead is considered (De Wreede et al., 2010; Geskus, 2019; Putter et al., 2007). Unlike classical survival models, MSM are used to model processes which go from an initial state (for instance "healthy") to a terminal (also referred to as absorbing) state (for example "dead"), but where more than two states are considered, some being transient. Thus, MSM offer a complete and informative representation of the occurrence of intermediate events on the pathway to some final event, notably via transition probabilities and transition intensities which govern movements between the different states depending on the state currently occupied and the time spent in that state (referred to as the sojourn time) (Andersen and Pohar Perme, 2008; Touraine et al., 2016).

In this thesis, a 3-state model, assuming that an individual can either be "healthy", "ill" (diagnosed with cancer), or "dead" is considered. See Fig. 1.3 for a visual representation of the model, often referred in the literature to as the "(3-state) illness-death model" without recovery. Individuals are initially with no cancer detected, thus considered as healthy. Then, they may be diagnosed with cancer and die, or they may die without having been diagnosed with cancer. Note that this 3-state model is, in its mathematical concept, similar to the well-known SIR model (susceptible – infected – recovered) in epidemiology (Anderson, 1991; Kermack and McKendrick, 1927). The

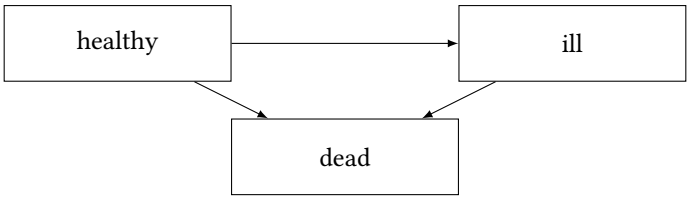


Figure 1.3: Visual representation of the 'illness-death model' without recovery for cancer patients

difference with our 3-state illness-death model is that a susceptible individual must go through the infectious state before being recovered, he or she cannot go directly from "susceptible" to "recovered". Main motivations for using a MSM are often to obtain (i) more biological insight into the disease or recovery process of a patient, and (ii) more accurate predictions than standard models neglecting intermediate states. Indeed, by incorporating intermediate events, predictions are adjusted in the course of time, giving more precise information about survival time (De Wreede et al., 2010; Geskus, 2019).

When considering MSM, the following notions must be distinguished: (1) Markov and Semi-Markov, and (2) homogeneous and non-homogeneous.

- Markovian models depict transitions solely based on the current state, disregarding previous states.
- Semi-Markovian models incorporate not only the current state but also the duration spent in that state.
- Homogeneous models assume that transitions between states remain constant over time.
- Non-homogeneous models allow transitions between states to vary over time.

In the context of cancer research, a homogeneous Markov model assumes the same mortality regardless of the time elapsed since diagnosis, which contradicts observed mortality patterns. Therefore, a Semi-Markov model, considering time since diagnosis, is preferred. Additionally, as transitions may depend on patient's age, non-homogeneous modeling (which accounts for age-dependent transitions), is essential. Consequently, a non-homogeneous Semi-Markov illness-death model is used in this thesis to consider both age and time since diagnosis, and thereby enhancing the precision of our calculations.

Chapter 4 presents a more detailed explanation, together with illustrations, of the concept of MSM.

1.3.9 Expected present value

The field of actuarial science plays a crucial role in assessing and managing risks across various domains. One prominent application of actuarial science in the context of this thesis is in the realm of mortgage insurance, where actuaries are tasked with calculating premiums to safeguard lenders against potential default risks associated with mortgage loans. Mortgage insurance serves as a protective mechanism for lenders, enabling them to mitigate losses in the event of borrower default.

The actuarial determination of prices for mortgage insurance entails the computation of statistical measures related to the frequency and amounts of future cash flows, where premiums represent the anticipated value of future benefit cash flows,

evaluated at present value for a specified interest rate structure. These cash flow probabilities are contingent upon the survival of the policyholder within the period of reimbursement. Survival of a policyholder is evaluated thanks to life tables (also referred to as mortality or actuarial tables), which present how mortality impacts individuals from the general population across different ages, and also usually stratified by calendar year and sex. Life tables allow to compute several statistics, such as, among others, the probability that a policyholder alive at age x will reach age $x + t$, denoted ${}_tp_x$ in the actuarial literature and defined as follows

$$ {}_tp_x = \frac{l_{x+t}}{l_x}, $$

with l_{x+t} and l_x corresponding to the number of individuals living at the beginning of age $x + t$ and x, respectively. Hence, the probability distribution of the future lifetime for a policyholder of any age can be deduced from life tables. Moreover, the complementary of ${}_tp_x$, that is, the probability that a policyholder alive at age x does not reach age $x + t$, denoted ${}_tq_x$, is defined as

$$ \begin{aligned} {}_tq_x &= 1 - {}_tp_x \\ &= \frac{l_x - l_{x+t}}{l_x}. \end{aligned} $$

Note that the one-year survival probability of an individual aged x is conventionally denoted p_x, rather than ${}_1p_x$. The same convention is applied to its complement, q_x, which denotes the one-year death probability at age x (also known as the mortality rate at age x). For a more exhaustive coverage of the use and practice of life tables, the reader is referred to Keyfitz and Caswell (2005).

To make the parallel with survival analysis,

$$ {}_tp_x = \exp\left(-\int_0^t \mu_{x+s}\mathrm{d}s\right), $$

and

$$ {}_tq_x = \int_0^t {}_sp_x\mu_{x+s}\mathrm{d}s, $$

where μ_{x+s} is the force of mortality, or hazard rate, at age $x + s$.

As premiums for a mortgage insurance are usually paid by the insured at different times (which are referred to as periodic premiums, and which are often paid annually or monthly), a way to evaluate the current value of these monetary amounts available at different times is required. An important concept in actuarial sciences, which is used for this purpose, is the concept of present value. The present value can be considered as the value, in current money, of a series of cash flows that are available at different periods of time (Spedicato et al., 2013).

Moreover, when the occurrence of these monetary amounts is uncertain, which happen when the payments are paid with some given probabilities (contrarily to contracts purchased by a single premium in which case there is no uncertainty regarding premium income), it is the expected present value which is of prime interest. For a good comprehension of the concept of expected present value used in the context of this thesis, let us define the expected present value from a general point of view. For this, suppose i denoting the (constant) annual interest rate, and $v = (1 + i)^{-1}$ representing the discount factor. Considering also a series of payments $\boldsymbol{c} = (c_1, \ldots, c_k)$ due with probability $\boldsymbol{p} = (p_1, \ldots, p_k)$, at times $\boldsymbol{t} = (t_1, \ldots, t_k)$ (and with payments at dates $\{1, 2, \ldots, k\}$), the expected present value of those benefits is

$$\text{EPV} = \sum_{j=1}^{k} \frac{c_j \cdot p_j}{(1+i)^{t_j}} = \sum_{j=1}^{k} v^{t_j} \cdot c_j \cdot p_j. \tag{1.2}$$

Insurances of the person is divided into three main types of insurance products (Pitacco, 2014):

1. Life insurances and life annuities: benefits depend on survival and death of the insured;
2. Health insurances: benefits depend on the health status and related financial consequences, but also on the lifespan of the insured:
   - Sickness insurance covers medical expenses, benefits in the event of temporary or permanent disability, and possibly hospitalization benefits;
   - Accident insurance covers the risks (in particular, but not limited to, the risks of permanent disability and death) caused by an accident;
   - Disability insurance provides benefits in case of temporary or permanent disability. There are several types of covers among the disability insurance, e.g., the income protection (IP), which provides a periodic income to the policyholder if he or she is prevented from working by sickness or injury;
   - Critical illness insurance (CII), or dread disease (DD) insurance, provides benefits when the policyholder is diagnosed with a severe disease, as specified by the policy conditions (commonly; heart attack, cancer, stroke, and coronary artery diseases requiring surgery). This type of insurance often constitutes a rider benefit to a basic life policy including death benefit, and can be used to cover medical expenses or to provide protection against potential loss of income;
   - Long-term care insurance (LTCI) provides the insured with financial support, while he or she needs nursing and/or medical care because of chronic or long-lasting conditions or ailments.
3. Other insurances of the person: benefits depend on some specific events such as marriage, birth of a child, education of children, etc.

These insurance products may provide one-year, multi-year, or lifelong covers. For instance, car insurance (which belongs to accident insurance) is usually provided by one-year policies, while income protection is typically based on multi-year policies, and whole life sickness insurance covers the policyholder during his or her entire lifespan. The duration of the cover defines the insured period (also referred to as the coverage period), which corresponds to the time interval during which the insurance cover operates. In principle, a benefit is payable only if the claim falls within the insured period. Nonetheless, restrictions may apply, for instance when a deferred period is included in the policy. A deferred period, also known as elimination or probationary period, refers to the period of time following the policy issue during which the insurance cover cannot be exhausted by the policyholder. Moreover, the monetary benefit can be a lump sum (i.e., a single payment made at a particular time, for instance a lump sum paid at the time of cancer diagnosis so that the patient can use the amount to face out-of-pocket expenditures related to treatment) or follows an annuity-like structure (i.e., periodic payments, for example monthly payments so that the cancer patient can be compensated for the loss of income due to cancer). Last but not least, it is worth noting that these insurance products can be combined, adding more complexity to the actuarial structure and computation. However, as

most insurances of the person can be represented by a series of one or more payments whose occurrence, timing and present value are uncertain, they can be rewritten as particular expressions of Eq. (1.2). The insurance products considered in this thesis are no exception to the rule.

1.3.9.1 Common life and health insurance products

For completeness, we present the expected present value of the benefits for the main life and health insurance products in the area of insurances of the person, which will lay the foundations for a good understanding of the products considered in this thesis:

- Whole life insurance: a contract which promises payment of a lump sum when the insured dies
$$\overline{A}_x = \int_0^\infty (1+i)^t \cdot {}_tp_x \cdot \mu_{x+t} \mathrm{d}t,$$
with, as we recall, μ_{x+t} denoting the force of mortality, or hazard rate, at age $x+t$.

- Term life insurance: a contract which promises payment of a lump sum in case the insured dies within n years from the issue of the contract
$$\overline{A}^1_{x:\overline{n}|} = \int_0^n (1+i)^t \cdot {}_tp_x \cdot \mu_{x+t} \mathrm{d}t.$$
Note that the sum insured may be variable (e.g., a decreasing sum insured for a mortgage loan). Denoting $c(t)$ the amount of benefit in case of death at time t, the EPV of benefits becomes
$$\int_0^n c(t)(1+i)^t \cdot {}_tp_x \cdot \mu_{x+t} \mathrm{d}t.$$

- Deferred whole life insurance: a contract similar than the whole life insurance, except that the payment of the lump sum is deferred to the future (i.e., no payment in the first u years) and is contingent upon the survival of the insured
$${}_u|\overline{A}_x = \int_u^\infty (1+i)^t \cdot {}_tp_x \cdot \mu_{x+t} \mathrm{d}t.$$

- Whole life annuity-due: a contract which pays an amount at the beginning of each period while the insured is alive
$$\ddot{a}_x = \sum_{k=0}^{\infty} v^k \cdot {}_kp_x.$$

- Temporary life annuity-due: a n-year contract which pays an amount at the beginning of each period until the term of the contract or the death of the insured, whichever occurs first
$$\ddot{a}_{x:\overline{n}|} = \sum_{k=0}^{n-1} v^k \cdot {}_kp_x.$$

- Whole life annuity-immediate: a contract similar than the whole life annuity-due, except that the amount is paid at the end of each period
$$a_x = \ddot{a}_x - 1.$$

- Temporary life annuity-immediate: a contract similar than the temporary life annuity-due, except that the amount is paid at the end of each period

$$a_{x:\overline{n}|} = \sum_{k=0}^{n} v^k \cdot {}_kp_x.$$

- Deferred whole life annuity-due: a contract similar than the whole life annuity-due, except that the payment of the lump sum is deferred to the future (i.e., no payment in the first u years) and is contingent upon the survival of the insured

$${}_{u|}\ddot{a}_x = \sum_{k=u}^{\infty} \frac{1}{(1+i)^k} \cdot {}_kp_x.$$

- Pure endowment: a contract which promises payment of a lump sum at the end of the contract if the insured is alive at the end of the contract

$$A_{x:\overline{n}|^1} = v^n \cdot {}_np_x.$$

- Endowment: a contract which pays a lump sum at the minimum between the death of the insured and the term of the contract (i.e., it provides a combination of a term insurance and a pure endowment)

$$\overline{A}_{x:\overline{n}|} = \overline{A}^1_{x:\overline{n}|} + A_{x:\overline{n}|^1}.$$

The concept of expected present value for several financial products will be used subsequently in the following chapters.

The remainder of this thesis is structured as follows. Chapters 2 to 5, with each chapter corresponding to a paper, are presented. The final Chapter 6 concludes the thesis with a general discussion, the questions that remain unanswered and avenues for future research.

Waiting period from diagnosis for mortgage insurance issued to cancer survivors

2

This chapter corresponds, notwithstanding a few minor improvements, to an article carrying the same name as the chapter and published jointly with Pr. Catherine Legrand, Pr. Michel Denuit and Dr. Geert Silversmit in the *European Actuarial Journal* in 2021.

Abstract
Massart (2018) testimonial illustrates the difficulties faced by patients having survived cancer to access mortgage insurance securing home loan. Data collected by national registries nevertheless suggest that excess mortality due to some types of cancer becomes moderate or even negligible after some waiting period. In relation to the insurance laws passed in France and more recently in Belgium creating a right to be forgotten for cancer survivors, the present study aims to determine the waiting period after which standard premium rates become applicable. Compared to the French and Belgian laws, a waiting period starting at diagnosis (as recorded in national databases) is favored over a waiting period starting at the end of the therapeutic treatment protocol. This aims to avoid disputes when a claim is filed. Since diagnosis is often recorded in the official registry database, as is the case for the Belgian Cancer Registry, its date is reliable and unquestionable in case of claim. Based on 28,994 melanoma and thyroid cancer cases recorded by the Belgian Cancer Registry, the length of the waiting period is assessed with the help of widely-accepted tools from biostatistics, including relative survival models and time-to-cure indicators. It turns out for instance that a waiting period of 4 years after diagnosis is enough for 30-year-old thyroid cancer patients. This appears to be similar to the 3-year period starting at the end of treatment protocol according to the Belgian law in such a case.

Keywords: Term insurance, impaired lives, cancer, home loan, right to be forgotten.

2.1 Introduction

Property loans are often accompanied with mortgage insurance that pays the balance of the loan if the mortgagor dies. Coverage is usually awarded in the form of term insurance with decreasing sum insured, with the amount of death benefit diminishing as the debt decreases. This is common practice in Belgium, with about 170,000 new mortgage loans per year, mainly contracted by young adults acquiring their first family house (statistics from the Belgian Central Credit Register indicate that 36%

of new mortgage loans in 2017 were contracted by borrowers younger than 35 and about 68% were granted to borrowers younger than 45).

Based on answers to a health questionnaire, insurers evaluate applicant's health status and either impose surcharges in case of impaired lives or refuse to cover the risk. Filling such health questionnaires may create frustration for patients having survived cancer occurred many years ago. Having repeatedly to answer questions related to this disease has psychological consequences and being charged higher premiums or denied coverage generates a feeling of discrimination (Massart, 2018). This is often felt as a double penalty by cancer survivors.

The restricted access to insurance cover is often regarded as a barrier to property and home ownership (in case of house loan) and to entrepreneurship (in case of professional loan). This lead Belgian authorities to create the *Bureau du suivi de la tarification assurance solde restant dû* (`www.bureaudusuivi.be`) – *Opvolgingsbureau voor de tarifering schuldsaldoverzekering* (`www.opvolgingsbureau.be`) (which could be translated literally as the "Outstanding balance insurance pricing monitoring office") in 2014, in application of the law on insurance. This body reviews health questionnaires used by insurance companies selling mortgage insurance in Belgium and checks whether the proposed premium surcharges or cover denials are justified for impaired lives. In 2017, only 16% of the 454 cases submitted by insurance applicants to the Bureau resulted in improved policy conditions, as it can be read from the annual report published by the Bureau (2018).

Faced with a similar situation, France established in 2016 a "*droit à l'oubli*" (translated literally as "right to be forgotten" in the remainder of this text), that is, the right for an insurance applicant not to declare a previous cancer after a period of 10 years starting at the end of the therapeutic protocol. This 10-year waiting period is reduced to 5 years if the applicant suffered cancer before the age of 18. These periods of 10 and 5 years start from the date of the end of the therapeutic treatment, in absence of relapse within this period. The 10-year length of the waiting period is further shortened for several types of cancer (and other non cancer related pathologies), as detailed in the reference grid used in France (known as convention AERAS, see `www.aeras-infos.fr/cms/sites/aeras/accueil.html`) with reduced duration after which survivors have access to the right to be forgotten. After this period, insurance companies cannot take the pathology into account in risk assessment and cannot refuse the insurance, nor impose a premium surcharge because of the pathology.

A similar rule has been introduced in Belgium in a law dated April 4, 2019 (published in the Official Journal on April 18, 2019) for home-related and professional mortgage insurance. Despite clear similarities, the right to be forgotten established in Belgium differs from its French counterpart in an important way. Cancer survivors must still declare their pathology to insurance companies when applying for mortgage insurance in Belgium but the decision to grant coverage cannot be based on this information. Besides this right to not declare (in France) or to declare but without consequences (in Belgium) a cancer after a given waiting period, the premium surcharges may also be prohibited or limited for some cancer types. This has been implemented in Belgium in a Royal decree published in the Belgian Official Journal on June 14, 2019.

Although the establishment of such a right to be forgotten in Belgium is clearly an improvement for cancer survivors, there is most probably room for further reducing the waiting period for some cancer types. Also, there remains some ambiguity about what is considered as treatment and thus what marks the end of the therapeutic protocol. Since all cancer cases (and date of diagnosis) must be recorded in a national database in Belgium and many other EU countries, defining the start of the waiting period at the recorded date of diagnosis would certainly avoid endless discussions

when a claim is filed.

There is abundant literature on cancer survival in biostatistical and medical studies. However, this topic has been the subject of few actuarial papers beyond those dealing with the assessment of extra mortality after Haberman and Renshaw (1990) and Renshaw (1988), such as Dodd et al. (2015). Let us briefly discuss some of the contributions on pricing life insurance specifically for cancer patients that appeared in the actuarial literature.

Lemaire et al. (2000) considered term life pricing in the presence of a family history of breast or ovarian cancer. These authors found that while many women with a family history of breast or ovarian cancer can be accepted at standard rates, women with two family members with cancer or one first-degree relative with cancer at an early age show substantial mortality increases (up to 100%) and can thus probably only be accepted at higher premium rates. Moreover, the authors also found that mortality increases for women with the BRCA1 or BRCA2 gene mutation (a malfunction which results in cells more likely to develop additional genetic alterations that can lead to cancer) reach 150% and can thus possibly only be accepted at a premium rate that incorporates a severe mortality surcharge.

Using 10-year and 20-year term life products, Shang (2019) calculated the single premium for breast cancer patients and found that both the average and minimum premium of the sample cancer patients to be much higher than standard premium, with the minimum premium still being close to 40% of the sum insured for the least risky patients. In order to improve the affordability of insurance products for cancer patients, Shang (2019) also suggested to set a waiting period during which no death claims will be paid. Since many deaths happen during the first years after diagnosis, a waiting period reduces the mortality risk and thus the net premium. For instance, for a 20-year term life product with sum insured 10,000 contracted by a 40-year-old cancer patient, a waiting period of 1, 2 and 3 years reduces the average premium by, respectively, 24.85%, 41.48% and 51.05% compared to no waiting period, according to the calculations by Shang (2019).

The present paper concentrates on the determination of the length of the waiting period embedded in the right to be forgotten, that is, the minimum duration before the applicant can be covered at standard premium rate. To this end, we apply several widely-accepted tools from biostatistics in order to assess excess mortality. We concentrate on melanoma (ICD-10 C43) and thyroid (ICD-10 C73.9) tumors for the sake of illustration, leaving other types of cancer for future research. These two types of cancer (called cancer sites) were selected to get a significant number of incidences occurring before the age of 40 (mortgage insurance applicants being rather young) and as they lead to a fraction of the patients who have a chance of survival close to cancer-free patients. We were thus looking for cancers with a relatively high survival rate or high cured rate. Our analysis is based on 28,994 cancer cases recorded by the Belgian Cancer Registry (BCR): 19,848 melanoma and 9,146 thyroid tumors, diagnosed between 2004 and 2016.

Based on survival data recorded by the Belgian Cancer Registry, the present paper aims

- to show that for some types of cancer (with melanoma and thyroid as examples), survivors actually have a survival comparable to that of the general population, that is, excess mortality is negligible.
- to demonstrate that patients having survived long enough to some types of cancer (still with melanoma and thyroid as examples) can access life insurance market at standard insurance rates, contrarily to the common belief within the actuarial community. The technical waiting period appears to be relatively short, and shorter compared to the 10-year period specified in the law.

In addition, we promote a waiting period starting at diagnosis rather than at the end of the therapeutic treatment protocol in order to avoid disputes in case of death. Indeed, diagnosis is recorded in databases maintained by official bodies within the European Union. The results obtained in the present study appear to be particularly encouraging as they suggest a considerable shortening of the 10-year waiting period for some types of cancer.

The remainder of this paper is structured as follows. Section 2.2 presents the data used to perform the present study. Sections 2.3 and 2.4 apply several tools from biostatistics to assess the length of the waiting period, with a focus on the estimation of the survival of cancer patients in Section 2.3 and concentrating rather on the time after which we can consider the patients still alive as "cured" in Section 2.4. Comparisons with standard premium rates based on life tables generally used on the Belgian market are provided in Section 2.5. The final section (Section 2.6) concludes the paper with a discussion.

2.2 Data sources

2.2.1 Belgian Cancer Registry (BCR)

The Belgian Cancer Registry (BCR) is a national population-based cancer registry collecting data on all new cancer diagnoses in Belgium since the incidence year 2004. For the execution of this main task, BCR relies on its own specific legislation.

In this study, we restrict our analysis to two cancer sites, as explained in the introductory section. We also limit our analyses to patients from 20 to 69 years old at time of diagnosis since the right to be forgotten mainly concerns young adults and active life. A total of 19,848 cases of melanoma and 9,146 cases of thyroid cancer diagnosed between 2004 and 2016 were followed-up until the 1^{st} of July 2018. Follow-up thus varied from 2 years for patients diagnosed in 2016 to 14 years for those diagnosed in 2004. Patients without a national security number (INSZ/NISS) were excluded from our analyses, as we have no vital status on these patients. Patients lost to follow-up (mostly due to moving abroad) and patients still alive at the end of the follow-up period were censored. Table 2.1 summarizes the percentage of lost to follow-up before the 1^{st} of July 2018 in each remaining subgroup together with the number of included cases. The fraction of patients lost to follow-up per subgroup varied from 1.04% for women with melanoma cancer aged 50-69 to 4.41% for male melanoma cancer patients aged 20-34. The total fraction of patients lost to follow-up cases, regardless of gender, site or age group was 1.87%. The analyses were conducted separately for women and men, representing 65.13% and 34.87% of all cases respectively, and age at diagnosis was included as a covariate, continuous or categorical depending on the approach. Age at diagnosis ranges from 20 to 69 years and 3 age groups were considered: 20-34, 35-49 and 50-69.

2.2.2 General population

In order to estimate excess cancer mortality, mortality in the cancer population must be compared to the expected mortality in the general population. Belgian population life tables, obtained from Statbel (the Belgian statistical office), were used to estimate expected mortality in the general population. Gender-specific mortality rates for the period 2004-2018 (by single year of age) have been smoothed in two dimensions to remove erratic variations. The surface smoothing was performed with the SAS procedure PROC LOESS using local linear polynomials weighted by population size (Cleveland et al., 1988; Cleveland and Grosse, 1991; Cleveland et al., 1992).

Gender	Cancer site	Age at diagnosis	Lost to follow-up	Number of cases included
Women	Melanoma	20-34	3.44%	1,863
		35-49	1.07%	4,386
		50-69	1.04%	5,786
	Thyroid	20-34	3.07%	1,204
		35-49	2.66%	2,596
		50-69	1.67%	3,048
Men	Melanoma	20-34	3.45%	725
		35-49	2.29%	2,360
		50-69	1.84%	4,728
	Thyroid	20-34	3.30%	273
		35-49	2.49%	724
		50-69	1.69%	1,301
Total			1.87%	28,994

Table 2.1: Numbers of melanoma and thyroid cancer cases diagnosed in Belgium between 2004 and 2016 (BCR data) by gender, site and age group, with percentage of lost to follow-up.

2.3 Survival of cancer patients

2.3.1 Overall survival

In this paper, we analyze survival (i.e., time to death) for cancer patients beyond diagnosis according to a number of covariates summarized into the vector $\boldsymbol{Z}$. Specifically, T denotes the remaining lifetime at diagnosis. Given $\boldsymbol{Z} = \boldsymbol{z}$, T has probability density function $f(\cdot|\boldsymbol{z})$, distribution function $F(\cdot|\boldsymbol{z})$, survival function $S(\cdot|\boldsymbol{z}) = 1 - F(\cdot|\boldsymbol{z})$, and hazard rate, or force of mortality $\lambda(\cdot|\boldsymbol{z}) = f(\cdot|\boldsymbol{z})/S(\cdot|\boldsymbol{z})$. Contrarily to insurance studies, T denotes the remaining lifetime since diagnosis and age at diagnosis is included in the covariates (attained age is thus obtained by summing age at diagnosis and survival time). This is why we refrain here from complying with the international actuarial notation for survival probabilities and force of mortality (when computing premiums in Section 2.5, we will revert back to the actuarial notation).

The nonparametric Kaplan-Meier (1958) estimator is used here to estimate the overall survival function $S(\cdot)$, without distinguishing according to causes of death (Belot et al., 2019). Estimated overall survival probabilities according to gender and site (melanoma and thyroid) are detailed in Table 2.2 and illustrated in Figure 2.1. The 10-year overall survival probabilities range from 0.774 (with 95% confidence interval $[0.762 - 0.786]$) for men with melanoma cancer to 0.93 (with 95% CI $[0.922 - 0.938]$) for women with thyroid cancer (see Table 2.2). Remember that overall survival probabilities take into account all causes of deaths, that is, both cancer and non-cancer related deaths are considered.

In addition to the two curves for women and men in Figures 2.1 and 2.2, a third curve including both sexes (denoted "Belgium") is also drawn for the simple reason that since December 21, 2012, the European directive on equality between men and women also applies to outstanding balance insurance, so the gender no longer has an influence on the premiums. Therefore, we believe it is of interest for insurance companies to visualize these survival probabilities for both sexes combined.

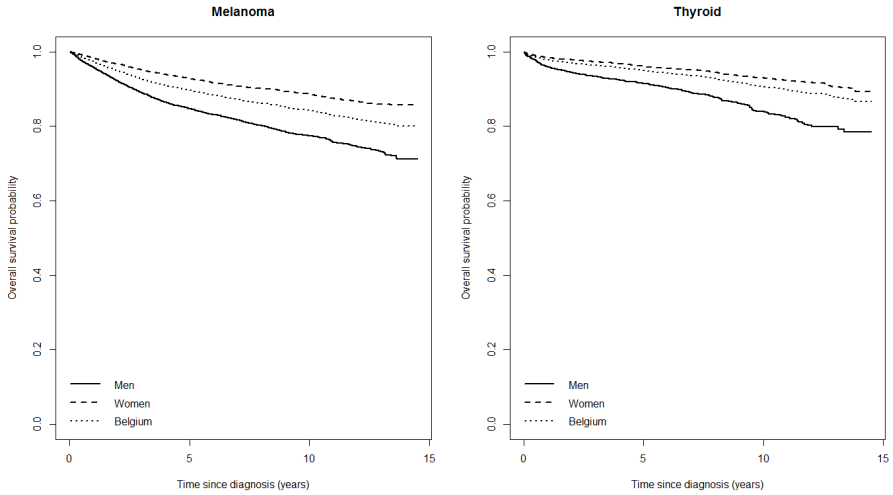


Figure 2.1: Estimated overall survival probability by gender and site using the nonparametric Kaplan-Meier (1958) estimator.

Sex	Cancer site	$\widehat{S}(t=5)$	$CI_{95\%}$	$\widehat{S}(t=10)$	$CI_{95\%}$
Women	Melanoma	0.928	(0.923-0.933)	0.885	(0.878-0.893)
	Thyroid	0.961	(0.956-0.966)	0.930	(0.922-0.938)
Men	Melanoma	0.847	(0.838-0.855)	0.774	(0.762-0.786)
	Thyroid	0.915	(0.903-0.927)	0.839	(0.820-0.860)

Table 2.2: Estimated overall survival probabilities by gender and site using the nonparametric Kaplan-Meier estimator at 5 and 10 years after diagnosis, $\widehat{S}(t=5)$ and $\widehat{S}(t=10)$, with their confidence interval (CI) at 95% level.

2.3.2 Relative survival

Information on cause of death is often unavailable or unreliable and not all deaths of cancer patients can be easily classified as a death due to the cancer of interest or due to another cause (Percy et al., 1981). Relative survival, which does not require information on the cause of death, provides a measure of the excess mortality experienced by cancer patients by comparing the mortality in the cancer population with the mortality in the general population. This led the relative survival to become the standard measure of patient survival for population-based cancer registries, as shown by its prominence around the world in studies related to cancer survival (Ries et al., 2002; Coleman et al., 1999; Berrino et al., 1999; Perme et al., 2012, 2016; Pavlič and Pohar Perme, 2019).

Relative survival models are divided into two types: (i) additive models and (ii) multiplicative models; see e.g. Pohar and Stare (2006). Despite the wide acceptance of multiplicative specifications within the actuarial community, it turns out that additive models are biologically more plausible in cancer studies and provide a better fit to the data (Dickman et al., 2004; Buckley, 1984; Hakulinen and Tenkanen, 1987; Esteve et al., 1990; Bolard et al., 2001). The additive specification is thus favored here. The hazard rate at time t since diagnosis for cancer patients with covariate vector $\boldsymbol{Z}$, is decomposed into two additive components: the population hazard based on available patient's characteristics $\boldsymbol{Z} = \boldsymbol{z}$, denoted as $\lambda_P(\cdot|\boldsymbol{z})$, and the excess hazard specific for the cancer of interest, denoted as $\lambda_E(\cdot|\boldsymbol{z})$. Formally,

$$\lambda(t|\boldsymbol{z}) = \lambda_P(t|\boldsymbol{z}) + \lambda_E(t|\boldsymbol{z}). \tag{2.1}$$

From this expression, relative survival model (2.1) can be written as

$$S(t|\boldsymbol{z}) = S_P(t|\boldsymbol{z})r(t|\boldsymbol{z})$$

where the relative survival function $r(\cdot|\boldsymbol{z})$ is defined as

$$r(t|\boldsymbol{z}) = \frac{S(t|\boldsymbol{z})}{S_P(t|\boldsymbol{z})}. \tag{2.2}$$

In words, the relative survival function $r(\cdot|\boldsymbol{z})$ corresponds to the ratio of the survival function of the studied group $S(\cdot|\boldsymbol{z})$ to the survival function of a comparable group (i.e., with the same characteristics) from the general population $S_P(\cdot|\boldsymbol{z})$ (Dickman et al., 2004).

In (2.1), $\lambda_P(\cdot|\boldsymbol{z})$ is usually estimated on the basis of external data such as population life tables, which are usually stratified according to the main factors affecting patient survival such as age, gender, and calendar year. As population life tables take into account all deaths, those due to the cancer of interest are thus also included. However, it is assumed that this does not influence the estimated ratio as mortality for a given cancer represents only a small fraction of the overall mortality. Correcting for this over-representation of the cancer being studied has, in practice, an insignificant effect on estimates of expected survival (Esteve et al., 1994). Oksanen (1998) shows that this holds even for common cancers such as prostate cancer.

Net survival is a measure of patient survival corrected for the effect of other causes of death (Dickman et al., 2004). It represents the (hypothetical) survival that would be observed if the only possible cause of death was the disease of interest (Berkson and Gage, 1950; Schaffar et al., 2017). If we are in a situation where we only have data from (a sample of) our sub-population of interest (e.g., cancer patients) but we have information on the cause of death, then estimation methods for net survival, sometimes referred then as marginal survival (see Geskus, 2015), can be estimated in a competing risks framework. However, unbiased estimate of this net or marginal (cause-specific) survival can only be obtained if one can assume independence of the

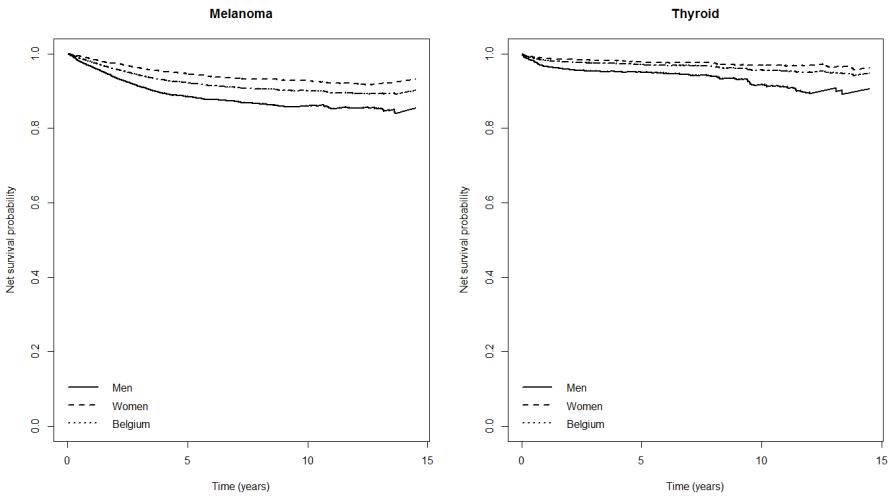


Figure 2.2: Net survival by gender and site using the nonparametric Perme et al. (2012) estimator.

censoring due to death from other causes (Latouche et al., 2013; Schaffar et al., 2017). On the other hand, net survival can also be estimated when the cause of death is unknown by making use of information from the general population. In this different framework, net survival can then be estimated using the relative survival method. Since the BCR does not collect information on the cause of death, we are in this latter situation. The net survival function is thus derived from the excess mortality hazard $\lambda_E(\cdot|\boldsymbol{z})$. Formally, it is defined as

$$S_n(t|\boldsymbol{z}) = \exp\left(-\int_0^t \lambda_E(u|\boldsymbol{z})du\right). \tag{2.3}$$

Depending on $\lambda_E(\cdot|\boldsymbol{z})$, $S_n(\cdot|\boldsymbol{z})$ may be a proper survival function but this is not necessarily the case.

There are several approaches to estimate net survival of a cohort of patients in a relative survival framework. Danieli et al. (2012) showed that only two of them provide unbiased estimates of net survival: (i) the nonparametric Perme et al. (2012) estimator and (ii) the excess risk based on an adjusted modeling on the demographic variables of the life tables. We used the nonparametric Perme et al. (2012) estimator to estimate net survival as recommended by Danieli et al. (2012) for population based studies.

Net survival probabilities by gender and site are illustrated in Figure 2.2, while net survival by age group and site is displayed in Figure 2.3. Numerical values are listed in Table 2.3. The net survival functions reach a plateau for both sites and genders, which is indicative for 'cure' of cancer. The estimated survival curves staying practically constant after 5 years since diagnosis indicates that the excess hazard of dying compared to the general population becomes negligible after only a few years after diagnosis. This suggests that a waiting period of moderate length would be enough to apply standard life insurance rates for the cancers under consideration.

Net survival by site and age group (both sexes combined, Figure 2.3) follows a quite expected path for melanoma cancer patients (left panel): younger patients have a better net survival compared to older patients. In particular, the 5-year and 10-year net survival probabilities for patients aged 20-34 were respectively 0.959

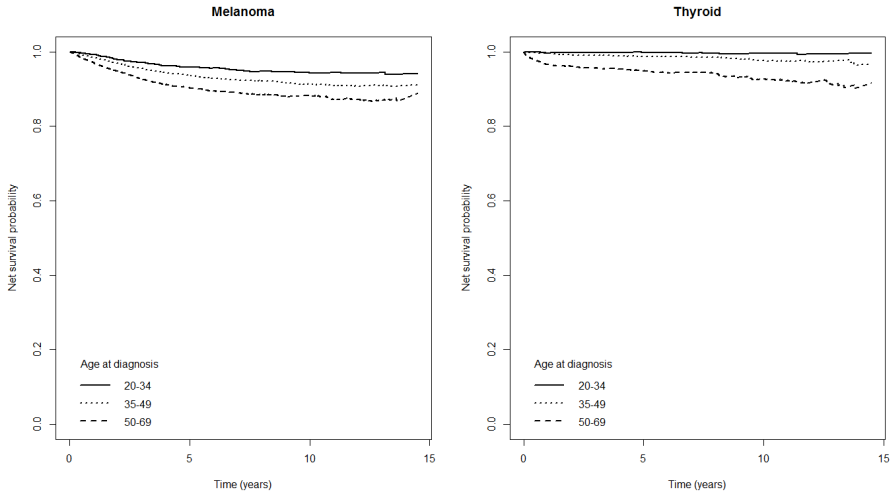


Figure 2.3: Net survival by age group and site using the nonparametric Perme et al. (2012) estimator.

Gender	Cancer site	Age at diag.	$\widehat{S}_n(t=5)$	$CI_{95\%}$	$\widehat{S}_n(t=10)$	$CI_{95\%}$
Women	Melanoma	20-34	0.974	(0.966-0.982)	0.963	(0.951-0.974)
		35-49	0.956	(0.949-0.963)	0.936	(0.925-0.946)
		50-69	0.927	(0.918-0.936)	0.907	(0.893-0.922)
	Thyroid	20-34	0.998	(0.995-1.000)	0.997	(0.990-1.000)
		35-49	0.992	(0.987-0.997)	0.990	(0.981-0.998)
		50-69	0.958	(0.948-0.968)	0.941	(0.924-0.959)
Men	Melanoma	20-34	0.921	(0.901-0.942)	0.894	(0.868-0.921)
		35-49	0.899	(0.886-0.913)	0.871	(0.853-0.890)
		50-69	0.871	(0.859-0.884)	0.849	(0.828-0.870)
	Thyroid	20-34	0.997	(0.987-1.010)	0.989	(0.968-1.010)
		35-49	0.972	(0.957-0.987)	0.932	(0.901-0.964)
		50-69	0.928	(0.908-0.949)	0.895	(0.858-0.933)

Table 2.3: Net survival probabilities by gender, site and age group using the nonparametric Perme et al. (2012) estimator at 5 and 10 years after diagnosis, $\widehat{S}_n(t=5)$ and $\widehat{S}_n(t=10)$, with their confidence intervals at 95% level.

(with 95% confidence interval $[0.951 - 0.968]$) and 0.943 (with 95% confidence interval $[0.932 - 0.955]$), meaning that for this age group (which is typically the age at which one starts a loan), patients' survival is close to that of the general population. Net survival by age group for thyroid cancer patients (right panel of Figure 2.3) yields even more promising results in the context of mortgage loans as the age group which is most likely to subscribe to such financial products (20-34 years) has a 5-year and 10-year net survival probabilities of respectively 0.998 (with 95% confidence interval $[0.995 - 1]$) and 0.996 (with 95% confidence interval $[0.989 - 1]$).

Net survival by gender, site and age group (Table 2.3) indicates a highly favorable outcome for women and men diagnosed with thyroid cancer and aged 20-34. For both subgroups, the 95% confidence interval for net survival at 10 years after diagnosis indicate that a net survival equal to that of the general population cannot be rejected at the 5% significance level.

Regarding the gap between women and men for the net survival curves (in particular for melanoma cancer patients), we do not have a clear explanation for these differences in survival probability. However, although stage distribution is very similar across gender, The Belgian Cancer Registry (2012), in agreement with Balch et al. (2001), showed that compared to men, females have more often melanoma on the arms or legs which has a better prognosis. This partly explains the difference in survival for melanoma cancer patients.

Using data provided by the French network of cancer registries (FRANCIM), Boussari et al. (2018) found very similar results for thyroid cancer patients diagnosed between 1995 and 2010. They obtained a 10-year net survival of 0.99 (95% CI [0.99-1.00]) for women aged 15-45, and 0.98 (95% CI [0.96-0.99]) for men of the same age group. Their results are also very close to ours when comparing patients aged 45-55: a 10-year net survival of 0.99 (95% CI [0.98-1.00]) for women, and 0.92 (95% CI [0.88-0.95]) for men. In another study including melanoma cancer patients diagnosed between 1989 and 2004 by the French registries, Jooste et al. (2013) obtained a 10-year net survival for men and women aged 15-45 of 0.81 (95% CI [0.78-0.84]) and 0.91 (95% CI [0.89-0.93]), respectively.

2.3.3 Proportional excess hazards

Previous sections suggested that some patients actually have a survival comparable to that of the general population. In the following sections, we confirm these findings from the point of view of the excess hazard.

Esteve et al. (1990) proposed a maximum likelihood method for estimating net survival via the modeling of the excess hazard. The excess hazard $\lambda_E(t|\boldsymbol{z})$ to be estimated is represented as

$$\log(\lambda_E(t|\boldsymbol{z})) = (\boldsymbol{\beta}^\top \boldsymbol{z}) + \log\left(\sum_{k=1}^{m} \tau_k I_k(t)\right) \tag{2.4}$$

where $\boldsymbol{\beta}$ is the log hazard ratio corresponding to the covariates, $I_k(t)$ the indicator function for the k^{th} interval (after splitting the follow-up time into short time intervals) and τ_k the net baseline hazard rate in that interval for patients with $\boldsymbol{z} = \boldsymbol{0}$.

A maximum likelihood approach to estimate parameters of model (2.4) is available in the `relsurv` package (Perme and Pavlič, 2018; Pohar and Stare, 2006) in R (R Core Team, 2017). Fitting two models to our data (one for each cancer site) with age group (50-69 years old taken as the reference category) and follow-up (with intervals of 5 years) as covariates, we obtain the results presented in Tables 2.4 and 2.5. The coefficient estimates $\widehat{\beta}$ and their standard errors are displayed in the second column,

Covariates	$\widehat{\beta}$ (S.E.)	p-value
Gender Women	-0.759 (0.060)	<0.001
Agegr 20-34	-0.741 (0.105)	<0.001
Agegr 35-49	-0.349 (0.064)	<0.001
fu [0,5)	-3.507 (0.045)	<0.001
fu [5,10)	-4.590 (0.105)	<0.001
fu [10,14]	-5.334 (0.420)	<0.001

Table 2.4: Results of model (2.4) fitted to melanoma cancer data.

Covariates	$\widehat{\beta}$ (S.E.)	p-value
Gender Women	-0.754 (0.169)	<0.001
Agegr 20-34	-2.864 (0.703)	<0.001
Agegr 35-49	-1.252 (0.213)	<0.001
fu [0,5)	-4.086 (0.129)	<0.001
fu [5,10)	-5.089 (0.314)	<0.001
fu [10,14]	-5.276 (0.867)	<0.001

Table 2.5: Results of model (2.4) fitted to thyroid cancer data.

p-values are reported in the last column. Note that the significance is largely impacted by the large sample size.

For both cancer sites, being a woman and being younger at the time of diagnosis are good prognostic factors.

2.3.4 Flexible parametric model

Model (2.4) is based on the assumption of proportional excess hazards, which constrains the hazard ratio to be constant over the follow-up time (Giorgi et al., 2005). Nonetheless, in cancer survival, the effects of prognostic factors often vary with time since diagnosis (Dickman et al., 2004; Quantin et al., 1999) and it is well known that the linearity assumption of covariates may be too strong and not always verified in practice (Mounier, 2015). As an example with the variable age, the effect on hazards of an increase of one year is often different for patients aged 18 or 60. This is why Remontet et al. (2019) and Fauvernier et al. (2019a,b) extended the model proposed by Esteve et al. (1990) to account for non-linear and non-proportional effects of covariates. This model also allows (i) a flexible modeling of the baseline hazard and (ii) a flexible interaction between several covariates adopting a multidimensional penalized splines approach. This leads to the specification

$$\log(\lambda_E(t|\boldsymbol{z})) = \sum_{j=1}^{J} g_j(t, \boldsymbol{z}) \tag{2.5}$$

where $g_j(\cdot,\cdot)$ are uni- or multidimensional penalized spline functions and each function $g_j(\cdot,\cdot)$ can be the marginal basis of time, the marginal basis of a covariate or the tensor product of the marginal bases of any number of elements of $(t, \boldsymbol{z})$ (Fauvernier et al., 2019b). This model has the advantage that the splines bring the flexibility needed for modeling the hazard and inclusion of penalty terms allows to control this flexibility for smooth estimation (as suggested by Eilers and Marx, 1996).

Several flexible models to estimate excess hazard were considered in this paper:

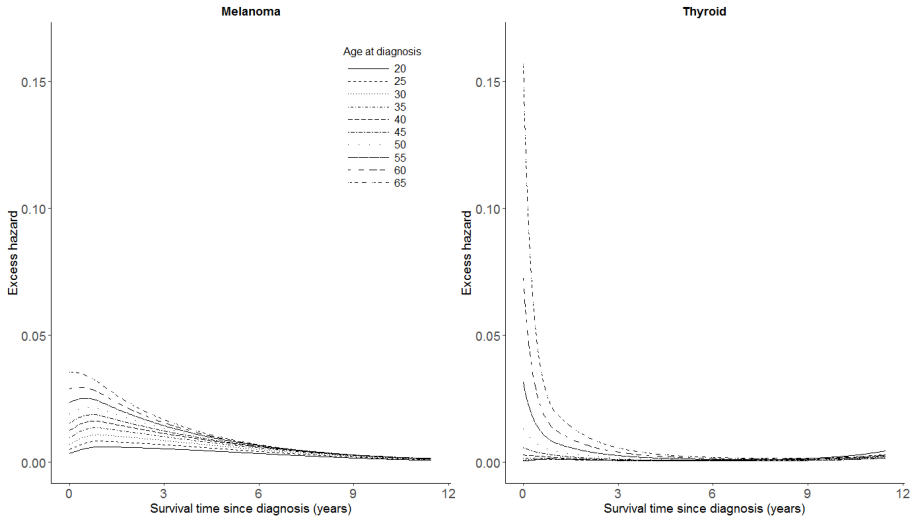


Figure 2.4: Excess hazard by age and cancer site estimated with a non-linear and non-proportional hazard model.

- baseline hazard only (BH) model:
  Eq. (2.4) without the $(\boldsymbol{\beta}^\top \boldsymbol{z})$ term, as it considers an excess hazard in each interval and no other covariates
- linear and proportional hazard (LPH) model:
  $\log(\lambda_E(t|\boldsymbol{z})) = f(t) + age$
- linear and non-proportional hazard (LNPH) model:
  $\log(\lambda_E(t|\boldsymbol{z})) = f(t) + age + g(t) \cdot age$
- non-linear and proportional hazard (NLPH) model:
  $\log(\lambda_E(t|\boldsymbol{z})) = f(t) + g(age)$
- non-linear and non-proportional hazard (NLNPH) model:
  $\log(\lambda_E(t|\boldsymbol{z})) = f(t) + g(age) + g(t) \cdot age$

with $f(t)$ the flexible parametric function for the baseline/reference hazard as a function of time and $g(\boldsymbol{z})$ the (non-)linear function of the covariates.

We fitted all these models and compared them based on a likelihood ratio test. The remaining of this section is based on the NLNPH model, deemed the best one according to likelihood ratio test. Excess hazard, assuming non-linear and non-proportional hazard for age at diagnosis, are estimated using the `flexrsurv` package (Clerc-Urmès et al., 2020) in R.

As previously suggested, for both cancer sites, excess mortality hazard increases with age at diagnosis and decreases with time since diagnosis (Figure 2.4). For thyroid cancer patients, excess hazard at the time of diagnosis is approximately 0.15 for patients aged 65 and is close to 0 for patients aged 40 and below. From 4 years to 10 years after diagnosis, excess hazard remains constant and is close to 0 for all ages. Beyond 10 years after diagnosis, excess hazard slightly bends up but remains small. This observation can also be an artefact of the spline function, which is "unbounded" at the end (a higher order degree for the spline function can produce this upwards bend). For melanoma cancer patients, excess hazard at the time of diagnosis is approximately 0.035 for patients aged 65 and is close to 0 for the youngest patients. For all ages, excess

hazard peaks at 1 year after diagnosis before decreasing until it becomes negligible, around 8 years after diagnosis.

2.3.5 Cure models

Cure models are a specific class of survival models which assume that a fraction of the subjects will never develop the event of interest, here death due to cancer. Such models have been used in different fields such as economics (e.g., time until an unemployed person finds a new job), engineering (e.g., time until a machine or device fails), finance (e.g., time until a bank goes bankrupt), marketing (e.g., time until a client buys a new product), for instance. In our context, cure models can be used to determine cancer patients who are considered as "long-term survivors" and those who are not (Maller and Zhou, 1996; Othus et al., 2012). The long-term survivors are still often referred to as "cured subjects" in the literature as a consequence of the name of this class of models. Cure models are particularly suitable for some cancer sites because if the treatment is successful, the patient will never suffer a relapse of the disease. It is also particularly suitable for less aggressive cancers such as, among others, the two considered in this paper because a considerable percentage of subjects exhibits long-term survival.

There are two main types of cure models in the literature: (i) mixture cure models (the most common type, based on the seminal work by Boag (1949) and Berkson and Gage (1952)) and (ii) non-mixture cure models (Andrei et al., 1996; Tsodikov et al., 2003; Chen et al., 1999); see Amico and Van Keilegom (2018) for a recent overview. In the present paper, we consider only the family of mixture cure models. In this approach, the patient population is considered as a mix of two types of patients, that is, long-term survivors who will never die of their cancer and the uncured patients who, if not censored, will die of their cancer (Lambert et al., 2006). In the global survival setting, the mixture cure model is specified as follows:

$$S(t) = \pi + (1 - \pi)S_u(t),$$

where π is the proportion of patients that are long-term survivors and $S_u(\cdot)$ the survival function of the uncured population. Both π and $S_u(\cdot)$ can then be modeled to depend on covariates. Cure models can be a useful alternative to standard survival models for cancers with a strong medical evidence and a confirmation in the data for the presence of long-term survivors (Legrand and Bertrand, 2019). In the relative survival setting, cure models also allow to determine the proportion of statistically cured cases and survival time of the fatal cases (Silversmit et al., 2017a).

The estimated proportion of cured cases and mean survival time of fatal cases for the two cancer sites considered (melanoma and thyroid), using a mixture cure model are given in Table 2.6. Note that NAs for young age groups are due to an insufficient number of cases.

The estimated cured proportion ranges from 84.99% ($CI_{95\%}$ [84.65, 85.33]) for men with melanoma cancer aged 50-69 to 99.86% ($CI_{95\%}$ [99.82, 99.90]) for women with thyroid cancer aged 20-34. The estimated mean survival time of fatal cases ranges from 0.47 years ($CI_{95\%}$ [-0.18, 1.12]) for women with thyroid cancer aged 20-34 to 10.57 years ($CI_{95\%}$ [1.32, 19.81]) for men with thyroid cancer aged 35-49. Moreover, higher age at diagnosis is correlated with lower cured proportions, except for men with thyroid cancer from the oldest age group.

Using data on 818,902 Italian cancer patients diagnosed between 1985 and 2005, Dal Maso et al. (2014) also found encouraging results for thyroid cancer patients, with an estimated cured fraction of 99% and 95% for women and men aged 15-45, respectively. Regarding melanoma cancer patients, the results on Italian data are somewhat lower than the estimations presented in this work, with a cured proportion

Gender	Cancer site	Age at diag.	Est. cured fraction	$CI_{95\%}$	Est. mean T	$CI_{95\%}$
Women	Melanoma	20-34	96.53	(96.33-96.72)	3.81	(3.46-4.22)
		35-49	93.28	(92.94-93.63)	4.49	(4.04-5.03)
		50-69	90.01	(89.49-90.53)	3.93	(3.50-4.46)
	Thyroid	20-34	99.86	(99.82-99.90)	0.47	(-0.18-1.12)
		35-49	99.03	(98.82-99.23)	4.87	(3.03-6.71)
		50-69	95.73	(95.39-96.07)	1.40	(1.02-1.78)
Men	Melanoma	20-34	88.77	(88.21-89.32)	3.54	(3.12-4.07)
		35-49	86.28	(85.77-86.78)	3.83	(3.52-4.20)
		50-69	84.99	(84.65-85.33)	2.94	(2.79-3.10)
	Thyroid	20-34	NA	NA	NA	NA
		35-49	91.96	(86.99-96.93)	10.57	(1.32-19.81)
		50-69	92.87	(92.34-93.41)	0.96	(0.67-1.25)

Table 2.6: Estimated cured fractions (in %) and mean survival time (in year) for the fatal cases by gender, site and age group, with their 95% confidence intervals. *Note: Est. mean T = Estimated mean survival time of fatal cases. NAs for young age groups are due to an insufficient number of cases.*

of 85% and 77% for women and men aged 15-45, respectively. The fact that we are using a more recent incidence period may partly explain these improved results.

2.4 Time-to-cure

The time-to-cure (TTC) is generally referred as the time after which patients can be considered as long-term survivors. Different approaches to define TTC have emerged in the literature. We highlight three of them. Firstly, Chauvenet et al. (2009) defined TTC_C as the time at which "almost" (that is, $1-\epsilon$, with ϵ small enough usually ranging from 0.1 to 0.01) all uncured patients would have died. From that time onwards, the number of deaths attributable to the cancer of interest becomes negligible.

Secondly, Dal Maso et al. (2014) define TTC_D as the shortest time after diagnosis at which the 5-year conditional net survival (defined as the ratio between net survival at time $t+5$ years and net survival at time t) is close to 1.

Thirdly, Boussari et al. (2018) defined TTC_B as the shortest time from which the conditional probability of being cured at a given time t after diagnosis knowing that the patient was alive up to time t is close to 1, that is TTC_B is the smallest value of t such that, for some given (small) value of ϵ

$$\frac{\pi}{S(t)} = \frac{\pi}{\pi + (1-\pi)S_u(t)} \geq 1-\epsilon, \tag{2.6}$$

where π is the proportion of cured patients and is estimated from the relative survival with the hypothesis of cure.

The main advantage of TTC is that, although the results may depend on the chosen definition, it is a simple and straightforward indicator to set the time after which a patient who had cancer should not be penalized anymore when applying for mortgage insurance.

In this paper, we focus on TTC_B for several reasons. First, TTC_B depends on both the cure proportion and the survival of the uncured, so it is less influenced by high early excess mortality. Second, TTC_B has the advantage of being an increasing function of time, therefore, it is not sensitive to a temporary plateau effect (that is, net survival curve flattening before decreasing again). Third, estimating TTC_B requires a

Gender	Cancer site	Age at diag.	$\widehat{\text{TTC}}_B$ $\epsilon = 0.05$	$CI_{95\%}$ $\epsilon = 0.05$	$\widehat{\text{TTC}}_B$ $\epsilon = 0.01$	$CI_{95\%}$ $\epsilon = 0.01$	Est. cure prop. (%)
Women	Melanoma	20-34	0.01	(0.00-1.18)	4.32	(3.30-5.35)	96.75
		35-49	0.81	(0.21-1.42)	5.65	(4.82-6.49)	94.16
		50-69	2.56	(2.04-3.08)	6.39	(5.66-7.12)	90.87
	Thyroid	20-34	0.01	(0.00-23.42)	0.01	(0.00-23.42)	99.75
		35-49	0.01	(0.00-14.45)	0.01	(0.00-14.45)	99.35
		50-69	0.01	(0.00-5.10)	4.82	(3.78-5.86)	94.93
Men	Melanoma	20-34	2.38	(1.92-2.85)	5.64	(4.90-6.39)	89.69
		35-49	2.78	(2.33-3.23)	5.86	(5.13-6.59)	88.12
		50-69	3.42	(3.00-3.85)	6.21	(5.51-6.90)	84.77
	Thyroid	20-34	0.01	(0.00-28.97)	0.11	(0.00-0.78)	98.88
		35-49	0.01	(0.00-16.03)	4.66	(3.19-6.13)	96.46
		50-69	0.91	(0.51-1.31)	7.14	(6.40-7.87)	90.84

Table 2.7: Estimated value of time-to-cure ($\widehat{\text{TTC}}_B$ in years, with $\epsilon = 0.05$ and $\epsilon = 0.01$) together with 95% confidence intervals and cure proportion (in %) by cancer site, sex and age group.

5-year shorter follow-up than TTC_D, which is a clear advantage of TTC_B over TTC_D (Boussari et al., 2018).

Time-to-cure TTC_B was estimated using the `rstpm2` package (Clements and Liu, 2019) in R. Table 2.7 presents estimated TTC_B (with $\epsilon = 0.05$ and $\epsilon = 0.01$) in years and the cure proportion in percentage for each subgroup.

In the following we interpret only TTC_B with $\epsilon = 0.05$. Results of TTC_B with $\epsilon = 0.01$ are still presented for the sake of comparison and completeness. From Table 2.7 we see that while TTC_B increases with age, ranging from a few days (0.01 year) for the youngest age groups to almost 3.5 years for the oldest age group of male melanoma cancer patients, cure proportion decreases with age, ranging from 99.75% for women aged 20-34 with thyroid cancer to 84.77% for men aged 50-69 with melanoma cancer. With this approach, the subgroups that stand out the most are the ones with a small TTC_B (and confidence intervals as narrow as possible to decrease uncertainty as much as possible) and a large proportion of long-term survivors. Among all subgroups considered, this is the case especially for women with melanoma cancer aged 20-34 ($\widehat{\text{TTC}}_B = 0.01$ with 95% confidence interval $[0.00-1.18]$ and estimated cure proportion = 96.75%).

Figure 2.5 illustrates the cure proportion obtained from a cure model and the TTC_B ($\epsilon = 0.05$) by age group and gender for both cancer sites. We can easily classify points into two clusters; the ones in the upper left corner with the best possible outcomes in terms of TTC_B and cure proportion and the others. Among melanoma cancer patients, women aged under 50 belong to this group with favorable outcomes, whereas for thyroid cancer patients, women of all ages and men aged under 50 belong to this group.

Similar results for thyroid cancer patients have been obtained by Boussari et al. (2018) on FRANCIM data. They also found that women aged 15-65 and men aged 15-45 have highly favorable outcomes, that is, a cure proportion close to 100% and a TTC_B close to 0.

2.5 Application to mortgage insurance

All results obtained so far suggest that, for melanoma and thyroid cancer patients, excess mortality becomes negligible after some waiting period. In this section, we determine the length of such a waiting period as the time needed to get back to standard premium rates. Henceforth, standard rates correspond to premiums computed

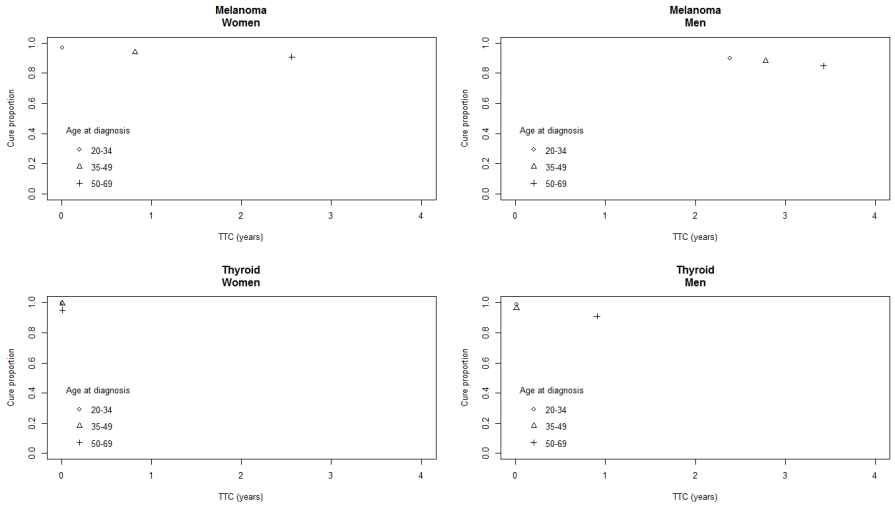


Figure 2.5: Cure proportion versus time-to-cure (in years, with $\epsilon = 0.05$) by gender and age group at diagnosis in patients diagnosed with melanoma and thyroid cancer.

according to life tables commonly used on the Belgian market:

- regulatory life table XK applying to insurance products comprising benefits in case of death (formally, XK defines minimum premium amount for policies with a positive sum at risk). This life table is conservative and generates a relatively high safety loading.
- experience market life table published by the National Bank of Belgium (NBB). These life tables reflect the mortality observed on the market, within portfolios of companies controlled by NBB. There is no safety loading and insurers are only allowed to apply premium rates resulting from NBB tables for relatively short periods of time (rates are subject to revision in case the observed mortality on the market changes over time).

We also include premium rates calculated from general population life table published by Statbel, for the sake of comparison. Term life premiums are smaller for NBB life tables and larger for XK life table, Statbel life tables falling in between because of the higher socio-economic profile of the insured population. Premium rates for cancer patients are computed from the excess mortality hazards estimated with the help of the flexible parametric model discussed in Section 2.3.4, according to the time elapsed since diagnosis.

Consider a mortgage insurance applicant aged x borrowing an amount of 100,000 at interest rate 2% for a duration δ. At time t, the amount of the loan that has not been amortized is denoted as $c(t)$. This loan is secured by mortgage insurance, repaying the lender the amount $c(t)$ in case the policyholder dies at time t. The expected present value (EPV) of benefits paid in case of death is thus equal to

$$\text{EPV} = \int_0^{\delta} c(t) v(0,t) {}_t p_x \mu_{x+t} dt \tag{2.7}$$

where $c(t)$ is the amount of benefit in case of death at time t (which is in our case, a decreasing sum insured corresponding to the amount of the loan not yet amortized), $v(0,t)$ is the present value at time 0 of a unit payment made at time t (the discount

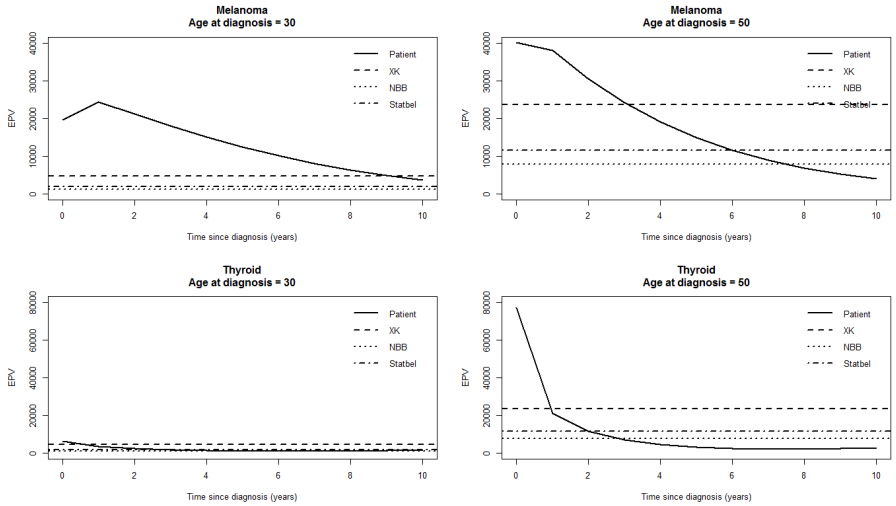


Figure 2.6: Expected present value (EPV) of a life insurance contracted by a 30 and 50-year-old cancer patient for a period of 20 years with interest of 1 percent and benefit of 100,000. Horizontal lines correspond to EPV calculated with XK, NBB and Statbel life tables.

factor), ${}_tp_x$ is the t-year survival probability for a policyholder aged x and μ_{x+t} is the force of mortality at attained age $x+t$.

In case the applicant suffered from cancer, we have to relate ${}_tp_x$ and μ_{x+t} entering the formula to the survival function and hazard rate estimated in the preceding sections. To this end, we assume that the applicant aged x has been diagnosed with cancer at age $x-w$. We then have

$$\begin{aligned} {}_tp_x &= \frac{S(w+t|\text{age at diagnosis} = x-w)}{S(w|\text{age at diagnosis} = x-w)} \\ \mu_{x+t} &= \lambda(w+t|\text{age at diagnosis} = x-w). \end{aligned}$$

with $S(\cdot)$ and $\lambda(\cdot)$ being estimated from the NLNPH model (eq (2.5)) discussed in Section 2.3.4.

This allows us to compute EPV(w) according to (2.7) for each candidate waiting period w after diagnosis and to select the smallest w such that EPV(w) becomes close to the value computed from the XK life table (assuming that cancer patients are priced with the regulatory life table XK and that the market can absorb the extra mortality burden corresponding to the difference between XK and NBB life tables).

Consider a home loan of duration 20 years. A cancer patient aged 30 and another aged 50 apply for mortgage insurance, with technical interest rate of 1% and a term of 20 years. These characteristics have been chosen as they represent a rather standard setting and other scenarios revealed similar patterns when considering different ages between 18 and 50. EPV based on XK, NBB and Statbel life tables have been computed to compare it with EPV(w) for a cancer patient diagnosed with melanoma or thyroid cancer at ages $30-w$ and $50-w$. Results are illustrated in Figure 2.6. Notice that EPV are presented for both genders combined, as XK life table applies to both sexes and insurance companies operating in the European Union are not allowed to account for gender in pricing.

Results show that EPV of a 30-year-old patient approaches EPV based on XK life table about 9 years after diagnosis for melanoma cancer patients and after slightly

less than 1 year for thyroid cancer patients. For a 30-year-old patient with melanoma cancer, the EPV never reaches within our time period the EPV based on NBB and Statbel life tables. For a 30-year-old patient with thyroid cancer, the EPV goes below the lowest EPV (based on NBB life table) as early as about 4 years after diagnosis. For a 50-year-old patient with melanoma cancer, EPV reaches the same level than the one based on XK life table 3 years after diagnosis, and reaches the lowest level (based on NBB life table) less than 8 years after diagnosis. Finally, for a 50-year-old patient with thyroid cancer, EPV reaches XK level 1 year after diagnosis, then stays below the lowest EPV 3 years after diagnosis. This is in line with the reduced waiting periods published in the Royal decree on June 14, 2019, where a duration of 3 years applies in this case. There is however a fundamental difference in the approach because the waiting period starts at diagnosis in the present study whereas it starts at the end of the treatment protocol according to the law.

The improvement for melanoma cancer patients is not as substantial as for thyroid cancer patients since the time period is close to 9 and 3 years for 30 and 50-year-old patients, respectively. Nonetheless, it remains advantageous for patients since the 9 and 3-year period starts from the date of diagnosis and not at the end of the therapeutic treatment. Note that, at first glance there seems to be an advantage for older patients with shorter waiting time, while all the other results seem rather to indicate an advantage for younger patients with shorter time to cure and higher cure rate. This actually comes from the fact that in absolute terms, younger patients have shorter time to cure and thus lower EPV than older patients. As shown in Figure 2.6, EPV for a 30-year-old patient is lower than for a 50-year-old patient regardless of the time since diagnosis and for both cancer sites. However, financial burdens (i.e., XK, NBB and Statbel levels) are much lower for younger people than for older people (since young people from the general population have a lower probability of dying than older people).

2.6 Discussion

Results derived in this paper are in line with the reduced waiting period specified in the Belgian legislation. Furthermore, results are also in line with the reference grid used in France (convention AERAS) as the time after which patients have access to the right to be forgotten according to this convention is relatively short (maximum 6 years after the end of the therapeutic protocol for the two cancers considered in this paper).

All analyzes in this paper are based on the time since diagnosis although the right to be forgotten implemented in Belgium and France is applicable after a certain time after the end of the therapeutic protocol. For the sake of clarity and easiness, an approach based on the time since diagnosis would undeniably be more favorable to patients. A right to be forgotten based on the date of diagnosis would indeed allow patients to know when exactly they can expect to benefit from this right. On the contrary, with the current approach based on treatment end date (which is unknown until the success of the treatment and can even change later in case of relapse), benefiting from this right is subject to a high level of uncertainty as durations of treatments are heterogeneous and unpredictable even within same cancer types and stages. Therefore, patients cannot currently easily estimate when (and if) they will be able to benefit from this right.

One of the main results of this paper is thus to promote the use of the date of diagnostic instead of the end of the therapeutic treatment for defining the waiting period. However, one could argue that the length of medical treatments may have significantly reduced over the period 2004-2018 and if this is the case, both approaches could be now closer than expected. While it could have been interesting to formally

compare both approaches, individual data on the type and length of treatment for each case is not reported in the Belgian Cancer Registry and such information is not readily available. Furthermore, the definition of the end of the treatment is in itself debatable and the duration of the treatment can be quite different depending on several factors, which is actually an argument in favor of considering the date of diagnosis. Moreover, durations of treatment are heterogeneous even within the same cancer type, usually unpredictable, and optimal durations are often still open to debates (Schvartsman et al., 2019). In any case, a reduction in treatment length due to the progress made in medical treatment of cancer would obviously lead to closer agreement between the two approaches. Since the date of diagnosis, as recorded in national registries, offers the advantage to not be subject to any discussion and to allow the patient to know from the start when the waiting period will end, we think that all parties (actuaries included) will benefit from using the date of diagnosis instead of the end of treatment for more convenience and less uncertainty.

Contrarily to other studies (like Yue et al., 2018), calendar time has not been included in the analysis conducted in the present paper, because of the limited amount of cases available (recall that we concentrate on younger ages because of the product under consideration, targeting young adults). That being said, we performed two subanalyses: one for the cohort 2004-2011 and one for the cohort 2012-2018 and compared the results. There are no real changes, so we conclude that there is no cohort effect. Moreover, simplicity of the system is crucial and given that medical treatments keep improving, the resulting bias of ignoring a potential cohort effect favors insurance providers. Notice that the approaches are dependent on factors such as changing diagnostic criteria and improved diagnostic methods. As these factors may vary over time irrespective of any improvement of the treatment and are different between populations, one can not compare excess risks across different time periods and populations (Lenner, 1990). To illustrate this, suppose that a medical advance allows a cancer to be diagnosed at a less severe stage (cases that are not as fatal as the ones detected with the previous methods) and perhaps also earlier with the consequence that more cases are detected (cases that would not have been detected with the previous detection methods are now detectable). These improvements will yield an increased survival rate, regardless of whether the treatment improved or not. This weakness of the survival rate has been pointed out in the literature extensively (Enstrom and Austin, 1977; Bailar III and Smith, 1986, among others). However, although not necessarily the case, earlier diagnoses will in most cases be associated with better efficacy of the treatment.

The melanoma and thyroid cancers may include a variety of types and could be diagnosed at different stages of severity. Moreover, significant gaps are observed between women and men. It is undeniable that including the information on stages of severity and gender in the NLNPH model would refine the analysis. However, these have been ignored on purpose, considering the specific application of the results for insurance practice. Since December 21, 2012, the European directive on equality between men and women also applies to outstanding balance insurance. The judgment of the European Court of Justice indeed stipulates that no discrimination can be made between men and women when establishing such insurance contract, so the gender no longer has an influence on the premiums nor on the coverage conditions of outstanding balance insurances. Therefore, the gender has been omitted not only for the sake of simplicity, but also because it is illegal for insurance companies to use that information. Concerning the stage of severity, it has also been omitted to ensure the simplicity and thus the legal safety of the system. Moreover, not including the stage may bias the analysis but not necessarily in favor of patients diagnosed at an advanced stage (perhaps a shorter waiting period would be indicated for them). In any case, if the method amounts to determining the time to wait before mortality returns

to normal, ignoring the stage is actually a conservative approach for the insurer.

To ensure that the coverage cost of cancer patients remains acceptable for the insurance industry, further constrains may be imposed in terms of sum insured, for instance. Also, it could be reasonable to impose that insurers charge a single premium for mortgage insurance to mitigate mortality risk. Last but not least, compensation could be performed at market level to avoid that some insurers face higher costs. Indeed, even if the coverage of mortality risk becomes affordable after a relatively short waiting period, premium rates reflecting the actual mortality of cancer survivors remain sometimes above the market premiums resulting from NBB life tables. This means that when it is the case, this extra cost must be fairly distributed among stakeholders: cancer patients, insurance industry, banking sector (as they sell the loans) and society as a whole.

Notice that cancer stage at diagnosis has not been taken into account in the present study. For patients with more advanced cancer stage, the waiting period will be more conservative because mortality will peak just after diagnosis and before reverting back to the general population level. Note also that *in situ* cancer cases (considered as pre-cancer) have not been included in the present study as they are not classified as cancer *per se* (Chang et al., 2007) like any other "regular" cancer cases which were not diagnosed as *in situ* before. Moreover, cancer is not one disease, but a family of many diverse diseases with different outcomes. Results in the present paper focus on melanoma and thyroid cancer patients, and cannot be applied to other cancer types. A natural extension of this work would be to repeat the analyses for all major cancer types. This would certainly be useful for implementing appropriate market rules but goes beyond the scope of this study which primarily aims to advocate a waiting period starting at diagnosis.

Cancer patient survival has improved over the last few decades, with an increasing proportion of patients being cured for many types of cancer (Andersson et al., 2011; Lambert et al., 2006). Providing coverage in case cancer is diagnosed or to long-term cancer survivors is therefore of prime importance, for the society but also for the insurance industry since proper coverage of such risks may well produce attractive returns.

Semi-Markov modeling for cancer insurance

3

This chapter corresponds, notwithstanding a few minor improvements, to an article carrying the same name as the chapter and published jointly with Pr. Catherine Legrand, Pr. Michel Denuit and Dr. Geert Silversmit in the *European Actuarial Journal* in 2022.

Abstract
Advancements in medicine and biostatistics have already resulted in a better access to insurance for people diagnosed with cancer. This materializes into the "right to be forgotten" adopted in several EU member states, granting access to insurance after a waiting period of at most 10 years starting at the end of the successful therapeutic protocol. This paper concentrates on insurance covers on a market where such a right has been implemented. Stand-alone products are considered, as well as guarantees included as a rider in an existing package. The cost of offering standard premium rates to all applicants in mortgage insurance related to property loans is also evaluated. The 3-state (healthy–ill–dead) Semi-Markov hierarchical model developed in Denuit et al. (2019) for long-term care insurance is adopted here for actuarial calculations. Semi-Markov transition intensities are estimated from cancer cases recorded by the Belgian Cancer Registry. The obtained results suggest that a new offer could develop, targeting the particular needs of cancer patients.

Keywords: Critical illness, medical insurance, right to be forgotten, multi-state models.

3.1 Introduction and motivation

Massart (2018) testimonial illustrates difficulties faced by cancer survivors to access life and health insurance products. See also Hendriks et al. (2021) for the particular case of childhood cancer survivors. However, progress in medicine over the last 20 years greatly improved the prognosis of several types of cancer. In parallel, many tools have been developed in biostatistics and epidemiology to study mortality and morbidity associated with cancer. These advances have already resulted in a better access to insurance products for people diagnosed with cancer. For example, this led France and Belgium to establish a "*droit à l'oubli*" (translated literally as "right to be forgotten" in the remainder of this text) granting access to insurance after a waiting period of at most 10 years starting at the end of the successful therapeutic protocol, in absence of relapse within this period. The 10-year period has even been shortened for several types of cancer with a good prognosis. We refer the reader to Soetewey et al. (2021) for an actuarial analysis of the "right to be forgotten" mechanism. Several initiatives purpose to extend this right throughout EU countries, by law or through a convention with insurance sector (as in Luxembourg). We refer the reader e.g. to Scocca and Meunier (2020) as well as to "Survivorship challenge 3.4: Lack of knowledge

of the stigma associated with cancer" listed in Lawler et al. (2021) purposing to take advantage of the existing legal framework in four EU member countries (France, Belgium, Luxembourg and the Netherlands) to investigate a pan-European legal framework on access to financial services for cancer survivors.

Although the establishment of such a "right to be forgotten" is a clear progress for cancer survivors, there is still room for improvement by filling the coverage gap during the waiting period opening this right. Social security and supplemental health insurance cover most medical costs related to cancer, but non-medical costs are usually paid out of pocket. The latter include lost income (the part not covered by Social security or supplemental disability insurance), travel to and from hospital (especially for patients living in rural areas), travel and family lodging expenses, deductibles and co-payments, non-conventional comfort treatments, private nursing care costs, and non-nursing help with activities of daily living. They can rapidly become a financial hardship, even for patients who are treated on an outpatient basis.

Some insurance companies market supplemental cancer insurance policies that pay lump sum benefits upon diagnosis of cancer or a temporary life annuity to face these out-of-pocket expenditures. These products developed in Asia, North America and UK (Bennett et al., 1998; Nielsen and Mayer, 2000). Using Taiwan National Health Insurance Database, Yue et al. (2018) priced two types of whole-life insurance products: (i) a (lump-sum) benefit paid when the insured is diagnosed with cancer for the first time (and the contract terminates after the benefit is paid) and (ii) an annual benefit paid after the insured is diagnosed with cancer and as long as he or she survives. Shang (2019) considered term life products and evaluated the extra cost related to cancer. Let us also mention that cancer is also typically included into the diseases covered by critical illness insurance policies.

The products considered in this paper are specifically related to the waiting period opening the "right to be forgotten", with temporary covers restricted to that period to fill the gap in coverage on a market where such a right has been implemented. First, stand-alone products are studied, including cancer insurance with lump sum payment at diagnosis, or temporary life annuity starting at diagnosis. In the latter case, periodic payments may correspond to insurance premiums of another product, or even to loan reimbursement. Then, riders included in a package are discussed. We consider term insurance with accelerated death benefit paid as a lump sum at diagnosis or as a temporary life annuity starting at diagnosis. The payment of rider reduces death benefit specified in term insurance. Finally, we discuss products granting access to some specific insurance cover (such as mortgage insurance) during the waiting period opening the "right to be forgotten". This is especially important at young age, to guarantee access to property and home ownership (in case of house loan) and to entrepreneurship (in case of professional loan) to cancer patients whose health status has improved but who cannot benefit from the "right to be forgotten" because the waiting period is not exhausted. This guarantee can be bought by parents for their children (as it is commonly the case in medical insurance, where extra premiums ensure that children can continue their parents' medical cover when they leave the household, whatever their future health status) or by young adults starting their professional career (in supplement to individual health insurance, for instance). The cover may be subject to some deferred period after diagnosis in order to lower its cost (without real impact since it is very unlikely that cancer patients consider buying a house or developing professional activities right after diagnosis). The price of this cover also allows to evaluate the cost of offering standard premium rates to all applicants in mortgage insurance related to property loans. This is the approach followed by Crédit Mutuel in France, which announced in November 2021 the end of health questionnaire when applying for a home loan (under some conditions, like age below 62, amount borrowed less than 500,000 euros, having Crédit Mutuel as

main bank since 7 years at least, this duration serving as an implicit waiting period as in the "right to be forgotten" mechanism, and electing domicile in the house to be bought). The bank even announces that this decision will have a retroactive effect for the ongoing loans. The loss in mortgage insurance premiums is evaluated to 70 millions euros per year for Crédit Mutuel. Our calculation help to assess the cost of this decision.

Cancer typically belongs to the set of critical illnesses and is thus also covered under several standard insurance policies. Long-term care or disability insurance policies also pay benefits in case of cancer. However, depending on the terms of the contract, a cancer patient may well have completed his or her treatment before the deferred period required to receive insurance benefits has been reached. Some patients even continue working while receiving chemotherapy or radiation therapy and thus do not qualify for loss of income. The cover comprised in the products considered in this paper is activated at diagnosis. These policies are targeted to help patients facing out-of-pocket expenditures while treatment occurs or to grant them access to other insurance products during the waiting period opening the "right to be forgotten". They thus constitute a new offer on a market where such a right has been implemented (as it is the case in Belgium or France, and may be generalized at EU level).

To illustrate our proposals, we consider the Belgian market where the "right to be forgotten" has been inserted in the Insurance Law in April 2019, with reduced waiting periods for some cancer types defined by Royal Decree in May 2019. Calculations are based on data available from the Belgian Cancer Registry (BCR), a national population-based cancer registry collecting data on all new cancer diagnoses in Belgium since the incidence year 2004. The paper considers three cancers with clear differences in terms of incidence, survival after diagnosis, and waiting periods defined by Royal Decree: melanoma (ICD-10 C43) with waiting period reduced up to 1 year after the end of the successful therapeutic protocol, thyroid (ICD-10 C73) with waiting period reduced up to 3 to 6 years after the end of the successful therapeutic protocol, and female breast (ICD-10 C50) cancer subject to the standard 10-year waiting period. Melanoma and thyroid cancer patients are known to have limited excess mortality compared to the general population (Soetewey et al., 2021). This has been recognized by reducing the waiting period opening the "right to be forgotten" in the Royal decree published in the Belgian Official Journal on June 14, 2019. The situation for female breast cancer patients is different with usually a high survival probability in the first years after diagnosis before it eventually decreases due to late cancer recurrences.

In this paper we perform our actuarial calculations in a 3-state Semi-Markov model assuming that a policyholder can be either "healthy", "ill" (diagnosed with cancer), or "dead". For the sake of easiness, only transitions from healthy to ill, healthy to dead and ill to dead are allowed so that cancer is assumed to be permanent (i.e., no recovery is possible). This non-reversibility greatly simplifies the computations (as the 3-state process is hierarchical) and appears to be reasonable for cancer insurance considered in this paper, with temporary cover restricted to the waiting period opening the "right to be forgotten" and benefits paid after the first diagnosis. For premium calculations, if mortality in cancer state reverts to the standard level after a sufficiently long period (like in cure models) then this is equivalent to a transition back to the initial healthy state. This model has been considered by Denuit et al. (2019) for long-term care insurance. Let us mention that Dębicka et al. (2015) also considered multi-state models to combine reverse annuity contracts with critical illness (in fact cancer) insurance for retired people. The cancer state is splitted into several stages to capture duration effects while remaining with a Markov structure. Here, we consider younger ages and focus on the waiting period opening the "right to be forgotten", performing calculations in the 3-state Semi-Markov model.

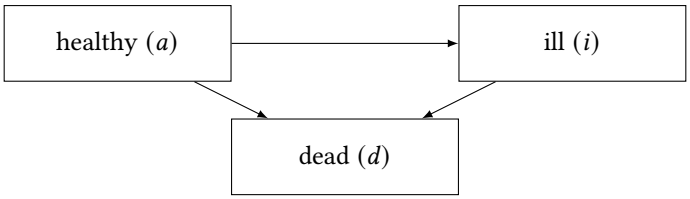


Figure 3.1: Semi-Markov 3-state model for cancer insurance.

The remainder of the paper is organized as follows. Section 2 describes the 3-state Semi-Markov model used for premium calculation. Cancer insurance products are described in Section 3, where corresponding premium rates are computed in the model of Section 2. The final Section 4 discusses the results and concludes.

3.2 Semi-Markov 3-state model

3.2.1 State space and transitions

Multi-state models offer a convenient representation for life and health insurance liabilities when benefits are associated to sojourns in, or transitions between different states (Dickson et al., 2013; Pitacco, 2014). We consider an individual aged x at policy issue, taken as time 0. His or her history is described by the stochastic process $\{X_t, t \geq 0\}$ where X_t gives the state occupied at time t. Here, t corresponds to contract seniority. In this paper, we consider that $X_t \in \{a, i, d\}$ where state a stands for "active" (healthy), state i stands for "ill" (cancer) and state d stands for "dead" as represented in Figure 3.1. At the time of diagnosis, individual moves from state a to state i. We do not allow for recovery (but mortality rates in state i may ultimately become similar to those applying in state a for cured individuals). Since benefits only relate to the first diagnosis, state i corresponds here to the first time a policyholder is diagnosed with cancer.

The time spent in state i is known to influence mortality so that we introduce the random variable Z_t, defined as the time spent in the state occupied at time t. Formally,

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

For an individual in state i at time t, Z_t is the time since diagnosis. Henceforth, we work under the Semi-Markov assumption: only the current state X_t and the time Z_t spent in the current state influence future transitions. This means that stochastic process $\{(X_t, Z_t), t \geq 0\}$ is a Markov process.

3.2.2 Transition intensities

Transition intensities quantify the instantaneous risk of making a given transition, depending on the state currently occupied and sojourn time. Assuming that Z_t only matters in state i, transition intensities are defined by the following limits:

$$\mu_{x+t}^{ai} = \lim_{h \to 0} \frac{\mathrm{P}[X_{t+h} = i | X_t = a]}{h}$$
$$\mu_{x+t}^{ad} = \lim_{h \to 0} \frac{\mathrm{P}[X_{t+h} = d | X_t = a]}{h}$$
$$\mu_{x+t;z}^{id} = \lim_{h \to 0} \frac{\mathrm{P}[X_{t+h} = d | X_t = i, Z_t = z]}{h}, z < t.$$

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

3.2.3 Data

3.2.3.1 Belgian Cancer Registry (BCR)

We consider the data available from the Belgian Cancer Registry (BCR), a national population-based cancer registry collecting data on all new cancer diagnoses in Belgium since the incidence year 2004. For the execution of this main task, the BCR relies on its own specific legislation (more information can be found on the BCR website, at `kankerregister.org`).

To illustrate our work, we restrict our analyses to three cancer types: melanoma (ICD-10 C43), thyroid (ICD-10 C73) and female breast (ICD-10 C50) cancer. These three cancer sites have been selected to evaluate the proposed insurance products in different scenarios. Melanoma and thyroid cancer patients are known to have a limited excess mortality compared to the general population. The situation for female breast cancer patients is different with usually a high survival probability in the first years after the date of diagnosis before it eventually decreases due to late cancer recurrences. We consider only female breast cancer as there are very few registrations regarding to male breast cancer. We also limit our analyses to patients aged 20 to 69 at diagnosis since the products considered in this paper target young adults and active life.

A total of 24,325 persons were diagnosed with melanoma, 10,789 with thyroid and 105,127 with breast cancer between 2004 and 2018, and were followed-up until April 1, 2020. Follow-up thus ranged from 2 to 16 years. Only one record per patient (with the earliest incidence date) within each cancer site was kept for patients with multiple primary diagnoses. This is in accordance with the insurance products under consideration which are activated at the time of the first diagnosis. A minority of patients without national security number were excluded from the analysis. Patients lost to follow-up (mostly due to moving abroad) and patients still alive at the end of the follow-up period were treated as censored observations.

Table 3.1 summarizes the number of included cases, number and proportion of deaths and percentage of lost to follow-up before April 1, 2020 per type of cancer, gender and age group. The fraction of patients lost to follow-up per subgroup varied from 1.06% for women with melanoma cancer aged 35-49 to 4.04% for male thyroid cancer patients aged 20-34. The total fraction of patients lost to follow-up cases, regardless of gender, site or age group was 1.4%.

3.2.3.2 General population

The products considered in this paper are sold to individuals before diagnosis (thus, in state a). This is in contrast with the study by Soetewey et al. (2021) which considered individuals in state i. Therefore, we also need mortality in the general population. Belgian population life tables are available from Statbel (the Belgian statistical office) and can be freely downloaded from the website `statbel.fgov.be`.

3.2.4 Estimation

Transition intensities are often assumed to be piecewise constant in order to ease actuarial calculations. Starting from state a, this means that the identities

Gender	Cancer site	Age at diagnosis	Lost to follow-up	Number of included cases	Number of deaths
Men	Melanoma	20-34	3.38%	857	83
		35-49	2.24%	2,812	354
		50-69	1.70%	6,000	1,268
	Thyroid	20-34	4.04%	322	6
		35-49	3.21%	841	58
		50-69	1.92%	1,563	297
Women	Melanoma	20-34	3.22%	2,174	70
		35-49	1.06%	5,267	317
		50-69	1.12%	7,215	874
	Thyroid	20-34	3.76%	1,435	10
		35-49	2.51%	3,029	78
		50-69	1.83%	3,599	390
	Breast	20-34	2.87%	2,685	423
		35-49	1.49%	29,007	3,419
		50-69	1.07%	73,435	12,730
Total			1.40%	140,241	20,377

Table 3.1: Number of persons diagnosed with melanoma, thyroid and female breast cancer in Belgium between 2004 and 2018 (BCR data) by gender, site and age group, together with the percentage of lost to follow-up and the number of deaths.

$$\mu^{ai}_{x+k+t} = \mu^{ai}_{x+k} \text{ and } \mu^{ad}_{x+k+t} = \mu^{ad}_{x+k} \tag{3.1}$$

hold for every integers x and k and fractional $0 \leq t < 1$. Transition intensity from state i is displayed as a function depending on the age at diagnosis and the time elapsed since diagnosis, i.e.

$$\mu^{id}_{x+\xi;z} = \widetilde{\mu}(x + \lfloor \xi - z \rfloor, \lfloor z \rfloor) \tag{3.2}$$

for some given function $\widetilde{\mu}$ with integer arguments, where $\lfloor \cdot \rfloor$ denotes rounding from below (i.e., the integer part). The arguments of $\widetilde{\mu}(\cdot, \cdot)$ represent, respectively, integer parts of age at entry in the ill state and time spent in that state. We thus work with age last birthday at diagnosis and the number of years since diagnosis, rounded from below. Of course, more accurate calculations can be performed by refining the time step, if needed.

When intensities are piecewise constant, they are easily estimated by the ratio of the observed number of transitions (diagnosis or death) to the corresponding exposure (in the state to be left). Precisely, consider a given integer age y and let N^{ai}_y be the number of transitions from state a to state i, that is, the number of diagnoses, among individuals aged y last birthday. Similarly, let N^{ad}_y be the number of transitions from state a to state d, that is, the number of deaths recorded among healthy individuals aged y last birthday. Let E^a_y denote the (central) exposure to risk in state a, that is, the time spent by all individuals aged y last birthday in state a. Because general population mortality statistics do not record exposures but only the number of individuals at the beginning of the period (or initial exposure to risk), we assume that transitions occur in the middle of the period. Under (3.1), the maximum likelihood estimators of μ^{ai}_y and μ^{ad}_y are respectively given by N^{ai}_y / E^a_y and N^{ad}_y / E^a_y and these values apply between ages y and $y+1$. Similar formulas hold true under (3.2) for estimating $\widetilde{\mu}(y, z)$ for integer values y and z, classifying transitions and recording exposures according to age last birthday y and integer number z of years since diagnosis.

Given that BCR covers the whole population, it would be possible to subtract from exposures and death counts available from Statbel the time lived and the number of deaths among cancer patients. In this way, the estimated μ_y^{ad} would only account for mortality not related to the cancer under consideration. However, in this paper, estimated intensities μ_y^{ad} have been obtained from the Belgian population life tables, so ignoring the influence of cancer mortality. The fact that population life tables include cancer mortality is not an issue as mortality for a given cancer represents only a small fraction of the overall mortality, and correcting for this over-representation of the cancer being studied has, in practice, an insignificant effect. The transition intensities from healthy to ill (cancer) and from ill to death can be estimated from BCR data (combined with general population exposures in the former case).

Even if the general shape of the mortality and incidence curves is generally clearly visible, erratic variations often remain. As long as these random departures do not reveal anything about the underlying mortality or morbidity pattern, they should be removed before entering actuarial calculations. This process is known as graduation in the actuarial literature. As there is no simple parametric model able to capture the structure of mortality and morbidity, actuaries generally use a generalized additive regression model with Poisson response distribution for transition counts, see e.g. Denuit and Legrand (2018). The method can be summarized as follows. Under (3.1), maximum likelihood inference can be equivalently conducted under the hypothesis that N_y^{ai} is Poisson distributed with mean $E_y^a \mu_y^{ai}$. The transition intensity μ_y^{ai} is then represented as $\ln \mu_y^{ai} = s(y)$ for some smooth function $s(\cdot)$ to be estimated from the data, under the assumption that N_y^{ai}, $y = 20, 21, \ldots, 69$ are mutually independent. The function $s(\cdot)$ is estimated with the help of a Poisson generalized additive model and the resulting estimate is used to produce the transition rates adopted to perform all actuarial calculations in the remainder of this paper. A similar procedure is followed to produce the other transition intensities entering the calculations. The resulting estimated transition intensities are visible in Figures 3.2-3.4.

Figure 3.2 displays the estimated intensities μ_y^{ad} as functions of attained age y for males and females. Belgian regulatory life tables XR and XK are also displayed there: life table XK defines minimum premium amount for life insurance policies with a positive sum at risk (thus comprising mainly death benefits) whereas life table XR defines minimum premium amount for policies with a negative sum at risk (thus comprising mainly survival benefits). Life table XK is conservative and generates a relatively high safety loading. Dating back to the 1990s, life table XR does not comprise any safety loading anymore for women (but since it only defines minimal premium amounts, insurers remain free to charge higher premiums to remain solvent). We recognize on Figure 3.2 the exponential increase in mortality at adult ages (the accident hump is not visible because actual values are displayed along the vertical axis, without log transform).

Figure 3.3 displays the estimated intensities μ_y^{ai} as functions of attained age y for the three types of cancer considered in this paper, separately for males and females. We can see there that incidence curves greatly differ among the three cancer types under consideration. In particular, after age 30, incidence for women breast cancer largely exceeds the one for melanoma and thyroid. For males, incidence rates are closer at young ages but exhibit different age trends.

Estimated $\widetilde{\mu}(y, z)$ (with y and z corresponding to integer part of age at diagnosis and time since diagnosis, respectively) is displayed in Figure 3.4. We can see there that mortality increases with age at entry and sojourn time for both genders and all three considered cancer sites. Mortality, however, increases less rapidly with sojourn time for young patients compared to old patients. We also see that, for patients below age 40, mortality remains low even after a long period after diagnosis (i.e., for large values of sojourn time). Note that since we computed these quantities considering

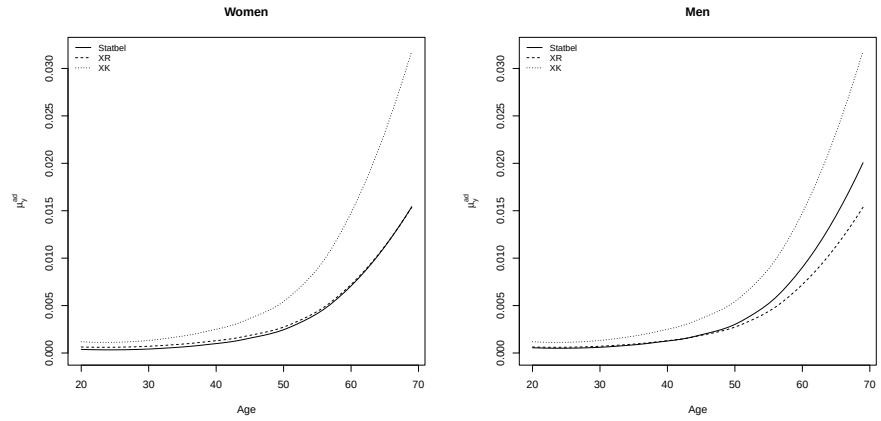


Figure 3.2: Estimated transition intensities μ_y^{ad} as functions of attained age y. General population (Statbel, continuous line) and insurance regulatory life tables XR (broken line) and XK (dotted line).

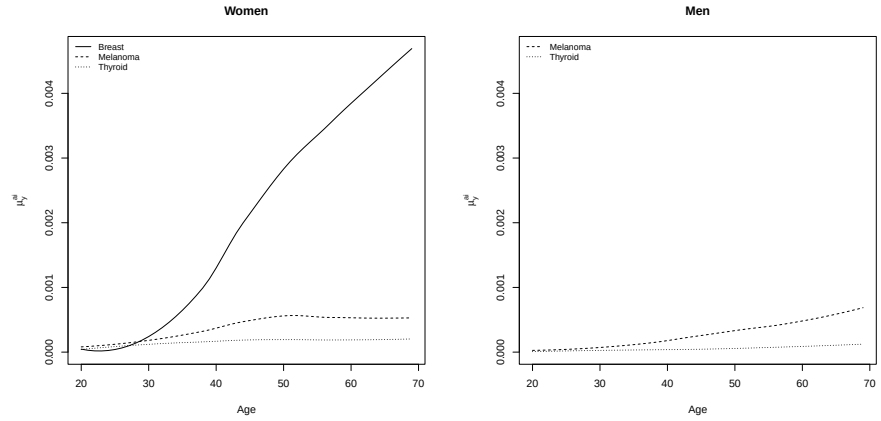


Figure 3.3: Estimated transition intensities μ_y^{ai} as functions of attained age y, for different cancer types.

all causes of death, they account for both mortality from the cancer of interest and mortality from other causes. This is in line with the application to insurance since benefits do not vary according to the cause of death for the products considered in this paper.

3.2.5 Transition probabilities

The following probabilities are useful to perform actuarial calculations. Considering an individual who is healthy at age $x + t$, that is, who is in state a at time t, the probability of being in state i at time $t + u$ is denoted as

$$ {}_up_{x+t}^{ai} = \mathrm{P}[X_{t+u} = i | X_t = a], $$

the probability of being in state d at time $t + u$ is denoted as

$$ {}_up_{x+t}^{ad} = \mathrm{P}[X_{t+u} = d | X_t = a], $$

and the probability of being in state a at time $t + u$ is denoted as

$$ {}_up_{x+t}^{aa} = \mathrm{P}[X_{t+u} = a | X_t = a]. $$

Since the time spent in state i influences future transitions, the random variable Z_t also enters the transition probabilities from that state. Precisely, considering an ill individual aged $x + t$ who has been diagnosed at time $t - z$, that is, who is in state i at time t since time $t - z$, the probability of being in state d at time $t + u$ is denoted as

$$ {}_up_{x+t;z}^{id} = \mathrm{P}[X_{t+u} = d | X_t = i, Z_t = z] $$

and the probability of being in state i at time $t + u$ is denoted as

$$ {}_up_{x+t;z}^{ii} = \mathrm{P}[X_{t+u} = i | X_t = i, Z_t = z]. $$

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

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

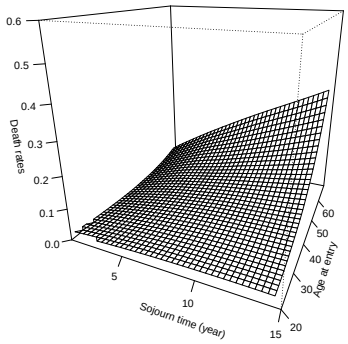
Sojourn probabilities are easy to compute when transition intensities are piecewise constant, as exponential functions of minus integrated exit rates (or cumulative hazards). This property will be used repeatedly in the next section.

3.3 Cancer insurance products

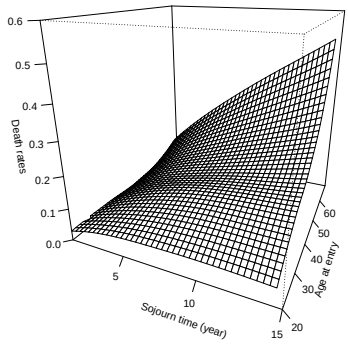
3.3.1 Notation and specific policy conditions

Henceforth, $v(s, t)$ is the present value at time s of a unit payment made at time t (with $s < t$ and $v(s, s) = 1$). We assume that the technical interest rate used in actuarial calculation is constant and we denote as δ the corresponding instantaneous force of interest, that is, $v(s, t) = \exp\big(-\delta(t - s)\big)$. Also, we denote as $\mu_{x+t}^{a\bullet}$ the exit intensity from state a, that is, $\mu_{x+t}^{a\bullet} = \mu_{x+t}^{ad} + \mu_{x+t}^{ai}$.

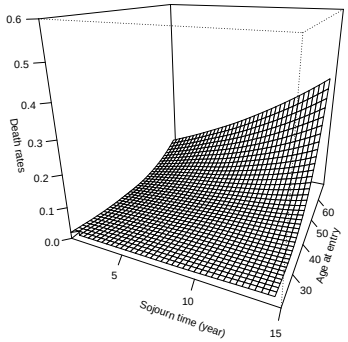
In this paper, we consider temporary covers where diagnosis has to occur within the next n years to get the insurance benefit. The insured period is thus the time interval $[0, n]$, in the sense that a benefit is payable only if the time of diagnosis belongs to this interval. In principle, the insured period begins at policy issue and ends at



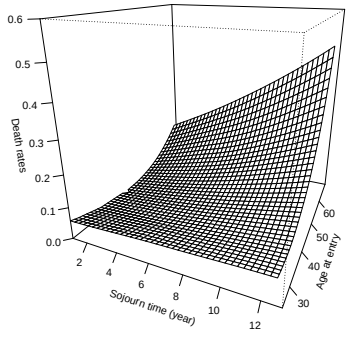
(a) Women with melanoma cancer



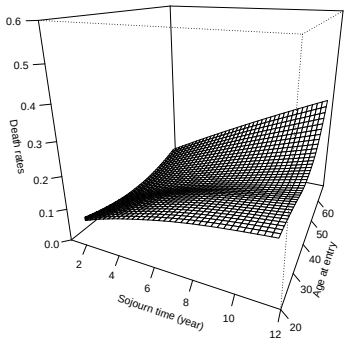
(b) Men with melanoma cancer



(c) Women with thyroid cancer



(d) Men with thyroid cancer



(e) Women with breast cancer

Figure 3.4: Mortality intensities according to age at diagnosis and sojourn time in the cancer state.

policy termination, subject to the following specific policy conditions. The waiting period (or "elimination" period) w is the period following the policy issue during which the insurance cover is not yet operating. The waiting period aims at limiting the effects of adverse selection, in particular because of pre-existing insured's health conditions. The waiting period considered here is the one specified in the contract, not to be confused with the waiting period opening the "right to be forgotten" fixed by the law. Sometimes, the benefit is not payable at diagnosis but only if the policyholder has survived a certain minimum period after diagnosis, called the deferred period and denoted as f. This policy condition acts as a deductible and essentially purposes to reduce the cost and hence the premium of the insurance product; premium reduction can be particularly significant in case high mortality immediately follows diagnosis.

All premium calculations are performed on a unisex basis, in accordance with EU regulation which prohibits any difference in insurance cover or amount of premium by gender. Transitions observed for males and females have therefore been combined to produce a set of intensities independent of gender, which are used for all calculations performed in this section. For computations with respect to breast cancer, transitions for women only have been considered as only female breast cancer cases are included.

Let us comment on the particular case of the breast cancer cover. Even if this contract targets female policyholders, male breast cancer, though rare, does exist. To be consistent with the anti-discrimination EU directive, the coverage has to be offered to both males and females and priced under unisex basis. Notice that a unisex tariff may well be based on women's data exclusively (the requirement is that the premium cannot differ between male and female policyholders). Even if breast cancers are generally rare for males, they often have a very poor prognosis. This is because (i) these cancers are often detected at a later stage compared to women, (ii) there is not much research devoted to breast cancer affecting males and (iii) available treatments against female breast cancer are difficult to adapt to treat male patients because of the marked difference in hormonal status. According to the BCR, the 5-year prevalence from 2013 to 2017 in Belgium is 398 for men compared to 47,423 for women.

Since this cover may raise some concerns related to gender-based discrimination, let us consider the Femina cover sold in Belgium by AG Insurance (one of the market leaders). This is a sickness insurance product with lump sum benefits paid after diagnosis of some specific cancers affecting women, including breast cancer that may also affect men. As its name indicates, this product clearly targets women. The policy must be issued before the age of 60, offering a lifelong cover (subject to a severe underwriting conditions). Health Minister had to answer some specific queries by members of the Belgian Parliament about possible discriminatory issues related to Femina (see Question 7 by Deputy Karin Jiroflée to Minister Kris Peeters about Femina insurance, ref. P2321, plenary session of October 5, 2017). Minister Peeters asked the Financial Services and Markets Authority (FSMA, protecting consumers in the financial sector, including bank and insurance) to determine whether this product complies with the anti-discrimination EU Directive. The conclusion of the FSMA study was provided during the meeting of Parliament Economy Commission of March 28, 2018. Since AG Insurance indicated that Femina cover could also be bought by men, no discriminatory issue was found in relation to the product. This shows that even if products are marketed to address specific needs of the male or female population and are priced accordingly using data from the targeted population, this does not violate the EU Directive as long as both genders can access the cover at the same conditions (even if it is very unlikely that males will ever buy the Femina insurance cover).

3.3.2 Stand-alone covers

3.3.2.1 Lump sum

A first possibility is to pay a lump sum at diagnosis. The beneficiary can use this amount to face out-of-pocket expenditures related to treatment. The expected present value of a unit lump sum paid at diagnosis is given by

$$\overline{A}^{a;a\to i}_{x;n\rceil} = \int_0^n v(0,t)\,{}_tp_x^{aa}\mu_{x+t}^{ai}\mathrm{d}t.$$

In case the contract specifies a waiting period w, the integral is over (w,n) instead of $(0,n)$. Since diagnosis is recorded in the BCR database, the payment date t is easy to check.

When transition intensities are piecewise constant, we get

$$\begin{aligned}\overline{A}^{a;a\to i}_{x;n\rceil} &= \mu_x^{ai}\frac{1-\exp\left(-\delta-\mu_x^{a\bullet}\right)}{\delta+\mu_x^{a\bullet}} \\ &\quad+\sum_{j=1}^{n-1}\mu_{x+j}^{ai}\exp\left(-\sum_{k=0}^{j-1}\mu_{x+k}^{a\bullet}-j\delta\right)\frac{1-\exp\left(-\delta-\mu_{x+j}^{a\bullet}\right)}{\mu_{x+j}^{a\bullet}+\delta}.\end{aligned} \tag{3.3}$$

In principle, the payment could be deferred. In many policies, the benefit is not payable until the need has lasted a certain minimum period called the deferred period (Pitacco, 2014). Here, this would mean that the lump sum is not paid at diagnosis but the payment is deferred later on. This may not be desired by the customers buying the product considered here so that we do not consider this possibility. Deferred periods may nevertheless be useful to lower premiums in case they are too expensive.

The values of $\overline{A}^{a;a\to i}_{x;n\rceil}$ obtained in the Semi-Markov 3-state model are displayed in Figure 3.5 for ages $x\in\{20,21,\ldots,40\}$, coverage period $n=20$ years, and yearly interest rate 1%, that is, $\delta=\ln 1.01$. Without discounting, that is, setting $\delta=0$ or $v(s,t)=1$ for all $s<t$, $\overline{A}^{a;a\to i}_{x;n\rceil}$ is the probability of being diagnosed with cancer for a healthy individual aged x over the next n years. Remember that for melanoma and thyroid cancers, it should be interpreted as the probability on the whole population while for breast cancer, it should be interpreted as the probability only among women. We can see on Figure 3.5 that premium amounts remain rather low for melanoma and thyroid cancers, but considerably increase for breast cancer because of larger incidence within the Belgian population (culminating at 0.017 per unit of sum insured at age 40).

Remark 3.3.1. *The single premium $\overline{A}^{a;a\to i}_{x;n\rceil}$ can be converted into a periodic one by dividing it with*

$$\overline{a}^{aa}_{x;n\rceil} = \int_0^n v(0,t)\,{}_tp_x^{aa}dt, \tag{3.4}$$

with the understanding that the premium is payable until diagnosis or death.

3.3.2.2 Temporary life annuities

Insured benefits can also consist in a temporary life annuity starting at diagnosis. This provides cancer patient with a periodic income to face out-of-pocket expenditures. The corresponding expected present value if payments are made continuously at a constant unit rate as long as cancer patient survives is given by

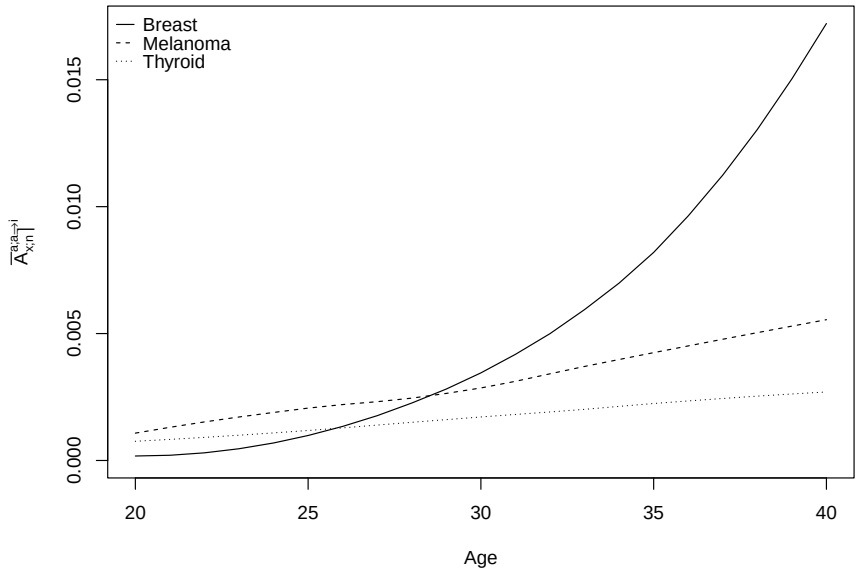


Figure 3.5: Values of $\overline{A}^{a;a\to i}_{x;n\rceil}$ as function of age $x \in \{20, 21, \ldots, 40\}$ for different cancer types with $n = 20$ and yearly interest rate 1%.

$$\overline{a}^{\mathrm{ai}}_{x;n\rceil} = \int_0^n {}_t p^{aa}_x \mu^{ai}_{x+t} v(0,t) \overline{a}^{ii}_{x+t;0} \mathrm{d}t \tag{3.5}$$

where

$$\overline{a}^{ii}_{x+t;0} = \int_0^m {}_s p^{ii}_{x+t;0} v(t, t+s) \mathrm{d}s \tag{3.6}$$

with m denoting the maximal payment duration. Here, m is given in policy conditions and may vary with the waiting period opening the "right to be forgotten" by the law, that is, it may depend on cancer type.

The duration m could be determined in two ways, at least. Either we adopt the reduced waiting periods specified by Royal decree but it would then be necessary to add the duration of treatment since the "right to be forgotten" in the law starts at the end of a successful treatment protocol. Or, we take the duration of the modified "right to be forgotten" since diagnosis as determined by Soetewey et al. (2021). While it could have been interesting to formally compare both approaches, individual data on the type and length of treatment for each case is not reported in the BCR and such information is not readily available. Even if it were available, the definition of the end of the treatment remains unclear (and this is precisely the reason why Soetewey et al. (2021) suggested to let the waiting period start from diagnosis, to avoid endless disputes when a claim occurs). Moreover, durations of treatment are heterogeneous even within the same cancer type, usually unpredictable. Optimal durations are often still open to debates, see e.g. Schvartsman et al. (2019), making it hard to include the duration of treatment in the actuarial computations. In any case, a reduction in treatment length due to the progress made in medical treatment of cancer would obviously lead to closer agreement between the two approaches. Since

the date of diagnosis, as recorded in national registries, offers the great advantage of not being subject to any discussion and to allow the patient to know from the start when the waiting period will end, we think that all parties benefit from using the date of diagnosis instead of the end of treatment. For these reasons, we favor the second approach in the present paper.

When transition intensities are piecewise constant, we get

$$\begin{aligned}\overline{a}^{ai}_{x;n\rceil} &= \sum_{j=0}^{n-1} \int_j^{j+1} {}_tp^{aa}_x \mu^{ai}_{x+t} v(0,t) \overline{a}^{ii}_{x+t;0} \mathrm{d}t \\ &= \sum_{j=0}^{n-1} {}_jp^{aa}_x v(0,j) \int_0^1 {}_tp^{aa}_{x+j} \mu^{ai}_{x+j+t} v(j,j+t) \overline{a}^{ii}_{x+j+t;0} \mathrm{d}t\end{aligned} \tag{3.7}$$

for $j \in \{0, \ldots, n-1\}$ and $t \in [0,1)$, where

$$\begin{aligned}\overline{a}^{ii}_{x+j+t;0} &= \int_0^m {}_sp^{ii}_{x+j+t;0} v(j+t, j+t+s) \mathrm{d}s \\ &= \sum_{k=0}^{m-1} \int_k^{k+1} {}_sp^{ii}_{x+j+t;0} v(j+t, j+t+s) \mathrm{d}s \\ &= \sum_{k=0}^{m-1} {}_kp^{ii}_{x+j+t;0} v(j+t, j+t+k) \\ &\qquad \int_0^1 {}_sp^{ii}_{x+j+t+k;k} v(j+t+k, j+t+k+s) \mathrm{d}s \\ &= \sum_{k=0}^{m-1} \exp\left(- \sum_{l=0}^{k-1} \widetilde{\mu}(x+j,l) - k\delta \right) \int_0^1 \exp\big(- s\big(\widetilde{\mu}(x+j,k) + \delta\big)\big) \mathrm{d}s \\ &= \frac{1 - \exp(-\delta - \widetilde{\mu}(x+j,0))}{\delta + \widetilde{\mu}(x+j,0)} + \sum_{k=1}^{m-1} \exp\left(- \sum_{l=0}^{k-1} \widetilde{\mu}(x+j,l) - k\delta \right) \\ &\qquad\qquad \frac{1 - \exp(-\delta - \widetilde{\mu}(x+j,k))}{\delta + \widetilde{\mu}(x+j,k)} \\ &= \overline{a}^{ii}_{x+j;0}\end{aligned} \tag{3.8}$$

which shows that $\overline{a}^{ii}_{x+j+t;0}$ can be taken out of the integral in the expression of $\overline{a}^{ai}_{x;n\rceil}$. Hence, we have

$$\begin{aligned}\overline{a}^{ai}_{x;n\rceil} &= \sum_{j=0}^{n-1} {}_jp^{aa}_x v(0,j) \overline{a}^{ii}_{x+j;0} \int_0^1 {}_tp^{aa}_{x+j} \mu^{ai}_{x+j+t} \exp(-t\delta) \mathrm{d}t \\ &= \mu^{ai}_x \overline{a}^{ii}_{x;0} \frac{1 - \exp(-\delta - \mu^{a\bullet}_x)}{\delta + \mu^{a\bullet}_x} \\ &\qquad + \sum_{j=1}^{n-1} \mu^{ai}_{x+j} \overline{a}^{ii}_{x+j;0} \exp\left(- \sum_{k=0}^{j-1} \mu^{a\bullet}_{x+k} - j\delta \right) \frac{1 - \exp(-\delta - \mu^{a\bullet}_{x+j})}{\delta + \mu^{a\bullet}_{x+j}}.\end{aligned} \tag{3.9}$$

Values of $\overline{a}^{\mathrm{ai}}_{x;n\rceil}$ are computed as a function of age at policy issue $x \in \{20, 21, \ldots, 40\}$ with $n = 20$ and a yearly technical interest rate of 1%, that is, $\delta = \ln 1.01$. The duration m is taken to be equal to 9 years for melanoma and 1 year for thyroid, in accordance with the reduced waiting periods fixed in the Belgian law for these cancers. For breast cancers, m is taken to be the standard waiting period of 10 years. The numerical values are displayed in Figure 3.6. We can see there that the product is cheaper

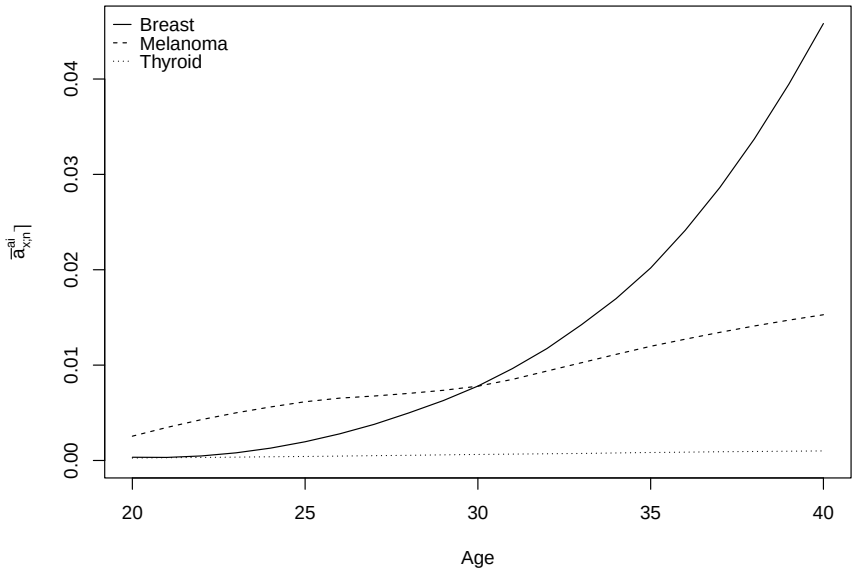


Figure 3.6: Values of $\overline{a}^{\text{ai}}_{x;n\rceil}$ as function of age x for different cancer types with $n = 20$ and yearly interest rate 1%.

for melanoma and thyroid cancers from age 30, thanks to lower incidence rates and payment duration. Moreover, premiums increasing more rapidly with age are obtained for breast cancer, because of higher incidence rates and payment duration.

3.3.3 Combined products

Combined products correspond to insurance packages where cancer insurance supplements a reference cover. Several examples are described hereafter.

3.3.3.1 Premium exemption

A first possibility is to use the temporary life annuity starting at diagnosis for paying the premiums of a reference cover, or even to reimburse a loan secured by mortgage insurance, for instance. Calculations are performed as explained before, with benefits matching amounts of premium or loan reimbursement.

3.3.3.2 Term-life insurance with cancer acceleration benefit

Cancer insurance can be added as a rider to a term-life insurance policy. In this case, the amount of death benefit is (totally or partially) converted into a lump sum paid at diagnosis. Specifically, let

$$\overline{A}^{a;a\to d}_{x;n\rceil} = \int_0^n v(0,t)\,{}_tp_x^{aa}\mu_{x+t}^{ad}\mathrm{d}t \tag{3.10}$$

be the expected present value of a unit lump sum paid at death occurring in state a. In practice, this amount is replaced with the XK premium if the latter is higher. Let c_{ad} be the amount of death benefit for a policyholder in state a. A proportion $\alpha \in [0,1]$

of the death benefit may be paid at diagnosis and the remaining $1-\alpha$ at death in state i. For a unit death benefit, the expected present value of insurance benefits is then given by

$$\begin{aligned}\overline{A}^{(\alpha)}_{x;n\rceil} &= \overline{A}^{a;a\to d}_{x;n\rceil} \\ &\quad + \int_0^n {}_tp_x^{aa}\mu_{x+t}^{ai}\left(\alpha v(0,t) + \int_0^{n-t} {}_zp_{x+t;0}^{ii}\mu_{x+t+z;z}^{id}(1-\alpha)v(0,t+z)\mathrm{d}z\right)\mathrm{d}t.\end{aligned} \tag{3.11}$$

When transition intensities are piecewise constant, we can compute the pure premium as follows. First, the part of the pure premium corresponding to death benefits for a healthy individual writes

$$\begin{aligned}\overline{A}^{a;a\to d}_{x;n\rceil} &= \mu_x^{ad}\frac{1-\exp(-\delta-\mu_x^{a\bullet})}{\delta+\mu_x^{a\bullet}} \\ &\quad + \sum_{j=1}^{n-1}\mu_{x+j}^{ad}\exp\left(-\sum_{k=0}^{j-1}\mu_{x+k}^{a\bullet} - j\delta\right)\frac{1-\exp(-\delta-\mu_{x+j}^{a\bullet})}{\mu_{x+j}^{a\bullet}+\delta}.\end{aligned} \tag{3.12}$$

Second, the accelerated death benefit payable at diagnosis is covered by $\alpha\overline{A}^{a;a\to i}_{x;n\rceil}$ where the expression for $\overline{A}^{a;a\to i}_{x;n\rceil}$ can be found in (3.3). The remaining part of death benefits involves the following integral

$$\begin{aligned}&\int_0^n {}_tp_x^{aa}\mu_{x+t}^{ai}\int_0^{n-t} {}_zp_{x+t;0}^{ii}\mu_{x+t+z;z}^{id}v(0,t+z)\mathrm{d}z\mathrm{d}t \\ &\qquad = \int_0^n {}_tp_x^{aa}\mu_{x+t}^{ai}v(0,t)\overline{A}^{i;i\to d}_{x+t;\overline{n-t}\rceil}\mathrm{d}t\end{aligned} \tag{3.13}$$

where

$$\overline{A}^{i;i\to d}_{x+t;\overline{n-t}\rceil} = \int_0^{n-t} {}_zp_{x+t;0}^{ii}\mu_{x+t+z;z}^{id}v(t,t+z)\mathrm{d}z \tag{3.14}$$

is the present value of a unit benefit payable at death before time n for an individual being diagnosed with cancer at time t. Then, (3.13) can be rewritten as

$$\sum_{j=0}^{n-1}\exp\left(-\sum_{l=0}^{j-1}\mu_{x+l}^{a\bullet} - j\delta\right)\int_0^1\exp\left(-t\left(\mu_{x+j}^{a\bullet}+\delta\right)\right)\mu_{x+j}^{ai}\overline{A}^{i;i\to d}_{x+j+t;\overline{n-j-t}\rceil}\mathrm{d}t, \tag{3.15}$$

with the understanding that an empty sum is zero, where

$$\begin{aligned}\overline{A}^{i;i\to d}_{x+j+t;\overline{n-j-t}\rceil} &= \sum_{k=0}^{n-j-2}\exp\left(-\sum_{l=0}^{k-1}\widetilde{\mu}(x+j+l;l)\right)\widetilde{\mu}(x+j+k;k)v(j,j+k) \\ &\qquad \int_0^1\exp\big(-z\big(\widetilde{\mu}(x+j+k;k)+\delta\big)\big)\mathrm{d}z \\ &\quad + \exp\left(-\sum_{l=0}^{n-j-2}\widetilde{\mu}(x+j+l;l)\right)\widetilde{\mu}(x+n-1;n-j-1)v(j,n-1) \\ &\qquad \int_0^{1-t}\exp\big(-z\big(\widetilde{\mu}(x+n-1;n-j-1)+\delta\big)\big)\mathrm{d}z\end{aligned} \tag{3.16}$$

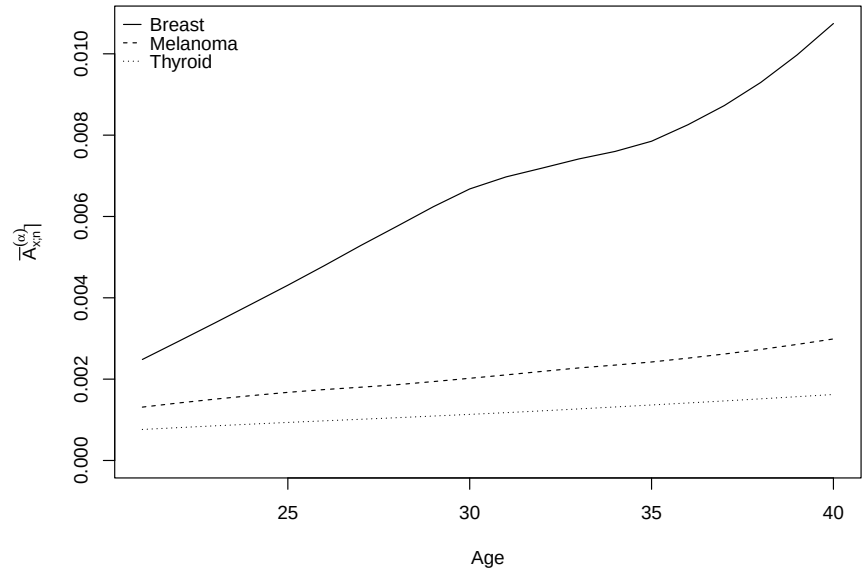


Figure 3.7: Values of $\overline{A}^{(\alpha)}_{x;n\rceil}$ as function of age x for different cancer types with $n = 20$, yearly interest rate 1% and $\alpha = 50\%$.

and

$$\begin{aligned}
&\int_0^1 \exp\big(-z\big(\widetilde{\mu}(x+j+k;k)+\delta\big)\big)\mathrm{d}z \\
&\qquad = \frac{1-\exp\big(-\widetilde{\mu}(x+j+k;k)-\delta\big)}{\widetilde{\mu}(x+j+k;k)+\delta} \\
&\int_0^{1-t} \exp\big(-z\big(\widetilde{\mu}(x+n-1;n-j-1)+\delta\big)\big)\mathrm{d}z \\
&\qquad = \frac{1-\exp\big(-(1-t)\big(\widetilde{\mu}(x+n-1;n-j-1)+\delta\big)\big)}{\widetilde{\mu}(x+n-1;n-j-1)+\delta}.
\end{aligned} \tag{3.17}$$

Premiums $\overline{A}^{(\alpha)}_{x;n\rceil}$ are computed as a function of age at policy issue $x \in \{20, 21, \ldots, 40\}$ with $n = 20$ and a yearly technical interest rate of 1%, that is, $\delta = \ln 1.01$. We take unit death benefit and $\alpha = 50\%$. The numerical values are displayed in Figure 3.7. We can see there that the product is cheaper for melanoma and thyroid cancers for all considered ages, thanks to lower incidence rates. Furthermore, higher premiums increasing with age are obtained for breast cancer, because of higher incidence rates.

Remark 3.3.2. *A temporary life annuity starting at diagnosis can also be financed by "accelerating" the payment of (part of) the death benefit. Specifically, the temporary life annuity starting at diagnosis (that is, at entry in state i) is payable continuously at rate b_i. Denoting as c_{ad} the amount of death benefit in case of transition from a to d, the residual amount of death benefit for an insured in state i dying after having spent a duration z in state i is given by*

$$c_{id}(t,z) = \max\{c_{ad} - b_i z, 0\} = (c_{ad} - b_i z)_+. \tag{3.18}$$

Setting the duration payment m equal to c_{ad}/b_i and converting the death benefit into a temporary life annuity starting at diagnosis, the expected present value of insurance benefits is given by

$$c_{ad}\overline{A}^{a;a\to d}_{x;n\rceil} + b_i\overline{a}^{ai}_{x;\overline{c_{ad}/b_i}\rceil} + \int_0^n {}_tp_x^{aa}\mu_{x+t}^{ai}\left(\int_0^{c_{ad}/b_i}(c_{ad}-b_iz)_zp_{x+t;0}^{ii}\mu_{x+t+z;z}^{id}v(0,t+z)dz\right)dt. \tag{3.19}$$

3.3.4 Cover option

Property loans are often accompanied with mortgage insurance that pays the balance of the loan if the mortgagor dies. Coverage is usually awarded in the form of term insurance with decreasing sum insured, with the amount of death benefit diminishing as the debt is reimbursed. This is common practice in France and Belgium. If the insurer refuses to cover the risk of premature death then the bank does not lend the money, resulting in a barrier to property and home ownership (in case of house loan) and to entrepreneurship (in case of professional loan).

The product considered in this section is issued in state a and offers the beneficiary the option to obtain mortgage insurance at standard conditions even if he or she has been diagnosed with cancer (subject to a deferred period f). The contract stipulates the characteristics of the loan (amount borrowed, amortization plan, maximal loan-to-value of the acquired building, etc.), or puts some limits. The sum insured is the difference between the actual single premium and the reference single premium computed from XK life table. This can also be seen as the expected cost on the Belgian market of a decision like the one taken by Crédit Mutuel in France. We restrict our analysis to single premiums to reduce risk for the insurer.

Let Π^{XK}_{x+t} be the reference single premium for mortgage insurance securing the loan described in the policy conditions, at age $x+t$. The actual premium, given the extra mortality related to cancer for a patient aged $x+t$ who has been diagnosed at time $t-z$ is denoted as $\Pi^{i}_{x+t;z}$. The sum insured is the difference between these two premiums. To avoid under-pricing, we consider that policyholders will exercise their option at the worst time for the insurer. This leads to the worst-case expected present value

$$\mathrm{EPV}_{\mathrm{wc}}(x,n,f) = \max_s \int_0^n {}_tp_x^{aa}\mu_{x+tf+s}^{ai}p_{x+t;0}^{ii}\Big(\Pi^{i}_{x+t+f+s;f+s} - \Pi^{XK}_{x+t+f+s}\Big)v(0,t+s+f)\mathrm{d}t \tag{3.20}$$

where n is the coverage period and f is the deferred period stipulated by the contract.

We consider here the same reference outstanding loan balance cover as in Soetewey et al. (2021). Specifically, we consider a home loan of duration 20 years (typical duration in Belgium). The mortgage insurance applicant aged x borrows an amount 100,000 at interest rate 2%. The technical interest rate for mortgage insurance is 1% and the insurance cover is over the full term of 20 years. These characteristics have been chosen as they represent a rather standard setting.

It is well documented that excess mortality generally decreases with time since diagnosis for most cancer types. Figure 6 in Soetewey et al. (2021) shows that $\Pi^{i}_{x+t;z}$ peaks at $z=0$ for thyroid cancer at ages 30 and 50 and melanoma at age 50 or $z=1$ for melanoma at age 30 before decreasing for larger values of z. The upper panel in Figure 3.8 displays premiums $\Pi^{i}_{x+t+k;k}$ and Π^{XK}_{x+t+k} for $x+t=30$ and 40, and time k since diagnosis in $\{0,1,\ldots,10\}$. The corresponding differences are shown in the lower panel in Figure 3.8. When the difference is negative, the cover described in this

section is not needed since cancer patients can be covered at standard premium rates. We can see there that the maximum is generally attained at $s = 0$ when $f \geq 2$.

Let us take $f = 2$ so that access to mortgage insurance at standard rate is granted two years after diagnosis. This is expected to address patients' needs since it is very unlikely that they wish to buy a house right after diagnosis, the two-year deferred period being devoted to the acute phase of treatment. Notice that this can be offered at no cost for thyroid cancer since $\Pi^{i}_{x+t;t}$ falls below Π^{XK}_{x+t} two years after diagnosis in that case. The worst-case expected present value is then equal to

$$\begin{aligned} \mathrm{EPV}_{\mathrm{wc}}(x,n,2) = &\sum_{j=0}^{n} v(0,j)\, {}_jp^{aa}_x \\ &\int_0^1 {}_tp^{aa}_{x+j}\mu^{ai}_{x+j}\, {}_2p^{ii}_{x+j+t;0} v(j,j+t+2)\Big(\Pi^{i}_{x+j+2;2} - \Pi^{XK}_{x+j+2}\Big)\mathrm{d}t \end{aligned} \tag{3.21}$$

where we have assumed that insurance premiums only depend on integer age. When transition intensities are piecewise constant, we get

$$\begin{aligned} &\sum_{j=0}^{n-1} \exp\left(-\sum_{l=0}^{j-1}\mu^{a\bullet}_{x+l} - j\delta\right)\mu^{ai}_{x+j}\exp(-2\delta) \\ &\quad\int_0^1 \exp\Big(-t\big(\mu^{a\bullet}_{x+j}+\delta\big)\Big)\exp\big(-\widetilde{\mu}(x+j;0)-\widetilde{\mu}(x+j+1;1)\big)\Big(\Pi^{i}_{x+j+2;2}-\Pi^{XK}_{x+j+2}\Big)\mathrm{d}t \\ &= \sum_{j=0}^{n-1} \exp\left(-\sum_{l=0}^{j-1}\mu^{a\bullet}_{x+l} - j\delta\right)\mu^{ai}_{x+j}\exp(-2\delta) \\ &\quad\frac{1-\exp\big(-\mu^{a\bullet}_{x+j}-\delta\big)}{\mu^{a\bullet}_{x+j}+\delta}\exp\big(-\widetilde{\mu}(x+j;0)-\widetilde{\mu}(x+j+1;1)\big)\Big(\Pi^{i}_{x+j+2;2}-\Pi^{XK}_{x+j+2}\Big). \end{aligned} \tag{3.22}$$

The upper panel in Figure 3.9 displays the sum insured $\Pi^{i}_{x+t+2;2} - \Pi^{XK}_{x+t+2}$ according to age $x + t \in \{20, 21, \ldots, 50\}$. We can see that the differences are negative for thyroid cancer so that the product is not needed in that case, as patients can be covered at standard rates after the deferred period. Amounts $\mathrm{EPV}_{\mathrm{wc}}(x, n, 2)$ are displayed in the lower panel of Figure 3.9 for $x \in \{20, 21, \ldots, 40\}$, with $n = 20$, yearly interest rate 1% and the reference outstanding balance cover. Only melanoma and breast cancer are considered since the cover is not relevant for thyroid cancer. The cost appears to be moderate for melanoma but much higher for breast cancer. This results again from the higher incidence of breast cancer compared to melanoma.

3.4 Discussion

In this paper, we have developed a Semi-Markov 3-state model for designing and pricing cancer insurance products on a market where the "right to be forgotten" has been implemented. Different covers are proposed and three cancer types are considered for illustration, with different incidence rates and survival prognosis. It is shown that insurance products can be developed to address the particular needs of patients during the waiting period opening the "right to be forgotten", but that costs greatly vary according to cancer type.

Insurance covers are typically limited to the first cancer occurrence. Several types of cancer can thus easily be combined into a single model to design products offering protection against more than just one cancer site. Considering the three

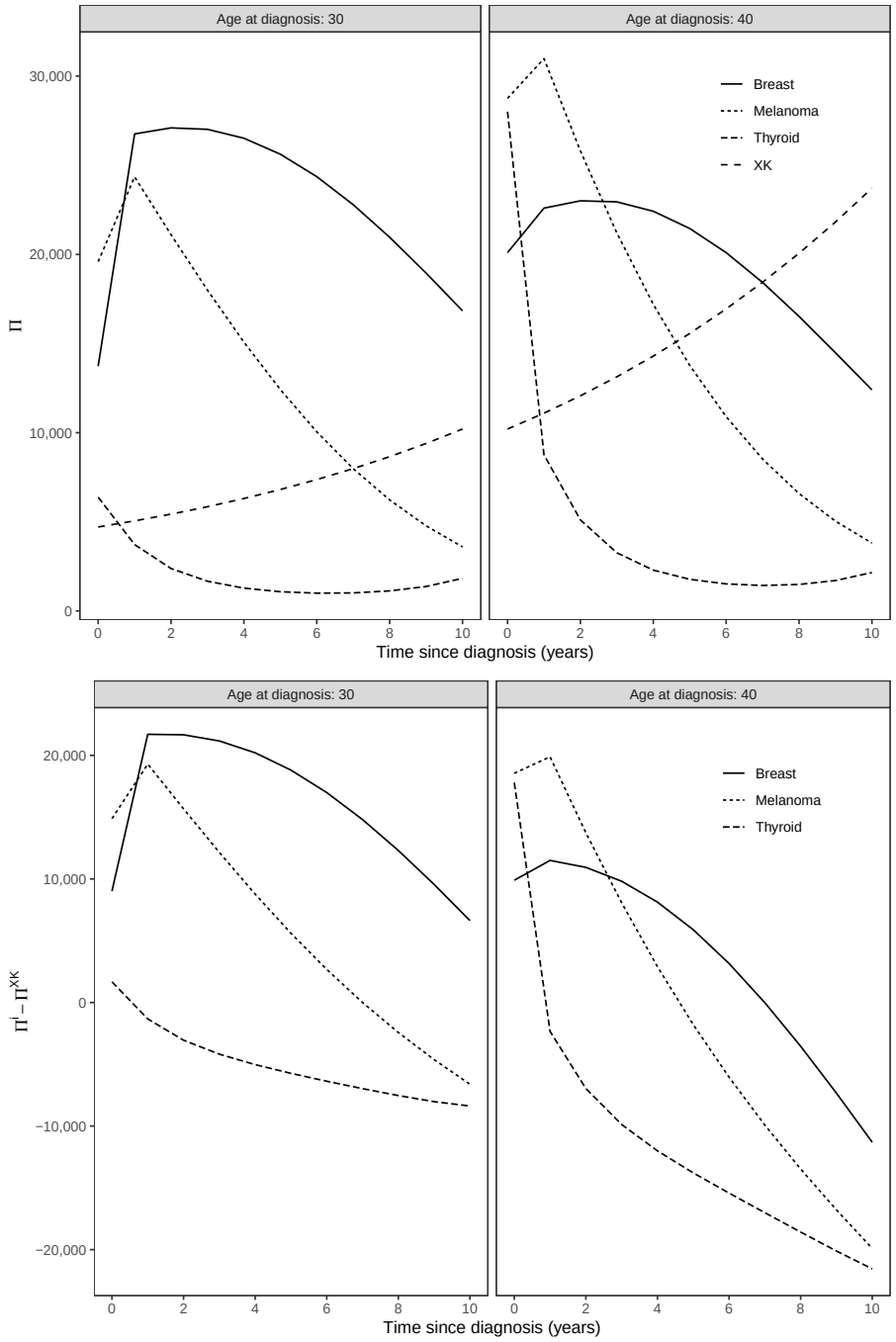


Figure 3.8: Values of premiums $\Pi^{i}_{x+t+k;k}$ and Π^{XK}_{x+t+k} for $x + t = 30$ and 40, and time k since diagnosis in $\{0, 1, \ldots, 10\}$, with $n = 20$, yearly interest rate 1% and the reference outstanding balance cover.

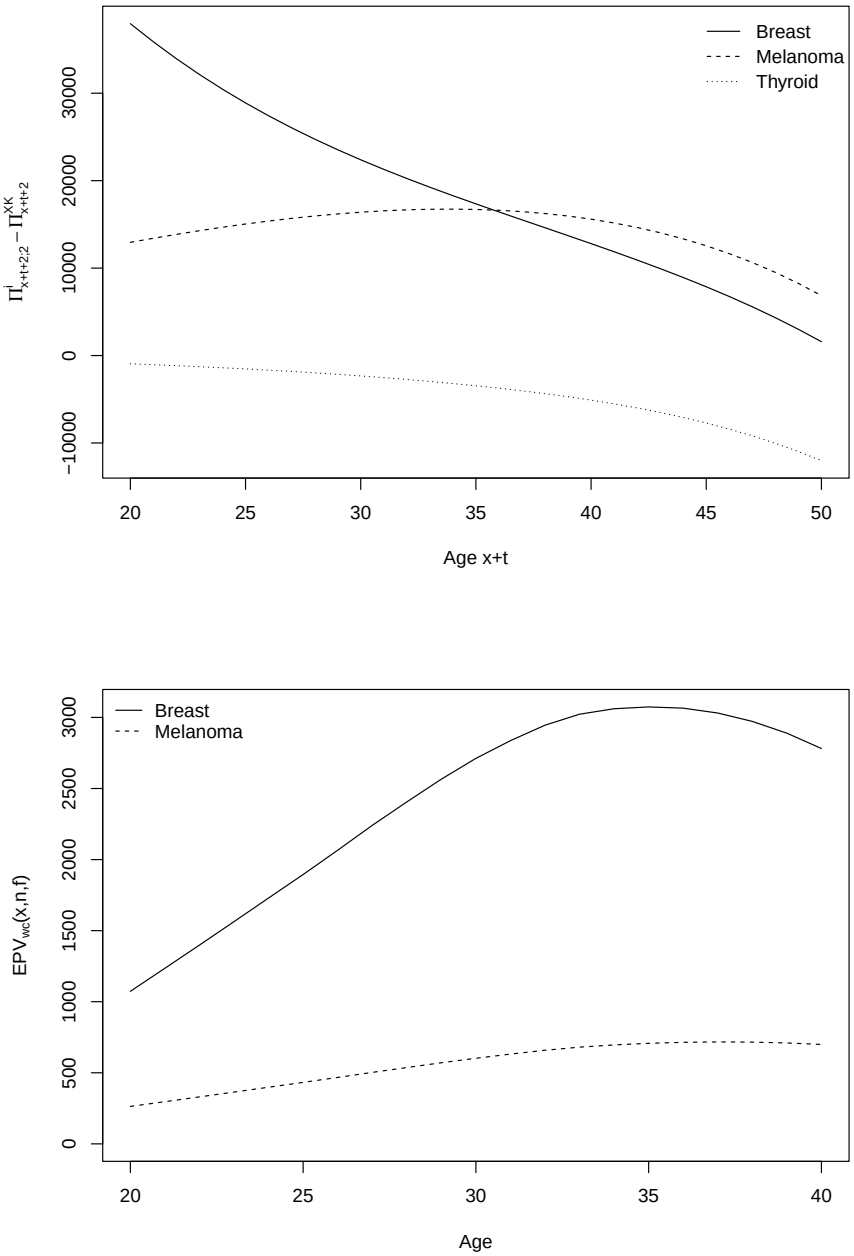


Figure 3.9: Differences $\Pi^{i}_{x+t+2;2} - \Pi^{XK}_{x+t+2}$ according to age $x+t \in \{20, 21, \ldots, 50\}$ in the upper panel and $\text{EPV}_{\text{wc}}(x, n, 2)$ for $x \in \{20, 21, \ldots, 40\}$, with $n = 20$, yearly interest rate 1% in the lower panel.

cancer types considered in this paper, this would result in a hierarchical Semi-Markov model with 5 states, with state i replaced with three states i_1, i_2 and i_3 corresponding respectively to thyroid, melanoma and breast cancer, with transitions from state a to states $\{i_1, i_2, i_3, d\}$ and from each i_1, i_2 and i_3 to state d. The extension to this more general setting is thus straightforward. Of course, distinguishing among cancer types is only relevant if coverage conditions vary with cancer site (for premium calculation, since distinguishing cancer types provides the actuary with a better understanding of cash flows, for instance to compute accurate reserves once a claim has been filed).

In Belgium, new diseases like, amongst others, HIV, some types of hepatitis and leukemia qualified for the "right to be forgotten". Even if the present paper restricts to cancer insurance, a more general critical illness approach would theoretically also be possible. The main difficulty however is the lack of nationwide registry for these diseases. An appropriate source of reliable and representative data must thus be identified to perform actuarial calculations assessing the actual costs of these extensions to the "right to be forgotten".

Health indices for disease incidence risk and duration in the Semi-Markov setting

4

This chapter corresponds, notwithstanding a few minor improvements, to an article carrying the same name as the chapter and submitted jointly with Pr. Catherine Legrand, Pr. Michel Denuit and Dr. Geert Silversmit in *BMC Medical Research Methodology* in 2024. Note that a section has been added at the end of the chapter (which is not present in the submitted manuscript). This final section presents the main findings published so far in the literature and which are related to the topic of this chapter.

Several concepts which have been introduced in Chapters 2 and 3 are used again in the present chapter. Chapters 2 and 3 are targeted to an actuarial audience, whereas the present chapter is intended for a biostastical public. Therefore, throughout this chapter, some notations introduced in Chapters 2 and 3 (and which are widely known in the actuarial literature) had to be adapted for a more biostatistical audience. A first adaptation, which remains a minor one, concerns the random variable giving the state occupied at time t. It was denoted X_t in Chapters 2 and 3. It is denoted $X(t)$ in the present chapter. Similarly, the random variable defining the time spent in the state occupied at time t was denoted Z_t and is now denoted $Z(t)$. Transition intensities from state i to state j, used to be denoted μ^{ij}, are denoted α_{ij} in the present chapter. Transition probabilities from state i to state j, p^{ij}, are now denoted p_{ij}. Moreover, the states were abbreviated as a, i and d (for $active$ (i.e., healthy), ill and $dead$), but are now referred to as 0, 1 and 2, respectively. A major difference concerns how age and time are defined. In Chapters 2 and 3, in order to comply with the usual notations used in the actuarial community, x corresponded to age and t referred to time. In this chapter, age is denoted as t (which can be seen as the time since birth). Moreover, in survival analysis, T is known to be a non-negative continuous random variable representing the real time to the event of interest. In this chapter, T_{ij} is the age at which an individual moves from state i to state j (which we recall are 0, 1 or 2).

Abstract
Over the last decade, the number of years of life lost (YLL) became a popular tool in biostatistics and epidemiology to measure discrepancies in life expectancy or mortality between a cohort of patients and the general population. Its prominence in the literature is primarily due to its ease of interpretation and because information on the cause of death is not required. Moreover, multi-state models are a powerful statistical approach to study the evolution of individuals between several "states".

Derived from data collected by the Belgian Cancer Registry, encompassing 161,007 cases of melanoma, thyroid, and female breast cancer, a 3-state (healthy–cancer–death) illness-death model is used to illustrate how it can be applied to cancer registry data to estimate the incidence risk, and the number of years of life lost due to cancer at different ages at diagnosis and given that the patient survived some years after diagnosis. Results suggest that the probabilities of being diagnosed with cancer over the next 20 years for a healthy individual remain rather low for melanoma and thyroid cancers for both sexes, but considerably increases with age for female breast cancer. Results also suggest that, for female breast cancer, the number of years of life lost before the age of 70 years due to cancer is highest when diagnosed at young ages and then decreases with age at diagnosis, whereas for melanoma and thyroid cancers, it peaks when diagnosed at later ages (between 35 and 55 years depending on the cancer and sex). It also turns out that the number of years of life lost before the age of 70 due to cancer is larger for men than for women for both melanoma and thyroid cancers. Last, it is found that, for melanoma and thyroid cancer patients diagnosed between the age of 20 and 70 years, once they have survived their cancer for 10 years, the number of years of life lost before the age of 70 due to cancer remains below one year. This indicates that, up to the age of 70 years, these patients lose a limited number of years of life due to cancer compared to the general population.

Keywords: Years of life lost; Multi-state models; Critical illness; Cancer mortality.

4.1 Introduction and motivation

Over the last decade, the number of years of life lost (YLL) became a popular tool in biostatistics and epidemiology to measure discrepancies in life expectancy or mortality. The idea behind YLL is to quantify the number of years of life a specific cohort of patients has lost due, for example, to a given disease, compared to the general population. This measure, as defined by Andersen (2013) and Andersen et al. (2013), has the advantage (compared to others such as the hazard ratio or excess hazard) that it is measured on a time metric (usually in years) making its interpretation easy for policy-makers and meaningful for gauging public health outcomes (Latouche et al., 2019).

It was first introduced to measure the reduction in life expectancy for a group of individuals compared to a hypothetical cohort where no one dies before a given age (Andersen, 2013). However, in most situations, it may seem more natural to measure the reduction in life expectancy for a group of individuals compared to a reference population (where some years of life are lost because of some standard or background mortality rates). In this sense, YLL can be used to estimate the number of years a specific cohort of patients (cancer patients, for instance) are expected to lose compared to the general population (i.e., the reference population to which the cancer cohort is compared). The difference between the life expectancy of the general population and the one of the considered cohort of patients corresponds to YLL. This measure is sometimes referred to as excess YLL because it is the number of years of life patients lose in excess of that seen in the general population. The larger this measure, the more important the societal burden of the disease or condition.

Similarly to the excess hazard, information about the cause of death is not required to estimate YLL, making it a practical measure for population-based studies in which the cause of death is often unavailable or unreliable (Percy et al., 1981). There are two types of YLL. First, the number of years of life lost by the entire cohort, which can be denoted YLL^c, and which is of interest if one wants to estimate at one point in time the global number of years of life lost due to a particular disease (see for instance

Aragon et al. (2008) who rank leading causes of premature death based on the total number of years of life lost due to each cause). This may be used to answers questions such as "How many years of life are lost in the population due to cancer?" (Andersson et al., 2013). It is of great interest to economists, governments and policy-makers to determine which condition or disease has the largest negative impact on citizens and society as a whole (for resource allocation, public health priorities, cancer control progress, etc.). Second, the number of years of life lost (on average) per individual, which we denote YLL^i, and which quantifies how many years of life a patient is expected to lose (see for example Belot et al. (2019) or Latouche et al. (2019)). It answers questions such as "How much does the life expectancy of an individual on average change if diagnosed with cancer?". See examples with common cancers in Chu et al. (2008) who measure health impacts on society using YLL^i. In this situation, YLL^i can be seen as an average per person, whereas YLL^c can be seen as the sum of the years of life lost for each individual in a patient cohort. See a comprehensive overview of the years difference measures in Manevski et al. (2023). Note that individuals do not necessarily lose years compared to the general population; they may also gain years. This is the case, for instance, in the study of the long-term survival of elite athletes for which survival may be better than that of the general population (Antero-Jacquemin et al., 2018).

From a general point of view, the major advantages of YLL^c and YLL^i are that (i) it is measured on a time metric (usually in years), facilitating its interpretation and communication (Baade et al., 2015; Licher et al., 2019), (ii) information on the cause of death is not needed to estimate it, and (iii) it can be computed for any time horizon and for a comprehensive list of causes of death. Andersen (2017) suggested several measures of life years lost among patients with a given disease in the framework of a (Markov or non-Markov) illness-death model, illustrated using data on Danish male patients with bipolar disorder. The main goal of the present study is to demonstrate how YLL^i can be easily estimated from a multi-state model and what the advantages are of doing so, with a focus on two applications using data on Belgian cancer patients. Their use in the context of the right to be forgotten will also be discussed.

Multi-state models (MSM) are a powerful statistical approach to study the evolution of individuals between several "states" (see Andersen et al. (2012) and Hougaard (1999) for a general review). MSM can be seen as an extension of classical survival analysis, in which only the transition from being alive to being dead is considered (De Wreede et al., 2010; Geskus, 2019; Putter et al., 2007). Unlike classical survival models, MSM are used to model processes which go from an initial state (for instance "healthy") to a terminal (also referred as absorbing) state (for example "dead"), but where more than two states are considered, some being transient. For example, considering that the "healthy" state is portioned into two or more intermediate states corresponding to specific stages of a disease (Meira-Machado et al., 2009). Thus, MSM offer a complete and informative representation of the occurrence of intermediate events on the pathway to some final event, notably via transition probabilities which have a natural interpretation (Andersen and Pohar Perme, 2008; Touraine et al., 2016).

In this paper, a 3-state model, assuming that an individual can either be "healthy", "ill" (diagnosed with cancer), or "dead" is considered. We will see that in our context, we actually only need to consider transitions from healthy to ill, healthy to dead and ill to dead. While excluding the possibility to transit from ill back to healthy can be interpreted as assuming that cancer is a permanent condition (which is debatable), we actually decided not to consider it following the parsimony principle since it would not bring any useful information in our context. Indeed, as it will be shown later, in our type of applications, distinguishing the health state of patients between diagnosis and death is actually not required. This non-reversibility greatly simplifies the computations, as in this case, our 3-state process is hierarchical and trajectories

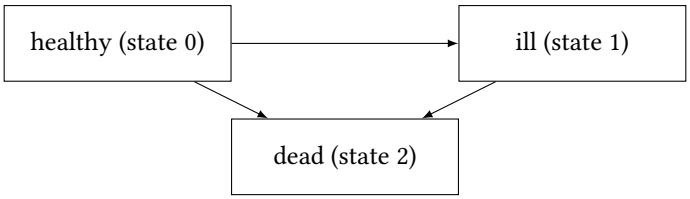


Figure 4.1: Visual representation of the 'illness-death model' without recovery for cancer patients

can be described in terms of just a few random variables (Denuit et al., 2019). See Fig. 4.1 for a visual representation of the model, often referred in the literature as the "(3-state) illness-death model" without recovery. More advanced types of MSM (known as reversible MSM) can be used in case recoveries are possible and has to be taken into account for the application considered. Note that this 3-state model is, in its mathematical concept, similar to the well-known SIR model (susceptible – infected – recovered) in epidemiology (Anderson, 1991; Kermack and McKendrick, 1927). The difference with our 3-state illness-death model is that a susceptible individual must go through the infectious state before being recovered, he/she cannot go directly from "susceptible" to "recovered".

The key contribution of this paper is thus to illustrate how disease incidence risk and YLL^i can be estimated based on a Semi-Markov 3-state MSM using cancer registry data, and what type of useful information can be obtained out of it. Furthermore, the main advantage of computing these quantities in a Semi-Markov context is that it allows to take into account the number of years a patient survived after the diagnosis. To the best of our knowledge, most studies refer to the number of years of life lost at the time of diagnosis, without taking the time survived since diagnosis into consideration. This is a major difference, given that time spent in the ill state is known to have an influence on survival for cancer patients.

The remainder of this paper is laid out as follows. Section 4.2 presents the data used to perform the present study. Section 4.3 details the methods and tools, with a focus on Semi-Markov MSM. Section 4.4 illustrates two useful MSM-based health indices. The final section (Section 4.5) concludes the paper with a discussion.

4.2 Data

For these applications, the data available from the Belgian Cancer Registry (BCR) are considered. The BCR is a national population-based cancer registry collecting data on all new cancer diagnoses in Belgium since the incidence year 2004. For the execution of this main task, the BCR relies on its own specific legislation (more information can be found on the BCR website, at `kankerregister.org`).

To illustrate our work, the methods were applied to three cancer types: melanoma (ICD-10 C43), thyroid (ICD-10 C73) and female breast (ICD-10 C50) cancer. These three cancer sites have been selected to evaluate the proposed method in different scenarios. Melanoma and thyroid cancer patients are known to have a limited excess hazard compared to the general population and high survival rates (CRUK, 2023b,c; NHS Digital, 2023; Soetewey et al., 2021). The situation for female breast cancer patients is different with usually a high survival probability in the first years after the date of diagnosis before it eventually decreases due to late cancer recurrences (CRUK, 2023a). Only female breast cancer is considered as there are too few registrations regarding male breast cancer.

Out of a total of 161,007 cases, melanoma, thyroid and breast cancer represent,

Sex	Cancer site	Age at diagnosis	Lost to follow-up	Number of included cases	Number of deaths
Men	Melanoma	20-34	3.72%	969	94
		35-49	2.66%	3,266	404
		50-69	1.70%	7,460	1,583
Total				11,695	2,081
Men	Thyroid	20-34	4.10%	366	6
		35-49	3.12%	961	67
		50-69	2.14%	1,773	379
Total				3,100	452
Women	Melanoma	20-34	3.62%	2,488	78
		35-49	1.47%	6,137	382
		50-69	1.35%	8,893	1,112
Total				17,518	1,572
Women	Thyroid	20-34	3.80%	1,607	14
		35-49	2.67%	3,449	107
		50-69	2.06%	4,085	484
Total				9,141	605
Women	Breast	20-34	2.76%	3,112	502
		35-49	1.78%	32,743	4,058
		50-69	1.31%	83,698	15,946
Total				119,553	20,506

Table 4.1: Number of persons diagnosed with melanoma, thyroid and female breast cancer in Belgium between 2004 and 2020 (BCR data) by sex, site and age group, together with the percentage of lost to follow-up and the number of deaths.

respectively, 29,213 (18.1%), 12,241 (7.6%) and 119,553 (74.3%) cases diagnosed between 2004 and 2020. Patients were followed-up until April 11, 2022, resulting in a follow-up ranging from 2 to 18 years. Only one record per patient (with the earliest incidence date) within each cancer site was kept for patients with multiple primary diagnoses. A minority of patients without national security number were excluded from the analysis. Patients lost to follow-up (mostly due to moving abroad) and patients still alive at the end of the follow-up period were treated as censored observations.

Table 4.1 summarizes the number of included cases, number and proportion of deaths and percentage of lost to follow-up before April 11, 2022 per type of cancer, sex and age group. The fraction of patients lost to follow-up per subgroup varied from 1.31% for women with breast cancer aged 50-69 to 4.1% for male thyroid cancer patients aged 20-34. The total fraction of patients lost to follow-up cases, regardless of sex, site or age group was 1.64%. Moreover, mean age at diagnosis was 50.5 years (standard deviation (SD) = 12.1), 48.1 years (SD = 12.4) and 54.6 years (SD = 9.5) for melanoma, thyroid and breast cancer, respectively.

In order to estimate the number of years of life lost, mortality in the cancer cohort must be compared to the expected mortality in the general population. Mortality in the general population is therefore also needed. The complete Belgian population is also required to estimate the transition from healthy to ill (which cannot be estimated based on the cancer registry data). These general population data come from the Belgian population life tables, which are available from Statbel (the Belgian statistical office) and can be freely downloaded from the website `statbel.fgov.be`.

Note that as population life tables take into account all deaths, those due to the cancer of interest are also included. Nonetheless, it is commonly assumed that the

fact that population life tables include cancer mortality is not an issue since mortality for a given cancer represents only a small fraction of the overall mortality. Correcting for this mortality of the cancer being studied has, in practice, an insignificant effect on survival of the general population (Esteve et al., 1994; Oksanen, 1998).

4.3 MSM and YLL for cancer patients

A MSM, which is a model for time-to-event data, consists of states and transitions between pairs of states that reflect the disease and death mechanism in medical applications. Main motivations for using a MSM are often to obtain (i) more biological insight into the disease or recovery process of a patient, and (ii) more accurate predictions than standard models neglecting intermediate states. Indeed, by incorporating intermediate events, predictions are adjusted in the course of time, giving more precise information about survival duration (De Wreede et al., 2010; Geskus, 2019).

When considering MSM, the following concepts must be distinguished: (1) Markovian and Semi-Markovian, and (2) homogeneous and non-homogeneous models. These concepts can be defined as follows

- Markovian: what happens next only depends on the current state, not on what happened before.
- Semi-Markovian: what happens next depends on the current state and how long ago it was reached (so the duration in that state).
- Homogeneous or time-homogeneous: transition between states do not depend on time (but time seen as age and not duration in the state, hence the name *time*-homogeneous).
- Non-homogeneous or time-inhomogeneous: transition between states may depend on time (seen as age, not duration).

For a non-homogeneous Markov model, the time until the next state is allowed to depend on the current state and the individual's age (i.e., time). For a homogeneous semi-Markov model, the time until the next state is allowed to depend on the current state and the time since he/she entered this state (i.e., duration). For a non-homogeneous Semi-Markov model, both aspects (time and duration) are combined: the time until the next state is allowed to depend on the current state, the time since he/she entered this state, and his/her age.

Thus, in our context, assuming a homogeneous Markov illness-death model would mean to consider that the expected length of stay in the ill state of a cancer patient depends only on the current state. In other words, it would assume that two cancer patients have the same expected length of stay in the ill state (and thus, the same mortality), even if one has been diagnosed for one year and the other for 10 years. However, it is known that mortality for cancer patients (and thus expected length of stay in the ill state) varies with time since diagnosis (and thus sojourn time) (Soetewey et al., 2022). Therefore, the Markovian assumption does not hold for our situation, and a Semi-Markov assumption taking also into consideration the time spent in the ill state is preferable. Moreover, the non-homogeneous assumption is also preferable as transitions may depend on patient's age. In this non-homogeneous Semi-Markov case (also known as general Semi-Markov), the expected length of stay in the ill state of a cancer patient will thus depend on both the age and the time since diagnosis. This assumption is important because it allows to update the patient's life expectancy conditional on the fact that he/she survived up to that time and a given specific age. This is the reason why our calculations are performed in the context of a non-homogeneous Semi-Markov illness-death model.

The whole process from birth to death of any individual can be defined formally as a random process over time $X = [X(t), t \geq 0]$, where $X(t)$ gives the state occupied at age t. Here, t corresponds to the time since birth. In the irreversible illness-death process depicted in Fig. 4.1, $X(t)$ has values in state space $\mathcal{S} = \{0, 1, 2\}$ where state 0 corresponds to the "healthy" state, state 1 to the "ill" state and state 2 to the "dead" state. Individuals are initially with no cancer detected, thus considered as healthy. Then, they may be diagnosed with cancer and die, or they may die without having been diagnosed with cancer.

More formally, let's denote by T_{ij} the age at which the patient moves from state i to state j. For patients diagnosed with cancer at age T_{01} and who died at age T_{12}, we have

$$\begin{aligned} X(t) &= 0 \quad 0 \leq t < T_{01}, \\ X(t) &= 1 \quad T_{01} \leq t < T_{12} \text{ and} \\ X(t) &= 2 \quad t \geq T_{12}. \end{aligned}$$

For patients without cancer who died at age T_{02}, we have

$$\begin{aligned} X(t) &= 0 \quad 0 \leq t < T_{02} \text{ and} \\ X(t) &= 2 \quad t \geq T_{02}. \end{aligned}$$

Remember that it is assumed that a cancer patient stays in the "ill" state until he/she dies (i.e., the transition from state 1 to state 0 is not allowed). So, in fact the state "ill" should rather be understood as "having been diagnosed with a cancer".

In our context, we have to assume that the time spent in state i influences transition to the next state. Therefore, the random variable $Z(t)$ is introduced, and defined as the time spent in the state occupied at time t. Formally,

$$Z(t) = \max\{z \leq t | X(t) = X(t-h) \text{ for all } 0 \leq h \leq z\}.$$

For an individual in state i at time t, $Z(t)$ is the time since entry in the state (i.e., time from birth for $i = 0$ and time from diagnosis for $i = 1$). Henceforth, we work under the Semi-Markov assumption: the current state $X(t)$ and the time $Z(t)$ spent in the current state influence future transitions. This means that the stochastic process $[(X(t), Z(t)), t \geq 0]$ is a Markov process.

A fundamental concept in multi-state models is the transition intensities, which govern movements between the different states depending on the state currently occupied and the sojourn time. The following transition intensities fully describe the process in an illness-death model:

$$\alpha_{01}(t) = \lim_{h \to 0} \frac{\mathrm{P}[X(t+h) = 1 | X(t) = 0]}{h} \tag{4.1}$$

$$\alpha_{02}(t) = \lim_{h \to 0} \frac{\mathrm{P}[X(t+h) = 2 | X(t) = 0]}{h} \tag{4.2}$$

$$\alpha_{12}(t; z) = \lim_{h \to 0} \frac{\mathrm{P}[X(t+h) = 2 | X(t) = 1, Z(t) = z]}{h} \tag{4.3}$$

where $\alpha_{ij}(.)$ are the transition intensities between state i and state j ($i = 0, 1; j = 1, 2$). Transition intensities from state 0 depend on the time spent in that initial state through attained age. Furthermore, there is an influence of the duration of stay in state 1 so that transition intensities from state 1 depend on both attained age and

time z since diagnosis. In our context, $\alpha_{01}(.), \alpha_{02}(.)$ and $\alpha_{12}(.;.)$ are, respectively, the intensity of developing cancer, the death intensity without cancer and the death intensity with cancer. Also, the exit intensity from state 0 is denoted $\alpha_{0\bullet}(t)$, that is, $\alpha_{0\bullet}(t) = \alpha_{01}(t) + \alpha_{02}(t)$.

Transition probabilities are meaningful to estimate in addition to transition intensities. Considering an individual who is healthy at age t, that is, who is in state 0 at time t, the probability of being in state 1 at time $t + h$ is denoted as

$$p_{01}(t, t+h) = \mathrm{P}[X(t+h) = 1 | X(t) = 0],$$

the probability of being in state 2 at time $t + h$ is denoted as

$$p_{02}(t, t+h) = \mathrm{P}[X(t+h) = 2 | X(t) = 0],$$

and the probability of still being in state 0 at time $t + h$ is denoted as

$$p_{00}(t, t+h) = \mathrm{P}[X(t+h) = 0 | X(t) = 0].$$

Since the time spent in state 1 influences future transitions, the random variable $Z(t)$ also enters the transition probabilities from that state. Precisely, considering an ill individual diagnosed at age T_{01} and aged $t = T_{01} + z$, that is, who is in state 1 since the last $z = t - T_{01}$ years, the probability of being in state 2 at time $t + h$ is denoted as

$$p_{12}(t, t+h; z) = \mathrm{P}[X(t+h) = 2 | X(t) = 1, Z(t) = z]$$

and the probability of still being in state 1 at time $t + h$ is denoted as

$$p_{11}(t, t+h; z) = \mathrm{P}[X(t+h) = 1 | X(t) = 1, Z(t) = z].$$

As explained before, we do not need to consider the possibility to move back to the initial state, or to transition to an intermediate "recovery" state for our applications. Hence, transition probabilities $p_{00}(t, t+h)$ and $p_{11}(t, t+h; z)$ are in reality sojourn probabilities, i.e.

$$\begin{aligned} p_{00}(t, t+h) &= \mathrm{P}[X(t+u) = 0 \text{ for all } 0 < u \leq h | X(t) = 0] \\ p_{11}(t, t+h; z) &= \mathrm{P}[X(t+u) = 1 \text{ for all } 0 < u \leq h | X(t) = 1, Z(t) = z]. \end{aligned}$$

More generally, transition probabilities can be rewritten as

$$p_{ij}(t, t+h; z) = \mathrm{P}[X(t+h) = j | X(t) = i, Z(t) = z] \quad \forall i, j \in \mathcal{S}$$

and transition intensities can be rewritten as

$$\begin{aligned} \alpha_{ij}(t; z) &= \lim_{h \to 0} \frac{\mathrm{P}[X(t+h) = j | X(t) = i, Z(t) = z]}{h} && \forall i, j \in \mathcal{S} \\ &= \lim_{h \to 0} \frac{p_{ij}(t, t+h; z)}{h} && \forall i, j \in \mathcal{S}. \end{aligned}$$

While these transition probabilities and transition intensities give useful information on the evolution of the individuals, obtaining information about survival duration is also of great interest for clinicians and patients. Life expectancy at birth is a metric widely used in demography to measure the length of survival present in a population, and corresponds to the average number of years an individual is expected to live from birth (given that mortality rates remain constant in the future) (Chiang, 1984; Keyfitz and Caswell, 2005; Preston et al., 2001). Moreover, remaining life expectancy is the

average number of remaining years an individual is expected to live, starting from a certain age instead of birth. By computing remaining life expectancy starting at a certain age, it is meant to be conditional on survival to that certain age. If, in addition to estimate life expectancy from a given age instead of birth, it is also estimated up to a given time horizon, it is known as the restricted mean lifetime and it can be interpreted as the average number of years an individual is expected to live between two specific ages. In this paper, we will be particularly interested in taking into account both a starting age different than birth (so conditional on survival to some ages after birth) and a finite time horizon (so considering a given upper age τ). See Section 4.4 for more details about the choices of the starting age and τ.

As mentioned earlier, the number of years of life lost can be seen either at the cohort level (YLL^c) or at the individual level (YLL^i). When applied to cancer patients, on the one hand, YLL^c represents the total number of years of life lost by the cancer cohort. This is useful to compare, for instance, the societal burden of cancer with other diseases or between different countries. On the other hand, YLL^i can be interpreted as the average number of years of life lost that a cancer patient experiences from the time of diagnosis in comparison to an healthy individual of the same age (and possibly sex, year and other covariates such as ethnicity or socio-economic factors). This latter definition resonates more in the patient-clinician communication. In this paper, it is the YLL^i which is chosen and illustrated.

YLL^i in a certain time interval is the sum of life years lost due to (i) population mortality (governed by mortality rates in that reference population) and due to (ii) the cancer of interest. This quantity can be computed based on the estimated survival observed in the general population minus the estimated survival in the cohort of cancer patients considered. Formally, the number of years of life lost due to cancer starting from the age at diagnosis T_{01} until age τ is defined as

$$YLL^i(T_{01}) = \int_{T_{01}}^{\tau} \hat{S}_P(t)\mathrm{d}t - \int_{T_{01}}^{\tau} \hat{S}_C(s)\mathrm{d}s \tag{4.4}$$

where $\hat{S}_P(\cdot)$ denotes the classical survival function estimated via the population mortality rates, and $\hat{S}_C(\cdot)$ is the cancer survival curve (in general, estimated via the nonparametric Kaplan-Meier [1958] method but it could be estimated via another method as well) (Belot et al., 2019).

The lower bound T_{01} in the integrals represents age at diagnosis (so conditional on survival to age T_{01}) and the upper bound τ corresponds to the time horizon, chosen arbitrarily or such that it matches a certain cut-off. The number of years of life lost uses the age at diagnosis for each cancer patient as its starting point and estimate the expected remaining lifetime at that age using age-specific mortality rates. The number of years of life lost due to cancer is then estimated by matching the expected remaining lifetime for someone diagnosed with cancer with the life expectancy in the general population at that specific age. Age-specific mortality rates and life expectancy in the general population are generally available through life tables (as they are usually stratified by age). For life tables that are stratified by sex in addition to age, the number of years of life lost can be used to compare cancer patients to the general population of the same sex and age.

Our objective is to demonstrate how this quantity can be estimated from our MSM. The idea here is to start from our MSM to compute YLL^i using life expectancy, probabilities of developing the disease within a specific time period, and expected lengths of stay in each of the different states (also referred in the literature as the mean sojourn time, see Jackson (2007)). Following Eq. (4.4), estimation of YLL^i via a MSM starting from the age at diagnosis is denoted $\text{YLL}^i_{MSM}(T_{01})$ and corresponds to the number of years of life lost at the time of diagnosis for someone diagnosed at age T_{01}. Fig. 4.2 illustrates the approach, where $e(T_{02})$ is the remaining life expectancy

until the expected death of a healthy individual. One could argue that it does not make sense to speak about age at diagnosis T_{01} if the person has no cancer. However, in fact we compare what would have happened to a patient diagnosed at age T_{01} if he/she would not have had a cancer at the time he/she was actually diagnosed. We are now considering the hypothetical trajectory that a patient diagnosed at age T_{01} would have had if he/she had not had cancer and therefore if he/she had remained in state 0.

Someone with cancer:

Same if he/she would have no cancer diagnosed:

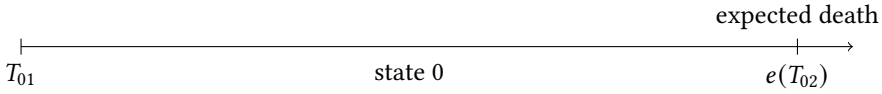


Figure 4.2: Representation of MSM to estimate YLL^i from diagnosis

In the context of a Semi-Markov multi-state model, the remaining life expectancy for a cancer patient diagnosed at age T_{01}, given the time z elapsed since diagnosis is

$$e_{11}^{\tau}(T_{01}+z;z) = \int_{T_{01}+z}^{\tau} p_{11}(T_{01}+z,s;z)\mathrm{d}s. \tag{4.5}$$

Since $t = T_{01} + z$, Eq. (4.5) becomes

$$e_{11}^{\tau}(t;z) = \int_{t}^{\tau} p_{11}(t,s;z)\mathrm{d}s. \tag{4.6}$$

Following Fig. 4.2, to define $\mathrm{YLL}^i_{MSM}(T_{01})$ in a Semi-Markov context we add the conditioning on z to have the number of YLL for someone diagnosed at age T_{01} but that would have already survived with his/her cancer for z years. In that case, we obviously have to update the life expectancy for the cancer patient (the fact that he/she lived already for z years gives an information on his/her life expectancy) and do the same for his "healthy" counterpart. This is denoted $\mathrm{YLL}^i_{MSM}(T_{01};z)$ and is defined as follows

$$\begin{aligned} YLL^i_{MSM}(T_{01};z) &= \text{remaining life expectancy at age } T_{01} \text{ for a healthy individual} \\ &\quad - \text{remaining life expectancy for a cancer patient diagnosed} \\ &\quad\quad \text{at age } T_{01}, \text{ given the time } z \text{ elapsed since diagnosis} \\ &= e(T_{01}) - e_{11}^{\tau}(T_{01}+z;z) \end{aligned} \tag{4.7}$$

Remaining life expectancy at age T_{01} for a healthy individual is usually found with life tables and population mortality rates. Here, the expected remaining lifetime until age τ for someone diagnosed with cancer is matched with the τ-restricted life expectancy in the general population at that specific age.

As often the case in practice, transition intensities are assumed to be piecewise constant in order to ease calculations but also given the information available in cancer registries. In that case, transition intensities are easily estimated by the ratio of the observed number of transitions (diagnosis or death) to the corresponding exposure (in the state to be left) (Soetewey et al., 2022). When (annual) piecewise constant transition intensities are considered, we get

$$e_{11}^{\tau}(t;z) = \sum_{k=0}^{\tau-t-1} \exp\left(-\sum_{l=0}^{k-1} \alpha_{12}(t+l;z+l)\right) \frac{1-\exp(-\alpha_{12}(t+k;z+k))}{\alpha_{12}(t+k;z+k)} \tag{4.8}$$

with $\sum_{l=0}^{k-1} \alpha_{12}(t+l;z+l) = 0$ if $l = 0$. The development of $e_{11}^{\tau}(t;z)$ is explained in Appendix A.1.

Remember that $YLL_{MSM}^{i}(T_{01};z)$ is defined at an individual level. In this sense, $YLL_{MSM}^{i}(T_{01};z)$ quantifies the number of years of life a patient diagnosed with cancer z years ago is expected to lose compared to someone who will never develop the disease. It can be seen as an insightful health indicator, complementary to other health indicators already used by clinicians and policy-makers. Indeed, it can be used to communicate about a patient's survival, but it can also serve as a measure of the burden of cancer for the whole society (with comparisons between diseases, countries or throughout the years for example).

4.4 Derived health indices - case of three specific cancer types

One of the main advantages of estimating YLL^i from a MSM is that several health indicators could be derived from it. The focus here is put on two different applications to illustrate its potential uses; (i) the cancer incidence risk and (ii) the number of years of life lost due to cancer given a certain time spent after diagnosis. Note that the first health indicator requires the 3 states. However, regarding the second one, we consider an individual of age T_{01} at diagnosis. This means that state 0 is no longer needed, since we are already in state 1. Also note that incidence refers to the number of new cases of a disease over a specified period, and can be expressed as a risk or an incidence rate (Noordzij et al., 2010). We are interested in the former, that is, the incidence risk that a subject within a population will develop a given cancer, over a specified follow-up period. This incidence risk, expressed as a probability, can be interpreted as an estimation of the risk of cancer in an individual subject over a certain time frame.

For these applications, our analyses are limited to patients aged 20 to 69 years old at time of diagnosis for two main reasons. First, childhood cancers can be seen as a category of cancer on their own and are often studied separately because they differ greatly from adult cancers. Second, τ has been set to 70 years, an age in which persons were censored if they had not died before to focus on active life from a public policy perspective. The estimate of YLL^i has therefore to be interpreted as the number of years of life lost before that specific age. This is analogous to the τ-restricted mean lifetime, which can be interpreted as the average number of years lived before time τ. Note that the choice of τ is arbitrary. In some settings, researchers may be interested in YLL^i before retirement's age applicable in a country. In our case, we are interested in potential implications for insurers in the context of the right to be forgotten, hence the upper limit of 70 years (people aged above are unlikely to contract a loan). Note the distinction between the maximum age at diagnosis (69 years) and the upper age

limit τ (70 years). This difference is explained by the fact that we include patients who have been diagnosed before their 70^{th} birthday (and thus who are still 69 years old at the time of diagnosis), while we are interested in the number of years of life lost before the age of 70 due to cancer. This is to avoid the possibility that a patient is diagnosed between his or her 70^{th} and 71^{th} birthday, while computing the number of years of life lost before he or she has reached the age of 70 years.

To display our results, the time since diagnosis z is set to 0, 5 and 10 years. $YLL^i(T_{01};0)$ corresponds to the number of years of life lost due to cancer at the time of diagnosis for a patient diagnosed at age T_{01}. $YLL^i(T_{01};5)$ and $YLL^i(T_{01};10)$ correspond to the same quantity computed after having survived to the cancer for respectively 5 and 10 years. A z of 5 and 10 years after diagnosis has been chosen to cover a relatively large period of time after diagnosis, while we refrain from setting it higher due to the limited follow-up period in our data.

4.4.1 Incidence risk

We start the applications with the estimation of the probability for the population of age t to be diagnosed with of each the three types of cancer we consider between age t and $t+n$. In other words, the probability of being diagnosed with cancer for a healthy individual aged t over the next n years is computed. This measure, similar to the incidence risk and again assuming yearly-constant intensities, is defined based on a MSM as follows

$$p_{01}(t,t+n) = \sum_{k=0}^{n-1} \alpha_{01}(t+k) \exp\left(-\sum_{l=0}^{k-1} \alpha_{0\bullet}(t+l)\right) \frac{1-\exp\big(-\alpha_{0\bullet}(t+k)\big)}{\alpha_{0\bullet}(t+k)} \tag{4.9}$$

with $\sum_{l=0}^{k-1} \alpha_{0\bullet}(t+l) = 0$ if $l = 0$.

The probabilities of being diagnosed with breast, melanoma and thyroid cancer over the next 20 years for a healthy individual obtained via the Semi-Markov 3-state model are displayed in Fig. 4.3, for ages $t \in \{20, 21, \ldots, 40\}$ and for each sex separately. Fig. 4.3 shows that incidence risk over a 20-year period remains rather low ($< 0.71\%$) for melanoma and thyroid cancers for both sexes, but considerably increases with age for female breast cancer (culminating at 5.12% at age 40).

4.4.2 Years of life lost from diagnosis

Results of $YLL^i_{MSM}(T_{01};z)$ as functions of age at diagnosis ($T_{01} \in \{20, 21, \ldots, 69\}$) and for $z = 0, 5$ and 10 years after diagnosis are presented by sex and cancer site in Fig. 4.4.

We can see that, for both sexes and all three cancers of interest, the longer the time survived after diagnosis (i.e., the greater the z), the lower $YLL^i_{MSM}(T_{01};z)$ (with an exception for women diagnosed with thyroid cancer at the age of 25 and below). For female breast cancer, $YLL^i_{MSM}(T_{01};z)$ is highest when diagnosed at the age of 20 and then decreases with age at diagnosis, whereas for melanoma and thyroid cancers, it peaks when diagnosed at later ages (between 35 and 55 years depending on the cancer and sex). For both melanoma and thyroid cancers, $YLL^i_{MSM}(T_{01};z)$ is larger for men than for women. Botta et al. (2019) who describe the impact of cancer during patients' entire lives found a similar pattern between women and men. Comparisons between sexes cannot be made for breast cancer as only female breast cancer is included. Among men, $YLL^i_{MSM}(T_{01};z)$ is globally lower for thyroid cancer than for melanoma cancer. Among women, $YLL^i_{MSM}(T_{01};z)$ is lowest for thyroid cancer and highest for breast cancer. Note also that, for patients diagnosed with melanoma or thyroid cancer

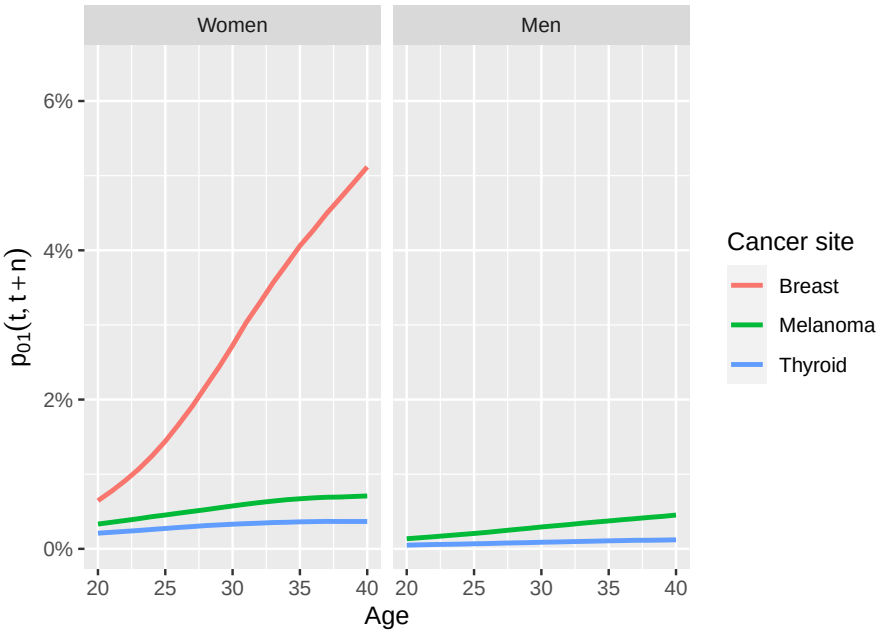


Figure 4.3: Probabilities of being diagnosed with breast, melanoma and thyroid cancer over the next $n = 20$ years for a healthy individual as function of age $t \in \{20, 21, \ldots, 40\}$

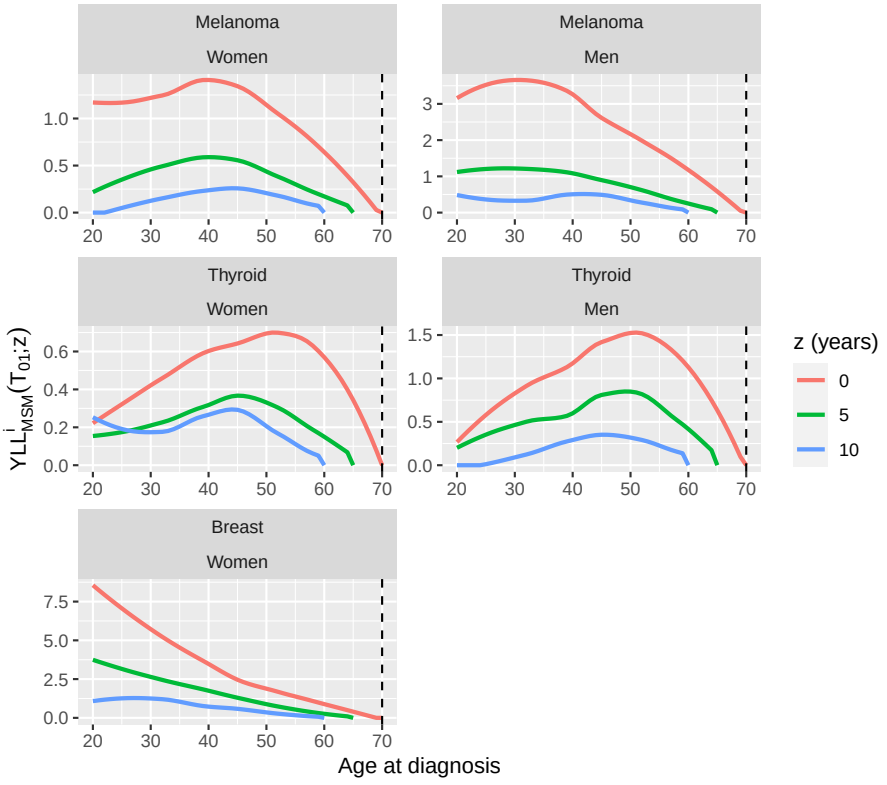


Figure 4.4: Number of life years lost at the individual level before the age of 70 years due to cancer, estimated from $z = 0, 5$ and 10 years after diagnosis, as a function of age at diagnosis

at all considered ages, $YLL^i_{MSM}(T_{01}; 10)$ remains below one year. This indicates that, once they have survived their cancer for 10 years, they lose (compared to the general population and up to the age of 70 years) a limited number of years of life due to cancer.

Remember that $YLL^i_{MSM}(T_{01}; z)$ is computed at the individual level with $\tau = 70$ years, so these figures give the number of years of life a patient diagnosed with cancer is expected to lose due to the disease before the age of 70 years (at the time of diagnosis, 5 and 10 years after diagnosis). This health indicator can, however, also be analyzed in relative terms, that is, in comparison with other cancers, diseases or conditions rather than in absolute terms. Indeed, knowing that a group of patients has more to lose (up to a certain age) in terms of years of life due to a specific disease compared to another one is more meaningful for policy-makers and clinicians. This comparison would allow, for example, to rank diseases in terms of burden to the society, that is, highlight those which are, until a chosen age, the most lethal and the ones which are the most harmless.

It is also worth noting that curves displayed in Fig. 4.4 would be different if another age was chosen for τ. Indeed, the higher the upper age limit τ, the more years of life an individual can lose. The decreasing trend of $YLL^i_{MSM}(T_{01}; z)$ at older ages can be explained partly by the fact that the survival of a cancer patient is approaching that of the general population, and partly by the fact that a cancer patient has simply less years of life to lose before the age of 70 years as he or she approaches that age.

4.5 Discussion

As it has been highlighted on several occasions in the literature, there are several approaches and methods to estimate the number of years of life lost due to cancer (Andersen, 2017). Sometimes, it even has different definitions and meanings depending on the context and the audience (Belot et al., 2019). It is therefore hard to compare YLL^i due to cancer across different studies, in particular when the upper age limit τ is different. In the present study, it is set to 70 years to focus on young adults and active life, while most studies set it at a higher age to consider the number of years of life lost during the entire lifetime (Centers for Disease Control and Prevention, 1993; Gardner and Sanborn, 1990). As mentioned above, the number of years of life lost before a given time horizon (70 years in our illustration) obviously depends on how far is this time horizon. Therefore, it is important to note that results found for the number of years of life lost from diagnosis until age 70 should not be taken as an evaluation of the risk from a medical or biological point of view. Such an information could however still be very useful in a situation where this time horizon would be meaningful, as could for example be the case from an actuarial or economical point of view. Indeed, in the context of the right to be forgotten for instance, the insurer is mainly interested in the survival until the end of the loan contracted. More generally, from a public policy perspective, one may be interested in the number of years of life lost before the age of retirement.

Although it is hard to compare results with existing literature, our results could be considered as in line with Silversmit et al. (2017b), who, also using Belgian data, found a YLL^i of 3.2 years for female breast cancer, 2.5 and 3.6 years for female and male melanoma cancer, and 1.5 and 2.5 years for female and male thyroid cancer, respectively. These results are obtained with as reference age the life expectancy from general population at age of diagnosis, which is mostly larger than 78 years. The interested reader is referred to Andersen et al. (2013); Andersen and Pohar Perme (2008); Andersson et al. (2013); Aragon et al. (2008); Baade et al. (2015); Belot et al. (2019); Botta et al. (2019); Capocaccia et al. (2015) for more methodologies and results in the context of cancer.

There is a vast literature on YLL and MSM in biostatistical and medical studies. The present paper illustrates their relevance for computing a measure of the number of years of life lost before a given age, chosen depending on the situation or the research question. Arik et al. (2023) have shown the implementation of years of life lost in the context of a multi-state model. However, it differs from the present study on several points: (i) it uses a Markov model (so transition intensities do not depend on the duration of stay in the current state), (ii) it is targeted to another age group as it uses data on women diagnosed with breast cancer aged 65–89 years, and (iii) it focuses on the number of years of life lost by the entire cohort. The present paper aims at filling this gap. Some useful applications of MSM-based calculation to derive health indices such as disease incidence risk and number of years of life lost due to cancer targeted to this public have been illustrated.

Most studies refer to the number of years of life lost or remaining life expectancy starting from the date of diagnosis as an estimate of the disease burden (Andersson et al., 2015, 2013; Baade et al., 2015, 2016; Licher et al., 2019; Syriopoulou et al., 2017). This is undoubtedly useful when considering patients who have just been diagnosed, the time at which a patient is most likely to be concerned about his/her survival. Nonetheless, its relevance should not be limited to quantifying the loss of survival at the time of diagnosis. For long-term survivors, it becomes even more pertinent when considering its evolution over time (Botta et al., 2019; Capocaccia et al., 2015). Indeed, there are many applications where one would be interested in the loss of survival due to cancer, given that the patient already survived some years after diagnosis. This is particularly useful for cancers where the amount of time survived since diagnosis has an influence on the patient's survival. This is actually the underlying basis behind the right to be forgotten (Mesnil, 2018; Scocca and Meunier, 2020; Soetewey et al., 2021). Implemented since 2016 in France and since 2019 in Belgium, it states that no difference can be made, in terms of access to an insurance product and the level of its premiums, between a healthy client and a cancer patient if he/she survived 10 years after the end of the therapeutic protocol. YLL^i over time since diagnosis can be interpreted as a measure of how close from being cured long-term survivors can be considered (Botta et al., 2019). A decreasing YLL^i over time since diagnosis shows some evidences that patients who are still alive are approaching the same mortality risks as of the general population. In this context, Capocaccia et al. (2015) proposed a cut-off of less than two years of life lost for colon cancer patients to be considered as statistically cured.

It is important to note that there has been improvements in treatment of advanced melanoma over the last decade, leading to a positive impact on quality of life and overall patient survival (Pasquali et al., 2018; Pedersen et al., 2023; Tichanek et al., 2023; Tromme et al., 2016). Obviously, the bigger the improvements in treatment and overall survival, the more the duration in the ill state is underestimated and the more the number of years of life lost is overestimated. This does not, nonetheless, undermine our analyses for multiple reasons. First, a better prognosis has no impact on the incidence nor on the incidence risk (i.e., the first application of the present study). Second, the largest improvements in treatment and overall survival concern advanced melanoma, so stages III and IV. These two advanced stages represent a limited share of all tumours considered here (8.96% and 4.16% for stages III and IV, respectively). Third, improvements in treatment are quite recent, limiting the impact on the obtained results. Fourth, in the context of the right to be forgotten and from an insurer's point of view, it is more conservative if the number of years of life lost due to cancer before a certain age is overestimated than if it was underestimated.

Melanoma, thyroid and female breast cancers may include a variety of cancer sub-types and could be diagnosed at different stages of severity, leading to differences in terms of survival. It is thus undeniable that including the information on stages of

severity would refine the analysis. This could be achieved, for instance, by stratifying the analyses by cancer stage. However, it has been omitted on purpose for the sake of illustration of the proposed approach.

Cancer is not one disease but a family of many diverse diseases with different outcomes. Results in the present paper focus on melanoma, thyroid and female breast cancer patients, and cannot, at this stage, be transferred to other cancer types. A natural extension of this work would be to repeat the analyses for all major cancer types. Arık et al. (2020) even showed, in a comprehensive study using UK data, that for female breast cancer there are regional differences in terms of cancer morbidity. Thus, the analysis could also be refined to a regional level instead of national level. This is not done in the present paper as it goes beyond the scope of this study which primarily aims to advocate a new method to estimate the number of years of life lost.

Cancer patient survival has improved over the last few decades, with an increasing proportion of patients being cured for many types of cancer (Andersson et al., 2011; Lambert et al., 2006; Silversmit et al., 2017a). Given the increasing numbers of people being diagnosed with cancer, informing patients and involved parties with relevant risk information is crucial (Baade et al., 2015). Providing precise and informative estimate of the reduction in the remaining life expectancy in case cancer is diagnosed or to long-term cancer survivors is therefore of prime importance, for patients, policy-makers and society as a whole. From the literature, it is clear that the number of years of life lost is an important addition to existing measures that give a complete picture of the impact of a cancer diagnosis. The methods proposed in this paper help to estimate this important health indicator from a multi-state model's perspective. This will undoubtedly help to assess when the excess mortality from cancer becomes negligible in cancer survivors, in turn allowing the right to be forgotten to be developed further.

In this study, the assumption is made that a cancer patient cannot become healthy again (i.e., transition from the ill to the healthy state is not possible). Although this assumption is believed to be reasonable for most cancers, one may argue that it does not always hold. However, in our context, the real transition of interest is more from ill to dead than from ill to healthy, following the reasonable paradigm that staying long enough in the ill state to die from something else is, at least from a statistical point of view, equivalent to be cured (cfr. the idea of "statistical cure" for example in Boussari et al. (2018); Jakobsen et al. (2020); Tralongo et al. (2017)). Also, the main objective of this study is to illustrate how the concept of MSM can be applied to estimate another well-known quantity in medicine and epidemiology, which has not yet been done so far. Using more advanced MSM to estimate the number of years of life lost is undoubtedly an interesting question, but left for future research.

For cancer patients, quality of life may be considered as important as the length of life itself (Shrestha et al., 2019). The number of years of life lost gives an easily interpretable measure about survival of cancer patients. However, other indicators such as, among others, the disability-adjusted life years (DALY) should also be considered, in particular for diseases or conditions that cause significant disability or do not result in death. Note that even though it is the number of years of life lost due to cancer that is estimated, the methods proposed in this paper is not limited to cancer and could be applied to several other diseases or conditions (diabetes and HIV, amongst others).

4.6 Additional notes

As mentioned earlier, there has been a proliferation of research on the topic of the number of years of life lost, in several countries and for several conditions or diseases. Findings from other studies cannot be compared to each others, nor to our results due to the fact that time horizons are different. Nonetheless, for the sake of completeness and for the interested reader, we highlight the main findings related to cancer research.

Andersen et al. (2013) found a YLL^i due to cancer (all types of cancer) ranging from 0.11 to 3.68 years for Danish males and from 0.21 to 1.62 years for Russian males, depending on the age at diagnosis and the method used. Also using Belgian data, Silversmit et al. (2017b) found a YLL^i of 3.2 years for female breast cancer, 2.5 and 3.6 years for female and male melanoma cancer, and 1.5 and 2.5 years for female and male thyroid cancer. These results are obtained with as reference age the life expectancy from general population at age of diagnosis, which is mostly larger than 78 years. Baade et al. (2015) found a YLL^i due to, respectively, melanoma and female breast cancer ranging from 3 years (at 40 years old) to 1 year (at 80 years old) and from 12.1 years (at 40 years old) to 1.6 years (at 80 years old). Capocaccia et al. (2015) obtained a YLL^i due to female breast cancer ranging from 8.7 years at age 40–44 to 2.4 years at ages 70–74. For patients diagnosed at age 45 years, Botta et al. (2019) found a YLL^i below 6 years for thyroid cancer in women and melanoma in men. Andersson et al. (2013) arrived at a YLL^i for female breast cancer ranging from 13 years (50-59 age group) to 2.2 years (80+ age group) and from 9.13 years (50-59 age group) to 1.84 years (80+ age group) for melanoma cancer. Finally, Belot et al. (2019) found a YLL^i due to colon cancer over a 10-year time window ranging from 4.14 to 4.77 years depending on the socioeconomic group.

Right to be forgotten for mortgage insurance issued to cancer survivors: Critical assessment and new proposal

5

This chapter corresponds, notwithstanding a few minor improvements, to an article carrying the same name as the chapter and submitted jointly with Pr. Catherine Legrand, Pr. Michel Denuit and Dr. Geert Silversmit in the *European Actuarial Journal* in 2023.

Like Chapters 2 and 3, the present chapter is targeted to an actuarial audience. Therefore, notations introduced in these two chapters are repeated as closely as possible in this chapter.

Abstract
Soetewey et al. (2021) proposed to determine the waiting period opening the right to be forgotten (RTBF) as the time after diagnosis needed for the premium to revert back to some acceptable level expressed by means of regulatory life tables. However, this approach requires data up to 30 years after diagnosis (10 years of standard RTBF plus the typical duration of the loan), or extrapolating the results up to that time horizon. When survival statistics are only available over a shorter duration, it turns out that the results may strongly depend on the extrapolation method. This is why an alternative method is proposed here, based on a constraint imposed to the premium. This constraint is then transposed into a target on the conditional observed survival and the waiting period follows. For the sake of robustness, results obtained with the proposed approach are compared to results obtained with Kaplan-Meier estimate taken as a nonparametric reference. Furthermore, the paper investigates the impact of the stage of the tumor at diagnosis on waiting periods.

Keywords: Term insurance, impaired lives, waiting period, home loan, cancer stage at diagnosis.

5.1 Introduction and motivation

Outstanding balance insurance is generally required by lenders to secure their loans. The borrower is the insured life: if he or she dies before the loan has been fully repaid then the insurer pays a death benefit corresponding to the balance of the loan. As

with any other term life insurance product, applicants with poor health conditions may be denied insurance or charged increased amounts of premium compared to standard conditions. In extreme cases, this may prevent them from accessing property or develop their business project. For this reason, several EU countries passed laws to ease access to mortgage insurance for long-term disease survivors.

The first initiative dates back to 2007, when France launched the AERAS Convention (AERAS is the acronym for "*s'Assurer et Emprunter avec un Risque Aggravé de Santé*" in French, which could be translated as "insuring and borrowing under poor health conditions"). This agreement, signed by the public authorities, banking and insurance sectors, and patients' and consumers' associations purposed to allow people cured from cancer or suffering certain chronic diseases to access insurance comprising benefits in case of death or disability, as well as to guaranteed income insurance. Considering long-term cancer survivors, the AERAS Convention included a "right to be forgotten" (henceforth abbreviated as RTBF), that is, the right for an insurance applicant not to declare a previous cancer after a period of 10 years starting at the end of the therapeutic protocol (reduced to 5 years for pediatric cancers). These periods of 10 and 5 years start from the date of the end of the therapeutic treatment, in absence of relapse within this period.

In Belgium, it was only in 2019 that the RTBF entered the insurance law. Based to a large extent on the reference tables published in the AERAS Convention, a Royal Decree dated May 26, 2019 lists certain types of cancer for which, depending upon entry criteria (such as cancer stage or age), the standard waiting period of 10 years from the end of active treatment is reduced. The RTBF has recently been adapted in Belgium, again following similar changes in France. As from November 2022, the standard waiting period opening the RTBF has been shortened from 10 years to 8 years, and it has been adopted that it will be reduced to 5 years as from January 2025. Moreover, also as from November 2022, the period is shortened to 5 years for cancer survivors who have been cured before the age of 21.

The RTBF has now also been installed in Luxembourg, The Netherlands, Portugal and Romania. It is being debated and advocated for at the European level to expand to the other EU countries as well. There are some ongoing discussions between Insurance Europe, the European Commission and the European Parliament on a possible EU-wide RTBF for cancer survivors. See for instance Scocca and Meunier (2020, 2022).

This paper aims to contribute to this evolution by proposing an actuarially sound methodology to evaluate a technically correct waiting period opening the RTBF. It starts with a critical assessment of the approaches proposed by Soetewey et al. (2021) and by Van Ginckel et al. (2022). It turns out that the results obtained from the method proposed by Soetewey et al. (2021) strongly depend on the extrapolation method. This is precisely shown in this paper, by modifying the extrapolation method and ending up with different waiting periods for some cancer types. This is clearly not acceptable in the context of the RTBF. To be precise, the problem is not with the method proposed by Soetewey et al. (2021) but comes from the limited follow-up period for patients in some cancer registries, including the Belgian one. This requires extrapolation to longer times since diagnosis and this step may induce higher uncertainty. If the length of the follow-up is enough, the method proposed by Soetewey et al. (2021) remains actuarially sound.

We also consider the approach proposed by Van Ginckel et al. (2022), which applies a biostatistical approach based on an arbitrary cut-off of 0.99 for the conditional net survival to propose reduced waiting periods for breast cancer. In Section 5.4.2, we show that, despite the apparent closeness to general population mortality in terms of survival, their method results in one-year death probabilities up to 10 times higher compared to general population at young adult ages. It is clear that such excess

mortality cannot be absorbed by mortgage insurance market without increasing premiums at standard conditions.

There is thus a need for another approach, escaping the problem faced with extrapolation in case of limited follow-up and controlling the resulting premium compared to some market reference. In this paper, we impose a constraint to the premium and transpose it into a target on the conditional observed survival probabilities. The main assumption retained throughout this paper is that mortality for cancer patients temporarily peaks after diagnosis before reverting back to a level comparable to the general population for survivors. We will demonstrate in this paper how the length of the waiting period opening the RTBF can be derived from the comparison of the conditional one-year survival probabilities of cancer patients with the corresponding probabilities at general population level. The main advantage is that the time from which the RTBF can be exercised can be estimated from the available data only, without the need to extrapolate mortality rates beyond 10 years. While cancer stage at diagnosis has not been taken into account in Soetewey et al. (2021), the present paper studies how the length of the waiting period opening the RTBF varies according to the extent of the tumor at diagnosis.

The remainder of this paper is structured as follows. Section 5.2 describes the mortgage insurance product considered in this paper. Section 5.3 presents the data used to perform the present study. Section 5.4 critically assesses the methods proposed by Soetewey et al. (2021) and by Van Ginckel et al. (2022). It is shown that the chosen extrapolation method for limited follow-up impacts on the length of the resulting waiting period opening the RTBF. Our alternative approach is detailed in Section 5.5, and results obtained with our approach are compared with results obtained via a method based on the Kaplan-Meier nonparametric estimator to demonstrate that the proposed approach is trustworthy. Waiting periods are then derived from the comparison between cancer patients' conditional one-year survival probabilities and conditional one-year survival probabilities of the general population. In Section 5.6, we illustrate these analyses considering Belgian data on melanoma, thyroid and female breast cancers according to the stage of tumor at diagnosis. The final Section 5.7 concludes the paper with a discussion.

5.2 Mortage insurance

The proposed approach is based on a representative mortgage insurance contract for the market under consideration. In this paper, we work with a simplified example which could be an appropriate starting point. Precisely, we consider a mortgage insurance applicant aged x borrowing an amount of capital κ at annual interest rate r for a duration n. The capital is reimbursed by constant yearly installments over the n years. The borrower pays the amount

$$\frac{\kappa}{a_{\overline{n}|}}, \text{ where } a_{\overline{n}|} = \sum_{k=1}^{n} \frac{1}{(1+r)^k},$$

back to the lender. Working with annual repayments compared to monthly ones is conservative from the insurer's point of view.

At time s, the amount of the loan that has not yet been amortized is denoted as c_s. Let $\lfloor s \rfloor$ denotes the integer part of $s \in [0, n]$, that is, the largest integer smaller than, or equal to s. At time s, $0 < s \leq n$, the present value of future payments is

$$c_s = c_{\lfloor s \rfloor}(1+r)^{s-\lfloor s \rfloor}$$

where $c_{\lfloor s \rfloor}$ is the outstanding balance of the loan at time $\lfloor s \rfloor$, right after the yearly installment has been paid, given by

$$c_{\lfloor s \rfloor} = \kappa \, \frac{a_{\overline{n-\lfloor s \rfloor|}}}{a_{\overline{n}|}}.$$

The loan is secured by a mortgage insurance, repaying the lender the amount c_s in case the policyholder dies at time s, $0 < s \leq n$. The net single premium is the expected present value (henceforth abbreviated as EPV) of insurance benefits, that is,

$$\pi_0 = \int_0^n {}_s p_x \mu_{x+s} c_s (1+i)^{-s} \mathrm{d}s$$

where i is the technical interest rate for the insurance contract, ${}_s p_x$ is the s-year survival probability for a policyholder aged x at policy issue and μ_{x+s} is the hazard rate, or force of mortality, at attained age $x+s$. Note that hazard rate and force of mortality are the same. For consistency, from now on, hazard rate will always be used in this paper. In accordance with actuarial notation, we denote as p_y the one-year survival probability at integer age y (that is, the probability of being alive at age $y+1$ given that the individual is alive at age y). Note that y is introduced to differentiate from x which refers to the age at policy issue. In this sense, although both are equal in terms of value, y refers to a generic age while x refer to the specific age at policy issue.

Premium calculation is often based on regulatory or experience life tables. In this paper, we consider that standard conditions correspond to premiums computed according to the Belgian regulatory life table XK applying to insurance products comprising benefits in case of death (formally, XK defines minimum premium amount for policies with a positive sum at risk). This life table is widely adopted by Belgian insurers. It is known to be conservative and to generate a relatively high safety loading. Insurers are also allowed to use experience life tables available from the website of the National Bank of Belgium (NBB). These life tables reflect the mortality observed on the market, within portfolios of companies controlled by NBB. There is no safety loading and insurers are only allowed to apply premium rates resulting from NBB tables for relatively short periods of time (5 years, and then rates are subject to revision in case the observed mortality on the market changes over time). In this paper, we only consider the XK life table for premium calculation since these tables can be guaranteed for the whole contract duration and their conservatism better reflects increased mortality levels due to the disease.

The XK life table published in a Royal Decree does not distinguish between male and female policyholders, in accordance with EU anti-discrimination directive. For this reason, the entire analysis is conducted in this paper by pooling male and female mortality data. Also, it only gives one-year survival probabilities p_y at integer ages y. In this paper, we work under piecewise constant hazard rate, assuming that

$$\mu_{y+s} = \mu_y = -\ln p_y \text{ for all } 0 \leq s < 1 \text{ and integer } y.$$

Let us compute π_0 under this assumption. To this end, we split the integral to get

$$\begin{aligned}
\pi_0 &= \sum_{k=0}^{n-1} \int_k^{k+1} {}_s p_x \mu_{x+s} c_s (1+i)^{-s} \mathrm{d}s \\
&= \sum_{k=0}^{n-1} {}_k p_x \int_0^1 {}_s p_{x+k} \mu_{x+k+s} c_k (1+r)^s (1+i)^{-k-s} \mathrm{d}s \\
&= \sum_{k=0}^{n-1} {}_k p_x (1+i)^{-k} c_k \mu_{x+k} \int_0^1 {}_s p_{x+k} (1+r)^s (1+i)^{-s} \mathrm{d}s.
\end{aligned}$$

Now,

$$\begin{aligned}\int_0^1 {}_sp_{x+k}(1+r)^s(1+i)^{-s}\mathrm{d}s &= \int_0^1 \exp\Big(-s\big(\mu_{x+k}-\ln(1+r)+\ln(1+i)\big)\Big)\mathrm{d}s \\ &= \frac{1-\exp\big(-\mu_{x+k}\big)\frac{1+r}{1+i}}{\mu_{x+k}-\ln(1+r)+\ln(1+i)},\end{aligned}$$

so that we finally get

$$\pi_0 = \sum_{k=0}^{n-1} {}_kp_x(1+i)^{-k}c_k\mu_{x+k}\frac{1-\exp\big(-\mu_{x+k}\big)\frac{1+r}{1+i}}{\mu_{x+k}-\ln(1+r)+\ln(1+i)} \tag{5.1}$$

where ${}_0p_x = 1$ and for $k \geq 1$,

$${}_kp_x = \prod_{j=0}^{k-1} p_{x+j} = \exp\left(-\sum_{j=0}^{k-1}\mu_{x+j}\right).$$

5.3 Data

The data available from the Belgian Cancer Registry (BCR) are considered in this paper. The BCR is a national population-based cancer registry collecting data on all new cancer cases diagnosed in Belgium since the incidence year 2004. Cancer registration has been made compulsory by law since 2006 in Belgium. The vital status is derived from linkage with the Belgian Crossroads Bank for Social Security up to April 11, 2022 and quality controls are performed regularly by BCR, ensuring the continuity and completeness of cancer registration in the country. More information can be found on the BCR website, at `www.kankerregister.org`.

To illustrate our work, three cancer types are considered: melanoma (ICD-10 C43), thyroid (ICD-10 C73) and female breast (ICD-10 C50) cancer (only female breast cancer is considered as there are too few registrations for male breast cancer). These three cancer sites have been selected to evaluate the proposed method in different scenarios. Melanoma and thyroid cancer patients are known to have a limited excess hazard compared to the general population (Soetewey et al., 2021). The situation for female breast cancer patients is different with usually a high yearly survival probability in the first years after the date of diagnosis before it eventually decreases due to late cancer recurrences. Moreover, it is known that mortality for patients diagnosed with any of these three cancer types varies with time since diagnosis (Soetewey et al., 2022), yielding appropriate illustrations of the right to be forgotten.

For these applications, our analyses are also limited to patients aged 20 to 69 at time of diagnosis for two main reasons. First, childhood cancers can be seen as a category of cancer on their own, and are often studied separately because they greatly differ from adult cancers. Second, the RTBF mainly concerns young adults and active life.

Out of a total of 161,007 tumors, melanoma, thyroid and breast cancer represent, respectively, 29,213 (18.1%), 12,241 (7.6%) and 119,553 (74.3%) cases diagnosed between 2004 and 2020. Patients were followed-up until April 11, 2022, resulting in a follow-up ranging from 2 to 18 years. Only one record per patient (with the earliest incidence date) within each cancer site was kept for patients with multiple primary diagnoses. A minority of patients without national security number were excluded from the analysis. Patients lost to follow-up (mostly due to moving abroad) and patients still alive at the end of the follow-up period were treated as censored observations. Censoring is assumed to be uninformative.

Sex	Cancer site	Age at diagnosis	Lost to follow-up	Number of included cases	Number of deaths
Men	Melanoma	20-34	3.72%	969	94
		35-49	2.66%	3,266	404
		50-69	1.70%	7,460	1,583
Total				11,695	2,081
Men	Thyroid	20-34	4.10%	366	6
		35-49	3.12%	961	67
		50-69	2.14%	1,773	379
Total				3,100	452
Women	Melanoma	20-34	3.62%	2,488	78
		35-49	1.47%	6,137	382
		50-69	1.35%	8,893	1,112
Total				17,518	1,572
Women	Thyroid	20-34	3.80%	1,607	14
		35-49	2.67%	3,449	107
		50-69	2.06%	4,085	484
Total				9,141	605
Women	Breast	20-34	2.76%	3,112	502
		35-49	1.78%	32,743	4,058
		50-69	1.31%	83,698	15,946
Total				119,553	20,506

Table 5.1: Number of persons diagnosed with melanoma, thyroid and female breast cancer in Belgium between 2004 and 2020 (BCR data) by sex, site and age group, together with the percentage of lost to follow-up and the number of deaths.

Table 5.1 summarizes the number of included cases, number and proportion of deaths and percentage of lost to follow-up before April 11, 2022 per type of cancer, sex and age group. The fraction of patients lost to follow-up per subgroup varied from 1.31% for women with breast cancer aged 50-69 to 4.1% for male thyroid cancer patients aged 20-34. The total fraction of patients lost to follow-up cases, regardless of sex, site or age group was 1.64%. Moreover, mean age at diagnosis was 50.5 years (standard deviation 12.1), 48.1 years (standard deviation 12.4) and 54.6 years (standard deviation 9.5) for melanoma, thyroid and breast cancer, respectively.

For the mortality in the general population, Belgian population life tables are available from Statbel (the Belgian statistical office) and can be freely downloaded from the website `www.statbel.fgov.be`.

5.4 Critical assessment

In this section, we revisit previous studies by Soetewey et al. (2021) and Van Ginckel et al. (2022) to underline their possible shortcomings.

5.4.1 Impact of extrapolation in case of limited follow-up

In this paper, we analyze survival time from diagnosis for cancer patients according to a number of covariates summarized into the vector z. Specifically, T denotes the remaining lifetime at diagnosis, so time from diagnosis to death. Given z, T has probability density function $f(\cdot|z)$, distribution function $F(\cdot|z)$, survival function $S(\cdot|z) = 1 - F(\cdot|z)$, and hazard rate $\lambda(\cdot|z) = f(\cdot|z)/S(\cdot|z)$. Contrarily to insurance

studies, T denotes the remaining lifetime after diagnosis and age at diagnosis is included as a covariate (attained age is thus obtained by summing age at diagnosis and survival time). The link with the international actuarial notation for survival probabilities and hazard rate is as follows: if the insurance applicant aged x has been diagnosed with cancer at age $x - w$ then

$$\begin{aligned} {}_sp_x &= \frac{S(w+s|\text{age at diagnosis} = x-w)}{S(w|\text{age at diagnosis} = x-w)} \\ \mu_{x+s} &= \lambda(w+s|\text{age at diagnosis} = x-w). \end{aligned}$$

Relative survival provides a measure of the excess mortality experienced by cancer patients by comparing the mortality in the cancer population with the mortality in the general population. Relative survival models are divided into additive and multiplicative models. Despite the wide acceptance of multiplicative specifications within the actuarial community, additive models are generally applied in cancer studies. The additive specification is thus adopted in this paper. The hazard rate $\lambda(t|\boldsymbol{z})$ at time t since diagnosis for cancer patients with covariate vector $\boldsymbol{z}$ is decomposed into two additive components: the population hazard based on available patient's characteristics $\boldsymbol{z}$, denoted as $\lambda_P(t|\boldsymbol{z})$, and the excess hazard specific for the disease of interest, denoted as $\lambda_E(t|\boldsymbol{z})$. Formally,

$$\lambda(t|\boldsymbol{z}) = \lambda_P(t|\boldsymbol{z}) + \lambda_E(t|\boldsymbol{z}). \tag{5.2}$$

In (5.2), $\lambda_P(\cdot|\boldsymbol{z})$ usually corresponds to general population life tables. Here, the covariate vector $\boldsymbol{z}$ corresponds to age and it is the same in $\lambda_P(t|\boldsymbol{z})$ and $\lambda_E(t|\boldsymbol{z})$.

Soetewey et al. (2021) adopted the flexible parametric model proposed by Remontet et al. (2019) and Fauvernier et al. (2019a,b) to (i) allow for a flexible modeling of the baseline excess hazard, (ii) account for non-linear and non-proportional effects of covariates and (iii) allow for a flexible interaction between several covariates adopting a multidimensional penalized splines approach. This leads to the specification

$$\ln \lambda_E(t|\boldsymbol{z}) = \sum_{j=1}^{J} g_j(t, \boldsymbol{z}) \tag{5.3}$$

where $g_j(\cdot,\cdot)$ are uni- or multidimensional penalized spline function. This model has the advantage that the splines bring the flexibility needed for modeling the hazard and the penalty terms control this flexibility for smooth estimation. Excess hazard was estimated using the `flexrsurv` package in R (Clerc-Urmès and Grzebyk, 2023), assuming non-linear and non-proportional hazard for age at diagnosis.

Let us proceed as in Soetewey et al. (2021) and determine the waiting period opening the RTBF as the smallest duration after diagnosis so that the expected present value of mortgage insurance benefits gets back to the premium determined according to XK life table. This is represented in Figures 5.1 and 5.2 for a patient diagnosed at the age of 30 and 50, respectively. The left panels are based on the estimation and extrapolation method adopted in Soetewey et al. (2021), implemented in the R package `flexrsurv`. The right panels are based on an alternative method implemented in the R package `rstpm2` (Jakobsen et al., 2020; Liu et al., 2017, 2018; Zhan et al., 2018). For this alternative method, we used a flexible parametric survival model with proportional hazards and 3 degrees of freedom for modelling the baseline log-cumulative hazard. These characteristics have been chosen to obtain excess hazards that are as similar as possible to the ones obtained with the `flexrsurv` package, and other scenarios revealed drastically different patterns for the two considered ages at diagnosis. The resulting waiting periods are listed in Table 5.2. They are obtained by considering that patients become insurable at standard conditions when the EPV

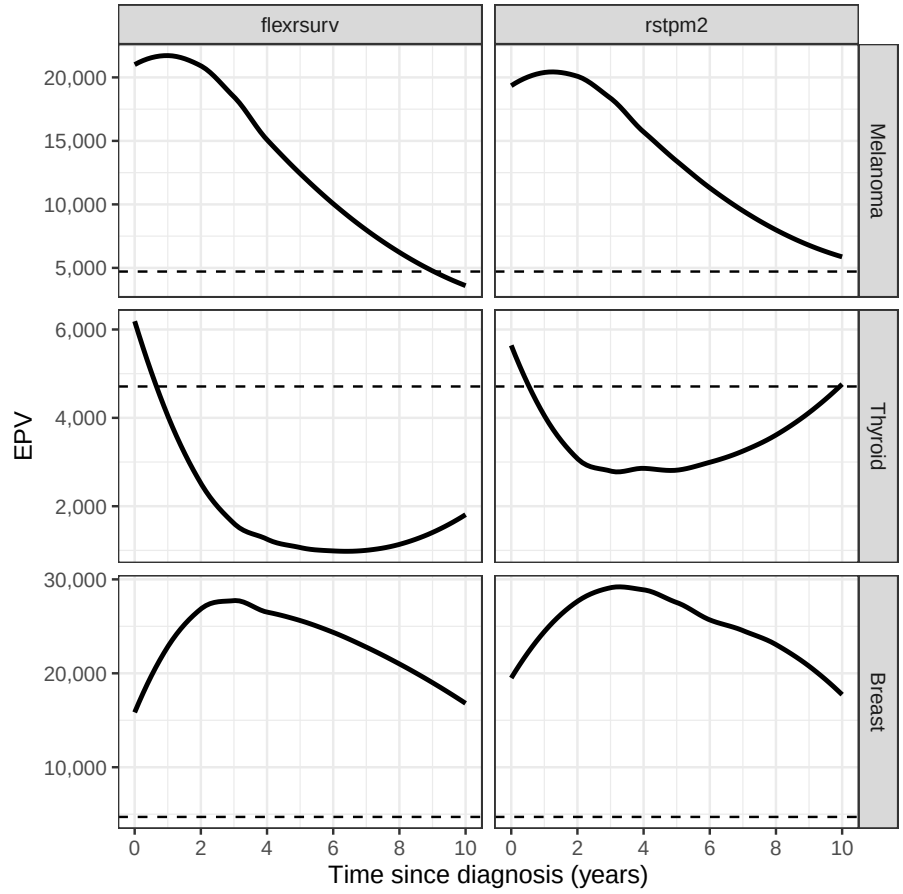


Figure 5.1: Expected present value (EPV) of a life insurance contracted by a 30-year-old cancer patient for a period of 20 years with interest of 1 percent and benefit of 100 000. Horizontal dashed lines correspond to EPV calculated according to XK life table.

reaches the level set by the XK life table. We can see in Table 5.2 that for melanoma and female breast cancer diagnosed at age 50, the waiting periods determined as in Soetewey et al. (2021) are smaller compared to the alternative extrapolation method. For thyroid cancer patients aged 30, the waiting period remains 1 year but EPV exceeds XK level a few years later according to the alternative extrapolation method. For melanoma cancer diagnosed at age 30, the reduced waiting period determined as in Soetewey et al. (2021) is contradicted by the alternative extrapolation method.

This example shows that conclusions may rely to a large extent on the extrapolation method, even more so when one considers different model parameters. This is not acceptable in the context of the RTBF. The aim of this paper is to propose a new approach, only using the available data (so without the need to extrapolate mortality rates beyond 10 years).

5.4.2 Conditional relative net survival

Van Ginckel et al. (2022) applied a pure biostatistical approach based on an arbitrary cut-off of 0.99 for the conditional net survival to propose reduced waiting periods for breast cancer. This section explains why their apparently sound methodology fails to convince actuaries.

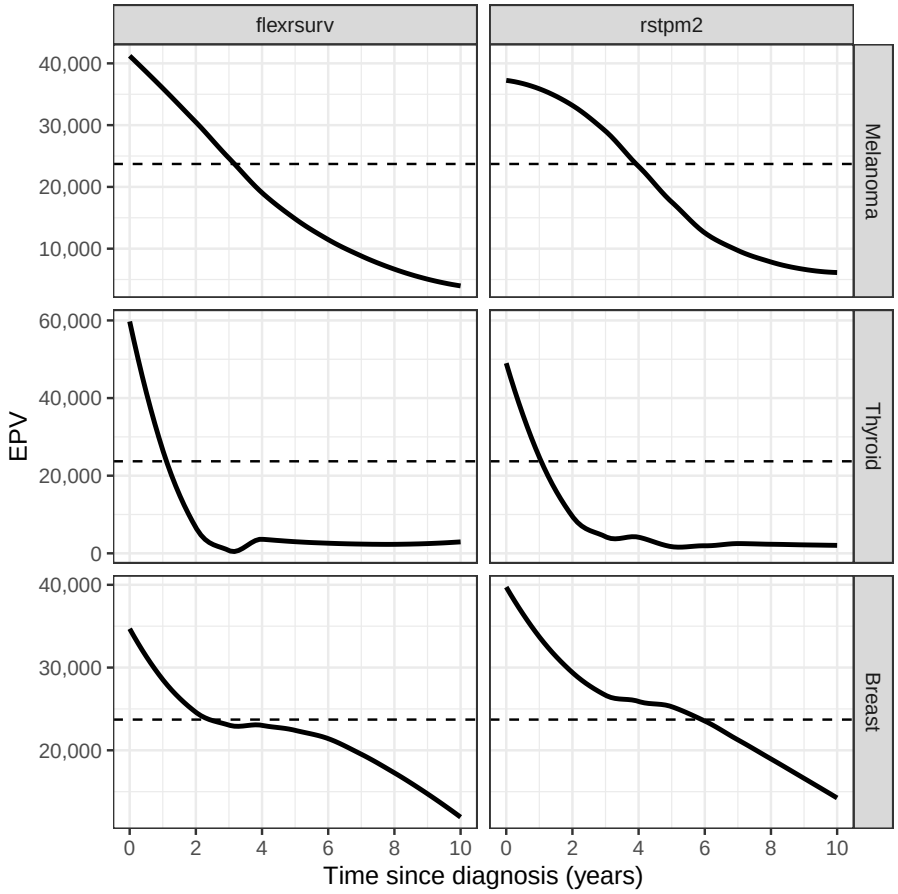


Figure 5.2: Expected present value (EPV) of a life insurance contracted by a 50-year-old cancer patient for a period of 20 years with interest of 1 percent and benefit of 100 000. Horizontal dashed lines correspond to EPV calculated according to XK life table.

Cancer site	Age at diagnosis	Waiting period (in years)	
		Soetewey et al. (2021)	Alternative extrapolation method
Melanoma	30	9	>10
Melanoma	50	3	4
Thyroid	30	1	1*
Thyroid	50	1	1
Breast	30	>10	>10
Breast	50	3	6

Table 5.2: Waiting periods by cancer site and age at diagnosis. The star indicates that EPV does not stay below XK level but start to increase a few years after diagnosis.

In accordance with actuarial notation, let $q_y = 1 - p_y$ be the one-year death probability at age y (that is, the probability of dying before age $y+1$ given that the individual is alive at age y). Probabilities, corresponding to general population mortality, are henceforth denoted as p_y^{NIS} and q_y^{NIS} where "NIS" refers to the National Institute of Statistics (Statbel based in Brussels; `www.statbel.fgov.be`). Likewise, denote as $p_{x,w}^{\text{CR}}$ and $q_{x,w}^{\text{CR}}$ these probabilities for an individual of age x who was diagnosed with cancer w years ago, so at age $x - w$. Here, "CR" refers to Cancer Registry established at national level.

The "conditional relative net survival" referred to in Section 4.2.1 of Van Ginckel et al. (2022) can be interpreted as the ratio $p_{x,w}^{\text{CR}}/p_y^{\text{NIS}}$. The reduced waiting period is then determined as the smallest w such that $p_{x,w}^{\text{CR}}/p_y^{\text{NIS}} > 0.99$. Their argument is that the resulting w ensures that surviving patients' mortality is very close to the general population one. However, when translated into premium calculation, this rule turns out to produce large increases. Indeed, considering that patients can be covered at standard conditions once they have survived w years after diagnosis, with one-year survival probability

$$p_{x,w}^{\text{CR}} = 0.99 p_y^{\text{NIS}} \tag{5.4}$$

means that

$$q_{x,w}^{\text{CR}} = 1 - p_{x,w}^{\text{CR}} = 1 - 0.99 p_y^{\text{NIS}} = q_y^{\text{NIS}} + 0.01 p_y^{\text{NIS}}.$$

Hence, the one-year death probability (driving the amount of premium for a one-year term insurance) is increased by 1% times the corresponding one-year survival probability. The impact of this rule greatly varies according to age x:

- if $q_y^{\text{NIS}} = 0.001$ then this results in an actual one-year death probability

  $$0.001 + 0.01 \times 0.999 = 0.01099$$

  which means that the one-year death probability (and hence the yearly term insurance premium) is multiplied by 10, approximately.

- if $q_y^{\text{NIS}} = 0.01$ then this results in an actual one-year death probability

  $$0.01 + 0.01 \times 0.99 = 0.0199$$

  which means that the one-year death probability (and hence the yearly term insurance premium) is multiplied by 2, approximately.

Considering the typical age range where mortgage insurance is sold, the rule retained by Van Ginckel et al. (2022) allows for mortality levels which largely exceed those corresponding to general population.

5.5 Proposed approach for limited follow-up

Clearly, the rule defining reduced waiting periods for the RTBF must be expressed in terms of premiums. The question about the RTBF centers on evaluating extra claim costs and sharing them among stakeholders in a fair and transparent way. This can only be achieved by computing actual premiums at age x in function of the time w elapsed since diagnosis, and by comparing them to the reference levels XK corresponding to regulatory expected costs.

Let π_0^{XK} be the amount of premium obtained from (5.1) when survival probabilities and death rates correspond to the XK life table. Then, formula (5.1) is used again to

obtain the additive increase in mortality compared to the general population level. Precisely, the additive mortality shift γ is the unique positive root of the equation

$$\pi_0^{\text{XK}} = \sum_{k=0}^{n-1} \exp\left(-\sum_{j=0}^{k-1}\left(\mu_{y+j}^{\text{NIS}}+\gamma\right)\right)(1+i)^{-k} c_k \left(\mu_{y+k}^{\text{NIS}}+\gamma\right) \frac{1-\exp\left(-\left(\mu_{y+k}^{\text{NIS}}+\gamma\right)\right)\frac{1+r}{1+i}}{\mu_{y+k}^{\text{NIS}}+\gamma-\ln(1+r)+\ln(1+i)}. \tag{5.5}$$

The solution is unique because the right-hand side of this equation is increasing in γ and the left-hand side is larger than the right-hand side when $\gamma = 0$ because the XK life table is conservative. The solution is therefore such that $\gamma > 0$.

Following the idea of (5.4), we propose to define the waiting period opening the RTBF as the smallest w such that

$$p_{x,w}^{\text{CR}} = \exp(-\gamma) p_y^{\text{NIS}}. \tag{5.6}$$

In this case, we recover a constraint on the conditional observed survival, but with the arbitrary 0.99 level replaced with $\exp(-\gamma)$ controlling premium. Following (5.6), the waiting period opening the RTBF is determined as the smallest w such that $p_{x,w}^{\text{CR}}/p_y^{\text{NIS}} > \exp(-\gamma)$. To apply this rule, p_y^{NIS} can easily be found within Belgian population life tables, available from Statbel. The calculation of $p_{x,w}^{\text{CR}}$ is explained in Appendix A.2.

Let us now apply this method to get the length of the waiting period opening the RTBF. To this end, survival probabilities of cancer patients, obtained via a flexible parametric model (using the `mexhaz` R package (Charvat and Belot, 2021) and based on a baseline hazard specified as the exponential of B-splines of degree 2 with a knot at 2.5 years of follow-up), are first compared with the observed survival probabilities obtained with the nonparametric Kaplan-Meier (1958) estimator. The results are displayed in Figure 5.3. It can be seen from Figure 5.3 that observed survival curves obtained via a flexible parametric model and via the Kaplan-Meier estimator are very similar for all cases under consideration (i.e., for both ages at diagnosis and for all three cancers of interest).

Secondly, conditional one-year observed survival probabilities, obtained via a flexible parametric model (denoted $p_{x,w}^{\text{CR}}$ and detailed in Section A.2) are compared with the conditional one-year observed survival probabilities obtained based on the Kaplan-Meier estimator (henceforth denoted as $p_{x,w}^{\text{KM}}$) in Figure 5.4. Here, probabilities $p_{x,w}^{\text{KM}}$ are computed with increments of 0.1 year and are referred as the KM-based method in the remainder of the text since these probabilities are computed based on the Kaplan-Meier estimator. The difference with a standard Kaplan-Meier estimator is that $p_{x,w}^{\text{KM}}$ correspond to conditional one-year observed survival probabilities (instead of simply observed survival probabilities). In practice, $p_{x,w}^{\text{KM}}$ are found by computing one-year survival probabilities using the `survival` R package (Terry M. Therneau and Patricia M. Grambsch, 2000), repeatedly for each subgroup of patients who survived at least $0, 0.1, 0.2, \ldots, 10$ years since diagnosis. The advantage of computing $p_{x,w}^{\text{KM}}$ this way is that the provided confidence intervals are usable, which is not the case if $p_{x,w}^{\text{KM}}$ are computed by dividing the survival probability at a given time by the survival probability one year earlier). The goal of comparing $p_{x,w}^{\text{CR}}$ with a counterpart based on a nonparametric reference such as the Kaplan-Meier estimator is to demonstrate that results obtained with the proposed approach are trustworthy.

Figure 5.4 shows that conditional one-year survival probabilities obtained via the flexible parametric model follow globally the same trend than the ones obtained via the KM-based method for all scenarios, except for the first year after diagnosis for

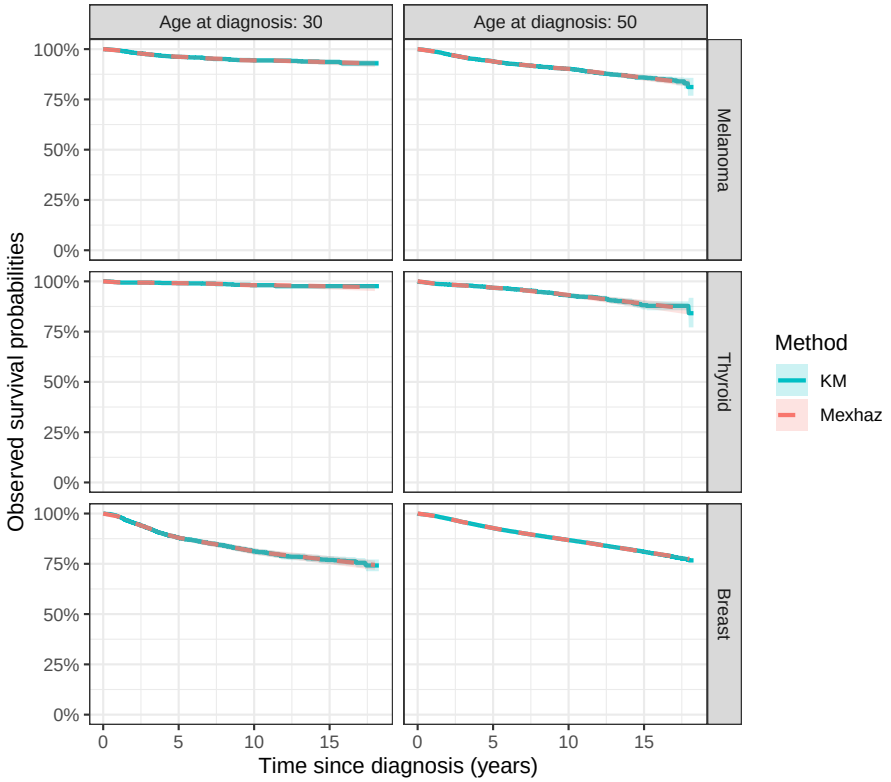


Figure 5.3: Survival probabilities (with 95% confidence interval) by cancer site and age at diagnosis. Mexhaz method corresponds to the probabilities obtained via a flexible parametric model (dashed line), whereas KM-based method corresponds to the ones obtained based on the nonparametric Kaplan-Meier estimator (solid line).

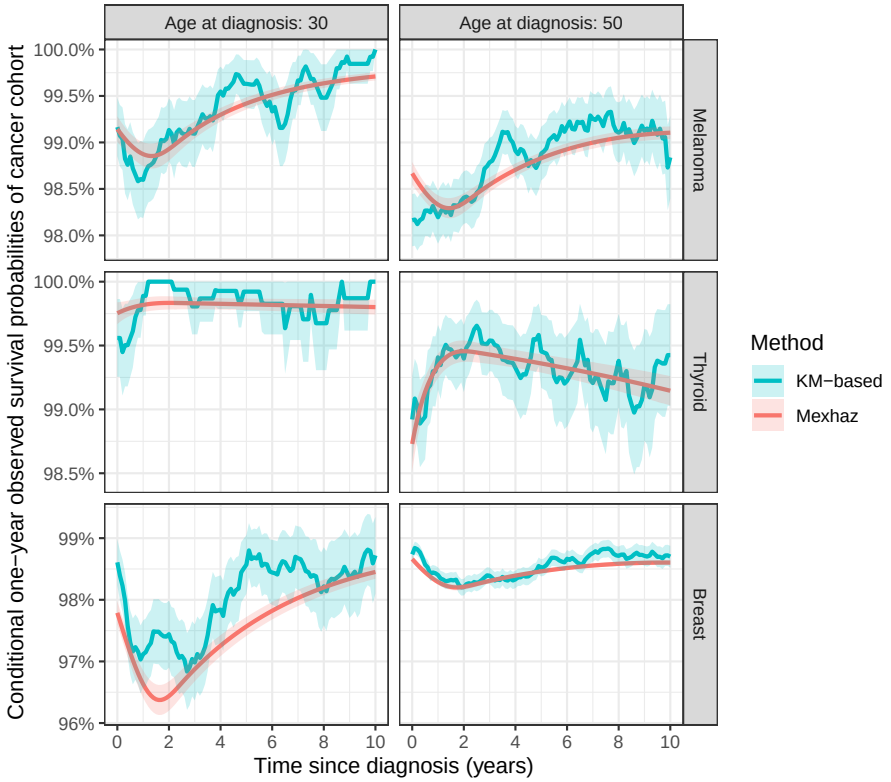


Figure 5.4: Conditional one-year survival probabilities (with 95% confidence interval) by cancer site and age at diagnosis. Mexhaz method corresponds to the probabilities obtained via a flexible parametric model, $p^{\mathrm{CR}}_{x,w}$, whereas KM-based method corresponds to the ones obtained based on the nonparametric Kaplan-Meier estimator, $p^{\mathrm{KM}}_{x,w}$.

Cancer site	Age at diagnosis	Waiting period (in years) Our approach	KM-based
Melanoma	30	> 10	10
Melanoma	50	6	6
Thyroid	30	1	Uncertain
Thyroid	50	1	2
Breast	30	> 10	> 10
Breast	50	> 10	> 10

Table 5.3: Waiting periods by cancer site and age at diagnosis computed via our approach and via the KM-based method.

patients diagnosed with melanoma cancer at age 50. We consider that conditional one-year observed survival probabilities are reasonably well estimated with our approach when compared to a nonparametric reference. Furthermore, for a given sample size, confidence intervals are narrower with our approach compared with the nonparametric Kaplan-Meier, a reason to prefer the new approach over the nonparametric reference.

To determine the waiting period opening the RTBF, conditional one-year survival probabilities obtained via the two approaches, that is, $p_{x,w}^{\text{CR}}$ and $p_{x,w}^{\text{KM}}$, are divided by the conditional one-year survival probabilities in the general population, that is, p_y^{NIS}. Results are displayed in Figure 5.5. For the sake of comparison, the waiting period opening the RTBF is determined as the smallest w such that $p_{x,w}^{\text{CR}}/p_y^{\text{NIS}} > \exp(-\gamma)$ or such that $p_{x,w}^{\text{KM}}/p_y^{\text{NIS}} > \exp(-\gamma)$. Dividing $p_{x,w}^{\text{CR}}$ and $p_{x,w}^{\text{KM}}$ by p_y^{NIS} allows the comparison with the additive correction $\exp(-\gamma)$. Here, $\gamma = 0.0014$ at age 30 and $\gamma = 0.0063$ at age 50. Notice the difference with the threshold of 0.99 set in Van Ginckel et al. (2022), as $\exp(-\gamma)$ equals 0.998601 and 0.9937198 for a patient diagnosed at age 30 and 50, respectively. Also note that, for patients aged 30 years at diagnosis, $p_{x,w}^{\text{KM}}$ is actually computed based on patients aged between 25 and 35 years at diagnosis. For patients aged 50 years at diagnosis, $p_{x,w}^{\text{KM}}$ is computed based on patients aged between 45 and 55 years at diagnosis. This is to include more patients and thus have more stable estimates. Indeed, samples of patients diagnosed at exactly 30 and 50 years old have a limited size, in particular for thyroid cancer. Considering patients aged from 25 to 35 and from 45 to 55 instead of patients of exactly 30 and 50 years old does not undermine our analyses, as patients within each age group are very similar in terms of survival.

Results are displayed in Table 5.3 and Figure 5.5. Remember that an increasing ratio is a sign of better prognosis for cancer patients. On the other hand, a decreasing ratio is a sign that survival for cancer patients declines over the years since diagnosis, so a sign of worse prognosis compared to the general population. Following this, and in order to be as conservative as possible, if the ratio of conditional one-year survival probability reaches the level of the additive correction more than once within the 10-year period after diagnosis, the waiting period is set as the largest time after diagnosis where the ratio of survival probabilities crosses the additive correction level. Notice also the emergence of small jumps when plotting the ratio of the conditional survival probabilities in Figure 5.5 resulting from the division by the one-year survival probabilities p_y^{NIS} in the general population.

From Table 5.3 and Figure 5.5, we can see that waiting periods are below 10 years for melanoma cancer patients aged 50 years at diagnosis, and thyroid cancer patients aged 30 and 50 at diagnosis. Waiting periods obtained via the KM-based method are below 10 years for melanoma and thyroid cancer patients aged 50 at the time

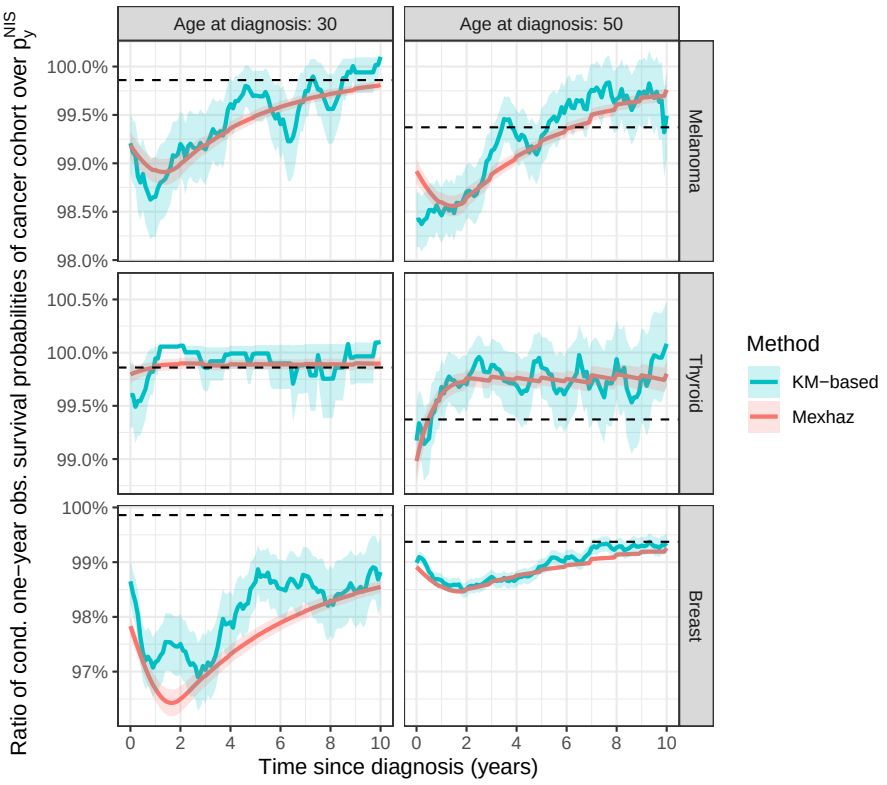


Figure 5.5: Ratio of conditional one-year survival probability (with 95% confidence interval) by cancer site and age at diagnosis, together with the additive correction $\exp(-\gamma)$ (horizontal dashed lines). Mexhaz method corresponds to $p_{x,w}^{\mathrm{CR}}/p_y^{\mathrm{NIS}}$, whereas KM-based method corresponds to $p_{x,w}^{\mathrm{KM}}/p_y^{\mathrm{NIS}}$.

of diagnosis. For all other scenarios, waiting periods are equal or above 10 years after diagnosis. When comparing the two approaches, waiting periods are relatively equivalent for all considered subgroups except for thyroid cancer patients diagnosed at 30 years old, who have an uncertain waiting period via the KM-based method (since the ratios of conditional one-year survival probabilities fluctuate around the level set by the additive correction from 0 to 10 years after diagnosis). Moreover, breast cancer patients diagnosed at age 30 have a waiting period above 10 years while it is equal to 10 years according to the KM-based calculation. Recall that, as advocated in Soetewey et al. (2021), these waiting periods start at the time of diagnosis, and not at the end of the therapeutic protocol as it is the case with the current legislation.

5.6 Impact of the stage of the tumor

One could argue that mortality and thus the waiting period opening the RTBF varies between cancer patients diagnosed at different tumor stages. As this information is available in BCR, this section refines the preceding analyses by cancer stage at diagnosis.

Information on both the clinical and pathological staging has been combined to define a final tumor stage. First, clinical staging is an estimate of the extent of the cancer based on results of physical exams, imaging tests, endoscopy exams, biopsies, and for some cancers, the results of other tests, such as blood tests. Second, the pathological staging (also called the surgical stage) is an estimate of the extent of the cancer that is based on the results of pathological examination of the resection piece after surgery. In some cases, the pathological stage is different from the clinical stage, for instance, if the surgery shows the cancer has spread more than was seen on imaging tests. A common practice is to combine these two methods to obtain a so-called combined stage. When the pathological stage is known, it is taken as combined stage, unless there is clinical evidence of metastasis. In case the pathological stage is unknown, the clinical stage is retained. Combining the clinical and pathological stage limits missing values (missing combined stage appears only when both the clinical and pathological stages are missing). This combined stage is considered in this section.

Stages I, II, III and IV were considered. Tumors with an unknown stage at the time of diagnosis, representing 7.5% of all tumors, have been ignored. Number of included cases, number of observed deaths, one-year and 5-year observed survival probabilities (obtained with the nonparametric Kaplan-Meier estimator) by cancer site and stage of the tumor are displayed in Table 5.4. Given the small number of observations for stages III and IV, these two stages have been combined for the analyses. Furthermore, to ensure simplicity and given that cancer patients diagnosed at stages I and II are relatively similar in terms of survival, these two stages have also been combined.

The present section is aimed at studying the appropriateness of stratifying the RTBF according to the stage of the tumor at diagnosis: waiting periods are computed separately for patients diagnosed at stages I–II and at stages III–IV using our proposed approach. This will serve as a comparison with results obtained before, where all stages are included. Note that, as the additive correction $\exp(-\gamma)$ depends only on age at diagnosis, it differs between patients diagnosed at 30 and 50 years old, but it is the same for all stages and it remains the same than when including all stages. Notice that the nonparametric Kaplan-Meier reference is no longer used because stratifying by stage reduces drastically the number of observations, in particular for stages III and IV. This rises the issue of the accuracy of the Kaplan-Meier estimator, and therefore reduces its usefulness in the context of the RTBF.

Results of the stratification by stage at diagnosis are displayed in Table 5.5 and Figure 5.6. Waiting period is lower for patients diagnosed at stages I–II compared

Cancer site	Tumor stage	Number of cases	Number of deaths	1-year survival prob. (95% CI)	5-year survival prob. (95% CI)
Melanoma	I	20,949	1,153	0.997 (0.996-0.997)	0.970 (0.968-0.973)
	II	3,019	777	0.980 (0.975-0.985)	0.802 (0.786-0.817)
	III	1,669	537	0.949 (0.939-0.960)	0.711 (0.688-0.735)
	IV	444	327	0.658 (0.615-0.703)	0.298 (0.257-0.345)
Thyroid	I	8,071	311	0.995 (0.994-0.997)	0.979 (0.976-0.982)
	II	946	73	0.989 (0.983-0.996)	0.961 (0.948-0.974)
	III	841	134	0.993 (0.987-0.999)	0.942 (0.926-0.958)
	IV	617	294	0.779 (0.747-0.812)	0.625 (0.587-0.665)
Breast	I	56,039	4,813	0.996 (0.995-0.996)	0.965 (0.964-0.967)
	II	37,935	5,979	0.992 (0.991-0.993)	0.926 (0.923-0.929)
	III	11,919	3,922	0.976 (0.973-0.979)	0.805 (0.798-0.813)
	IV	5,639	3,839	0.829 (0.819-0.839)	0.390 (0.377-0.404)

Table 5.4: Number of included cases, number of observed deaths, one-year and 5-year observed survival probabilities (with 95% confidence interval) by cancer site and stage of the tumor. Survival probabilities are obtained with the nonparametric Kaplan-Meier estimator.

to patients diagnosed at stages III–IV for all scenarios. In particular, compared to patients diagnosed at all stages, when including only patients diagnosed at stages I–II, waiting periods are reduced from 6 to 4 years for melanoma cancer patients aged 50, reduced from 1 to 0 year for thyroid cancer patients aged 50, and reduced from more than 10 years to 7 years for female breast cancer patients aged 50. For melanoma cancer patients aged 30, thyroid cancer patients aged 30 and breast cancer patients aged 30, waiting periods remain the same whether it is calculated by stage or for all stages combined. This shows that, for the three cancer sites considered, stratifying the analyses according to the stage has no impact on the waiting periods for patients diagnosed at the age of 30, but has an impact for patients diagnosed at the age of 50. This can be partly explained by the fact that, among patients diagnosed at a young age, a small proportion is diagnosed at stages III–IV. For instance, only 8.36% of patients aged 30 or below at the time of diagnosis are diagnosed at stages III–IV. Furthermore, we observe that the waiting period is above 10 years for patients diagnosed at stages III–IV for all scenarios.

Notice that 95% confidence intervals for stages I–II (Figure 5.6) are narrower than when all stages are considered (Figure 5.5), although the sample size is smaller when including only patients diagnosed at stages I–II. This is explained by the fact that, for the same sample size, standard errors for probabilities closer to 0% or 100% are smaller. Therefore, even though the sample size is smaller for stages I–II than for all stages combined, confidence intervals are narrower because probabilities are closer to 100% for this subgroup of patients. Also notice that confidence intervals do not widen with time since diagnosis, contrarily to what would be expected given that the sample size decreases with time since diagnosis. The following elements explain this phenomenon. We only consider age at diagnosis 20-69 years. Survival is high for this age range and the cancer types considered, so 10 years after diagnosis will not yet be long enough to see a clear increase in the length of the confidence intervals. And again, when survival probabilities approach 100%, the confidence intervals become smaller for a given number of observations. A similar pattern for the conditional net survival has been found in Van Ginckel et al. (2022) for female breast cancer. Calculations of the confidence intervals are further explained in Appendix A.2.

Cancer site	Age at diag.	Proposed approach			Soetewey et al. (2021)
		Stages I–II	Stages III–IV	All stages	All stages
Melanoma	30	> 10	> 10	> 10	9
Melanoma	50	4	> 10	6	3
Thyroid	30	1	> 10	1	1
Thyroid	50	0	> 10	1	1
Breast	30	> 10	> 10	> 10	NA
Breast	50	7	> 10	> 10	NA

Table 5.5: Comparison of waiting periods by cancer site and age at diagnosis resulting from our approach and from Soetewey et al. (2021).

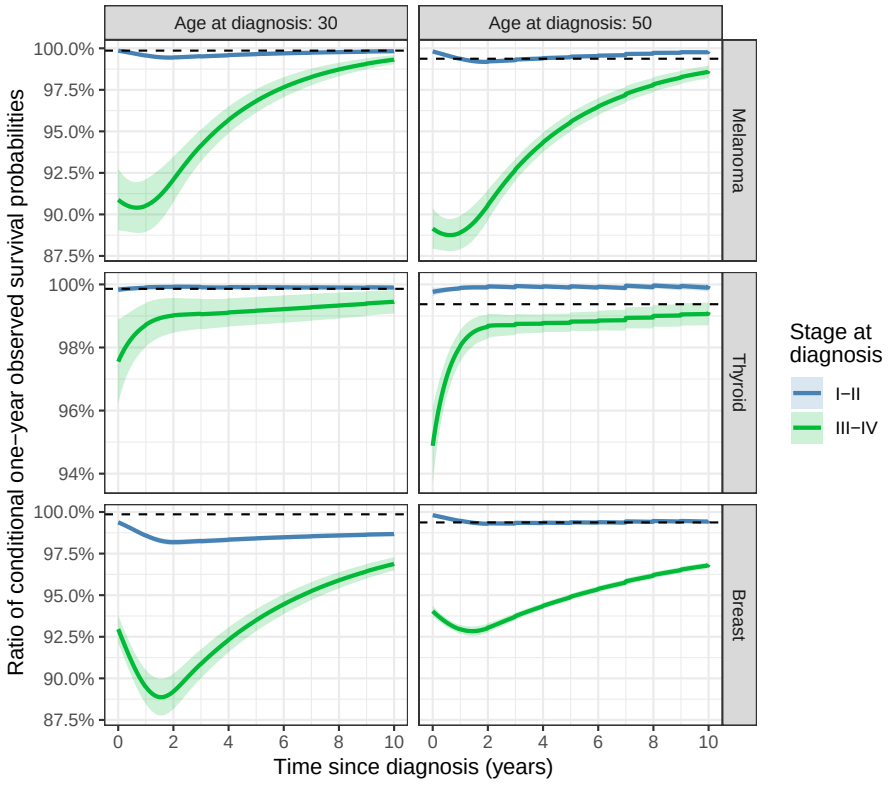


Figure 5.6: Ratio of conditional one-year survival probability obtained via the proposed approach (i.e., $p^{\text{CR}}_{x,w}/p^{\text{NIS}}_{y}$) by cancer site, age and stage at diagnosis. Horizontal dashed lines correspond to the additive correction $\exp(-\gamma)$.

5.7 Discussion

To sum up, let us compare waiting periods obtained with our approach with results obtained according to the method proposed by Soetewey et al. (2021), which are based on the time after diagnosis when the expected present value of a standard mortgage insurance reaches the same level than the one based on XK life table. A summary is displayed in Table 5.5. We can see that waiting periods are sensibly the same for thyroid cancer patients across all methods, while they are slightly higher when estimated via the approach proposed in this paper for melanoma cancer patients. Note that no comparison is made for breast cancer, as this cancer site was not considered in Soetewey et al. (2021).

Results in Table 5.5 are in line with the reduced waiting periods specified in the Belgian legislation. Furthermore, results are also in line with the AERAS convention (i.e., the reference grid used in France), which stipulates that the RTBF is maximum 6 years after the end of the therapeutic protocol for melanoma and thyroid cancers.

Nonetheless, an important difference is that in this paper, all waiting periods opening the RTBF are based on the time since diagnosis, rather than on the time since the end of the therapeutic protocol as currently implemented in the Belgian and French reference grids. As duration of cancer treatments are unpredictable and heterogeneous (even within the same cancer site and stage), a RTBF based on the date of diagnosis rather than based on the treatment end date will benefit both patients and insurers. Indeed, patients will know exactly when they can expect to benefit from this RTBF, and insurers will face less uncertainties (as the date of diagnosis is known and fixed, contrarily to the treatment end date which is difficult to establish and may change over time depending on the patient's health status) and less prone to debates (about, for instance, what is considered as treatment or not).

Although data used in the analyses cover a relatively long period of time (year of diagnosis ranges from 2004 to 2020) with diagnostic criteria and methods that have evolved and improved over that period, calendar time has not been included for two main reasons. First, the limited number of cases available (in particular since the focus is on young adults) prevents another division between different cohorts. Second, given that medicine and treatments progress with time, survival of cancer patients also improve with time. Thus, the resulting potential bias of omitting a cohort effect appears to be conservative, as the actual time for the patients to reach a survival comparable to that of the general population will decrease with improving treatments. In addition to that, population data are used whereas outstanding balance insurance applicants belong to the upper socio-economic class who usually have better prognosis, and individuals who contract a home or professional loan are generally in good health as individuals with poor health are unlikely to embark on such a project. These selection effects imply that analyses conducted in the present paper are conservative in many respects.

One could argue that waiting periods are expected to be shorter for patients diagnosed at stages III–IV compared to patients diagnosed at stages I–II, as we would expect when comparing patients diagnosed with pancreatic and breast cancer. The idea behind this reasoning is that the worse the prognosis, the quicker the patients die after diagnosis and thus the quicker only the survivors remain. Results of the stratification by stage show that it is not the case. The following arguments explain it. Statistical cure in the case of female breast cancer is not yet achieved within 15 years after diagnosis (except for stage I), while it is achieved for pancreatic cancer at around 5 years after diagnosis. Indeed, for female breast cancer, excess hazard is relatively constant and non negligible even after many years after diagnosis, with late recurrences occurring up to 20 years after diagnosis. On the contrary, for aggressive cancers, excess hazard is much less constant over the years after diagnosis, and in

the case of pancreatic cancer it becomes negligible around 5 years after diagnosis. Given the difference in excess mortality between breast and pancreatic cancer, it is reasonable to expect waiting periods to be shorter for pancreatic than for breast cancer. Although melanoma and thyroid cancers are nowhere near as aggressive as pancreatic cancer, the trend of the excess hazard for these two cancers is closer to pancreatic than to breast cancer, that is, excess hazard is not constant over the years after diagnosis, it becomes negligible only after some years after diagnosis and late recurrences are rare. This explains the shorter waiting periods for melanoma and thyroid cancers compared to female breast cancer. The same reasoning can be applied to the comparison of the waiting periods between stages of the tumor. One could expect that the more advanced the stage, the more quickly only the survivors remain and thus the shorter the waiting period. This holds only if statistical cure is reached at a given time after diagnosis (and in particular within 10 years after diagnosis to argue for a reduced waiting period opening the RTBF). For female breast cancer, the excess hazard for stages III–IV is higher than for stages I–II up to 15 years after diagnosis, resulting in waiting periods that are not shorter for stages III–IV compared to stages I–II.

Results obtained in the present study focus on melanoma, thyroid and female breast cancer patients for illustrative purposes. The approach developed in this paper can be applied to other cancer types or diseases. However, as just discussed, it cannot be used in case of late recurrences nor to chronic diseases to argue a shorter waiting period. For some cancers with late recurrences such as breast cancer, the waiting period resulting from our proposed approach when including patients diagnosed at all stages of the tumor is (much) longer than if it was proposed only to patients diagnosed at stages I–II. For melanoma and thyroid cancers, the waiting period resulting from our proposed approach when including patients diagnosed at all stages of the tumor is relatively similar than if it was proposed only to patients diagnosed at stages I–II. This demonstrates that, besides the fact that computing the waiting period should be done by stage for female breast cancer while it is not compulsory for melanoma and thyroid cancers, cancer is not one disease, but a family of many diverse diseases with different outcomes. Therefore, the proposed method should be applied on a case-by-case basis, that is, cancer by cancer. This is left for future research.

As mentioned in Section 5.1, the method proposed by Soetewey et al. (2021) remains actuarially sound if the length of the follow-up is long enough. It could be argued, however, that even when registry data have a sufficiently long follow-up period, the method proposed in this paper would still be preferable since a long follow-up means that some patients have been diagnosed a long time ago, and are thus not treated as well as nowadays. This argument is all the more valid the longer the follow-up time, as the longer the follow-up, the greater the potential increase in treatment efficacy between the beginning and end of the follow-up period.

The proposed approach can obviously be applied in other countries by replacing the databases by the appropriate ones. Moreover, other cancer sites and other diseases which qualify for the RTBF (e.g., HIV, some types of hepatitis and leukemia) are left for future research. This would undeniably be useful to improve the reference grids in Belgium and other countries, and ultimately, to improve access to such insurance products for other types of surviving patients.

Conclusion 6

In this thesis, we proposed new statistical methods pertaining to survival analysis and actuarial sciences. The objective of this work was mainly to propose new and effective tools to address some central questions arising in biostatistics and actuary, and to illustrate, using Belgian data, these methods in the context of the right to be forgotten in insurance. The remainder of the conclusion is structured as follows. We first propose a general discussion and briefly recall the major points and contributions of our studies. Certain inquiries arising from our research remain unresolved. We conclude this PhD thesis by delving into several of these questions, along with proposing potential directions for future investigation.

6.1 General discussion

The current research initiated from the premise that a growing number of patients having survived cancer faced difficulties to access mortgage insurance securing home loan. This problem was all the more acute as, thanks to medical advances, more and more patients were surviving their cancer. In response to this observation and backed by studies showing that mortality of some cancer types was indeed close to that of the general population after a few years after a successful treatment, France took the lead by introducing, in 2016, the right to be forgotten in the context of insurance. Noting the same phenomenon, and driven by the desire to reduce financial and emotional discrimination against cancer survivors, the right to be forgotten began to take effect in Belgium in 2019. Although the implementation of this right was already a major step forward for cancer survivors at that time, we believed that there was still room for improvement. Following this, through our research, we aimed at studying and improving the right to be forgotten in the context of insurance. In particular, we investigated the waiting period opening this right but starting from diagnosis, and with a focus on mortgage insurance issued to cancer survivors. This was done through several complementary approaches.

First, based on the underlying assumption that some patients who had cancer in the past exhibit a survival comparable to that of cancer-free individuals, we started our research by estimating the time after diagnosis after which melanoma and thyroid cancer patients could have access to a mortgage insurance at the same rate than cancer-free applicants. This time was based on when, after diagnosis, the expected present value of a standard mortgage insurance reached the same level than the one expressed by means of regulatory life tables. This was the main goal of Chapter 2. Through this chapter, it has been shown that the time from diagnosis after which melanoma and thyroid cancer patients can be covered at standard premium rate (and hence, have access to the right to be forgotten) is relatively short. More precisely, it has been demonstrated that, for thyroid cancer patients, a waiting period of 4 years after diagnosis is enough for 30-year-old cancer patients, and a waiting period of 3 years after diagnosis is enough for 50-year-old patients. Moreover, a waiting period of 8 years after diagnosis is sufficient for 50-year-old melanoma cancer patients. A major

contribution is that we promote a waiting period opening the right to be forgotten starting at the time of diagnosis instead of starting after a successful treatment. The date of diagnosis, registered in the national cancer registry, being much less ambiguous and subject to debates than the treatment end date (in particular given that duration of treatments can vary considerably and treatment success may affect the end date), this limits uncertainties for all parties and reduces disagreements in case of death. These findings align with the reduced waiting period outlined in Belgian legislation, where thyroid and melanoma cancer patients gain access to the right to be forgotten within a relatively short time frame; maximum 3 and 6 years after the end of the therapeutic protocol for thyroid and melanoma cancers, respectively. Note that it is undeniable that including the gender would have refined the analysis, as considerable gaps are observed between women and men in terms of survival for some cancer types. However, gender has been ignored on purpose considering the 2012's European directive on equality between women and men, stipulating that gender can no longer have an influence on the premiums nor on the coverage conditions of outstanding balance insurances.

Second, we continued our research based on the observation that the right to be forgotten was very binary, in the sense that an individual had access to it either entirely or not at all. Therefore, the goal was to develop financial products that allowed patients to be covered while waiting for the right to be forgotten to take effect, these products obviously being subject to additional premiums or adapted coverage conditions that reflect the aggravated risk the individuals represent. Towards that end, Chapter 3 introduced several insurance covers, with one of them granting access to a mortgage insurance during the waiting period opening the right to be forgotten. This product is especially important at young ages to guarantee access to property and home ownership to cancer patients whose health status has improved but who cannot benefit from the right to be forgotten because the waiting period is not exhausted yet. It has been shown that insurance products can be developed to address the particular needs of patients during the waiting period opening the right to be forgotten, but that costs greatly vary according to cancer type. More precisely, suppose a cover option which offers the policyholder the option to obtain mortgage insurance at standard conditions even if he or she has been diagnosed with cancer, but which cannot be executed before the end of a waiting period of 2 years starting from the date of diagnosis. It has been demonstrated that this product is not needed for thyroid cancer patients, as these patients can be covered at standard rates after the 2-year waiting period. Furthermore, it has been shown that the cost appears to be moderate for melanoma cancer patients, implying that such product could be proposed by insurance providers.

Third, we decided to look at the problem from another angle. The idea was to quantify the number of years of life a cancer cohort loses during the repayment period of a mortgage insurance. Given that a mortgage insurance has a finite horizon, any year of life lost due to cancer (i.e., in addition to the number of years the general population would have lost due to population mortality) during this repayment period is a year during which the client does not pay back the loan, whereas any year of life lost due to cancer after the loan has been fully reimbursed is not seen as a lost year from the insurer's perspective. Towards that end, the aim was to quantify the risk of developing cancer for a healthy individual and the number of years of life lost due to cancer given that the patient already survived some years after diagnosis. We expect that results can then easily be linked with the potential financial losses for insurance providers, which would then be able to assess whether or not the the mutualization mechanism could take effect for the studied cancers. Chapter 4, which thus focused on estimating the risk of developing cancer for a healthy individual and the number of years of life lost due to cancer given that the patient already survived some years

after diagnosis, conveyed two key messages. First, the probability of being diagnosed with cancer over a 20-year period remain below 1% for melanoma and thyroid cancers for both sexes. Second, for melanoma and thyroid cancer patients diagnosed between the age of 20 and 70 years, once they have survived their cancer for 10 years, the number of years of life lost before the age of 70 due to cancer remains below one year. For women diagnosed with breast cancer, once they survived 10 years after diagnosis, the number of years of life lost before the age of 70 due to cancer remains below 2 years. This indicates that, up to the age when most people have finished paying off their loan, melanoma, thyroid and female breast cancer patients who survived their cancer for at least 10 years after diagnosis lose a limited number of years of life due to their cancer compared to that of the general population. Furthermore, based on the age at which the number of years of life lost due to cancer start to decrease over the years, it has also been shown that, no matter whether the patient survived 0, 5 or 10 years after diagnosis, melanoma cancer patients approach the same mortality risks as of the general population (i.e., excess mortality decreases) from the age of 45. Thyroid cancer patients approach the general population mortality risks from the age of 50, whereas female breast cancer patients observe a decreasing excess mortality from the age of 20.

Fourth, Chapter 5 extended results found in Chapter 2 in two ways. First, the method proposed in Chapter 2 to determine the waiting period opening the right to be forgotten required data up to 30 years after diagnosis, or extrapolating results up to that time horizon. When survival data are only available over a shorter duration (which is likely with recent cancer registries), it turned out that results may strongly depend on the extrapolation method chosen. This is why an alternative method has been proposed, based on a constraint imposed to the premium, which is then transposed into a target on the conditional observed survival and the waiting period follows. A second extension was that the impact of the stage of the tumor at diagnosis on waiting periods has been investigated. This alternative method has shown that, when all stages of the tumor were considered, the waiting periods opening the right to be forgotten were sensibly the same for thyroid cancer (1 year after diagnosis for 30 and 50-year-old patients), while they were slightly higher when estimated via the alternative approach for melanoma cancer (more than 10 years, and 6 years after diagnosis for, respectively, 30 and 50-year-old patients). Results have shown to be still aligned with the waiting periods specified in the Belgian legislation (maximum 6 years after the end of the therapeutic protocol for melanoma and thyroid cancers). When analyses were stratified by the stage of the tumor at diagnosis, results have shown that the waiting period is lower for patients diagnosed at stages I–II compared to patients diagnosed at stages III–IV for all scenarios (the two considered ages at diagnosis and the three cancers of interest). In particular, compared to patients diagnosed at all stages, when including only patients diagnosed at stages I–II, waiting periods are reduced from 6 to 4 years for melanoma cancer patients aged 50, reduced from 1 to 0 year for thyroid cancer patients aged 50, and reduced from more than 10 years to 7 years for female breast cancer patients aged 50. For melanoma cancer patients aged 30, thyroid cancer patients aged 30 and breast cancer patients aged 30, waiting periods remain the same whether it is calculated by stage or for all stages combined. This shows that, for the three cancer sites considered, stratifying the analyses according to the stage has no impact on the waiting periods for patients diagnosed at the age of 30, but has an impact for patients diagnosed at the age of 50 (partly explained by the fact that, among patients diagnosed at a young age, a small proportion is diagnosed at stages III–IV). Furthermore, it has been shown that the waiting period is above 10 years for patients diagnosed at stages III–IV for all scenarios.

The methodologies and illustrations presented in this thesis collectively align towards the shared objective of reducing the waiting period opening the right to be

forgotten until it more closely reflects the real risk posed by patients suffering or having suffered from cancer. Based on the results provided in the present thesis, we also believe that this objective involves starting the waiting period from the date of diagnosis.

As of today, eight European Member States (France, Belgium, The Netherlands, Portugal, Romania, Spain, Cyprus and Italy) out of the 27 have implemented a legal framework to protect cancer survivors from financial discrimination when seeking a loan, a mortgage, or a life insurance. Thanks to progress in medicine, prognosis of several types of cancer has greatly improved over the last decades, supporting the implementation of legal frameworks protecting cancer survivors' rights. This thesis attempts to shed light on different methods that can be used to measure the viability and sustainability, both from a statistical and actuarial perspective, of a reduced waiting period opening the right to be forgotten. Furthermore, we hope that the illustrations presented in this work will convince countries currently lacking such rights to adopt the necessary protections in order to end financial discrimination against cancer survivors. This unified direction would certainly be a substantial advancement for long-term cancer survivors and the society as a whole, not to mention the insurance industry since proper coverage of such risks may well produce attractive returns.

As one could expect, some questions that came up during this work remain unanswered. We develop upon some of them in the next section, together with avenues for future research.

6.2 Future research

The field of cancer research is currently extremely lively, and the number of applications in the statistical and actuarial literature keeps on growing steadily. Additionally, the right to be forgotten is a topic of great importance to the governments, institutions and insurance providers of several European countries. Still, many questions and challenges remain to be faced, some of which are potential areas for further research. We list them below.

First, as already mentioned, we advocate for using the date of diagnosis instead of the end of the therapeutic treatment as a starting point for defining the waiting period opening the right to be forgotten. Nonetheless, it could be argued that the length of cancer treatments may have significantly decreased over the considered period (i.e., 2004-2022). A comparison between our approach and one based on the end of treatment would be valuable, in order to investigate the potential differences in terms of length of the waiting period resulting from the two approaches. However, the lack of individual treatment data in national registries currently impedes such an analysis. Additionally, the definition of treatment completion is somewhat subjective and treatment durations vary widely, further supporting the use of diagnosis date. In any case, a reduction in treatment length due to medical advancements would bring both approaches into closer alignment.

Second, calendar time has not been included in the analyses conducted in the present thesis. However, approaches presented in this thesis are dependent on factors such as changing diagnostic criteria and improved diagnostic methods. To illustrate this, suppose that a medical advance allows a cancer to be diagnosed earlier and perhaps also at a less severe stage, with the consequence that more cases are detected and that detected cases are not as fatal. These refined diagnostic methods will most likely imply an increased survival rate, regardless of whether the treatment improved or not. One could thus argue that not distinguishing between patients diagnosed recently and those diagnosed many years ago could hinder a potential cohort effect. Although this holds, there are two arguments which limits the benefits

of differentiating patients based on the diagnosis year or period. First, recall that we concentrate on young adults because of the products under consideration and cancer mainly affects people from a more advanced age. Combined with the fact that the obligation to report all cases of cancer to a registry is relatively recent in Belgium, this implies that there are, at least currently, a limited amount of cases available. Therefore, performing analyses for different cohorts separately may, at the moment, actually decrease the precision of the estimators and the stability of the results due to smaller sample sizes to such an extent that it would lead to unsound conclusions. Second, earlier diagnoses tend to be associated with better efficacy of the treatment and better prognosis. Given that medical treatments improved over the last decades, the resulting bias of ignoring a potential cohort effect favors insurance providers. We believe that, as long as patients survival increases (thanks to improved treatments for instance), ignoring a potential cohort effect will favor insurers, as the better the prognosis, the lower the risk of covering cancer survivors at the same rate than cancer-free clients. Along the same lines, one could argue whether or not results based on historical survival data can be generalized for the future, that is, used by insurance providers for the next generations of patients. In other words, can results obtained based on survival of patients diagnosed up to a few years ago be used for patients who will be diagnosed in a few or many years from now. On this matter, we believe that results remain valid and accurate as long as patients' survival stays similar across the years. In the case where survival improves over the years, results obtained in the present thesis could be considered as conservative and prudent from the perspective of the insurer. On the contrary, it is clear that if survival decreases throughout the years, results obtained based on data from the past decades may not be extrapolated to the future, and should therefore be applied with caution.

The 3-state Semi-Markov model, introduced in Chapter 3, is used for insurance covers typically limited to the first cancer occurrence. A promising avenue of research for the future would be to combine several types of cancer into a single model to design products offering against more than just one cancer. Considering the three cancer sites presented in this thesis, the 3-state model would result in a hierarchical Semi-Markov model with 5 states, with the ill state replaced with 3 states, corresponding to thyroid, melanoma and breast cancer. A natural extension to this general setting would be to construct a model with as many states as types of cancer, in addition to the active and dead state. Of course, distinguishing among cancer types is only relevant if coverage conditions vary with cancer site. In the same vein, more innovative solutions could be envisaged from an actuarial point of view (e.g., insurance products targeted to two individuals), and other methods from the biostatistical/epidemiological domain could be considered (e.g., a-splines, joint models, etc.).

Arık et al. (2020) showed, in a comprehensive study using UK data, that for female breast cancer there are regional differences in terms of cancer morbidity. Regional differences may occur for other cancer sites as well. In the present thesis, analyses have been conducted at the national level and not at the regional level, mainly because in Belgium it is the location of the hospital which matters the most rather than the region where the patient live, and because it goes beyond the scope of this thesis which primarily aimed at illustrating new methods in the context of the right to be forgotten. A valuable complement to this research would be to refine the analyses at a regional level instead of national level.

Fifth, in this thesis, the assumption is made that a cancer patient cannot become healthy again, that is, transition from the ill to the healthy state is not possible. Given that this assumption is believed to be reasonable for most cancers and since in our context, the real transition of interest is more from ill to dead than from ill to healthy, only irreversible multi-state models have been considered. Furthermore, this non-reversibility greatly simplifies the computations, as in this case, our 3-state process is

hierarchical and trajectories can be described in terms of just a few random variables (Denuit et al., 2019). However, one may argue that it does not always hold, and in that case more advanced multi-state models (such as, among others, reversible models) should thus be preferred. Another advantage of using a reversible model that allows the transition from ill to healthy is that cancer recurrence could be taken into account. This is undoubtedly an interesting question, but left for future research.

Sixth, for cancer patients, quality of life may be considered as important as the length of life itself (Shrestha et al., 2019). Therefore, other indicators than the number of years of life lost due to cancer such as, among others, the disability-adjusted life years (DALY) could also have been considered. Given that it is particularly useful for diseases or conditions that cause significant disability or do not result in death (such as diabetes or HIV), this seems to be a promising avenue for future research.

Seventh, it has been argued that the method proposed in Chapter 2 remains actuarially sound if the length of the follow-up is long enough such that no extrapolation is needed. Ignoring the fact that even when registry data have a sufficiently long follow-up period, the alternative method could still be preferable since a long follow-up means that some patients have been diagnosed a long time ago (and are thus not treated as well as nowadays), it could be interesting to compare both approaches to see, if there are indeed differences, what impact they have on the waiting periods.

Eight, an enduring challenge within the insurance domain concerns whether mortality rates observed in the general population can be used by insurers to set premium rates. This question comes from a well-recognized principle in the insurance literature that mortality rates for insured populations may differ (and likely be lower) from those of the general population due to factors like socioeconomic status, access to healthcare, lifestyle choices, etc. For insurance providers, accessing the mortality data specific to the insured population alone would more closely mirror real risks than using mortality data for the general population. In this thesis, when excess cancer mortality was required, mortality in the cancer population has been compared to the expected mortality in the general population. We did not account for the fact that mortality of the general population may not be identical to the one of the insured population. Mortality rates observed in the insured populations are, however, not so easy to estimate or obtain, mainly due to the fact that data provided by the National Bank of Belgium pertain to the number of insurance companies with whom the deceased policyholders held contracts, rather than the actual count of deaths. This implies that some individuals appear several times in the data, which makes it difficult to estimate precisely the number of deaths and thus the mortality rates. This promising avenue for research is currently being studied by my colleague Aurèle Bartolomeo.

Ninth, as it has been shown, a crucial aspect of conducting studies presented in the present thesis involves accessing follow-up data for both the general population and the cohort affected by the pathology of interest. Nonetheless, disease-specific registries are frequently either nonexistent, incomplete, or unreliable. When it comes to assessing the survival of individuals with a particular pathology lacking appropriate registry data, one potential approach is to utilize data from cohorts or registries in other countries and adapt them to the specific context of the country under investigation. A valuable extension would thus be to develop a way to facilitate the use of an existing registry from another country to project survival rates in the studied country. This is actually the research topic of my colleague Fanny Hoogstoel, for whom the aim is to illustrate the process of using a French registry for an application in Belgium.

Tenth, if we would like to go one step further when estimating the right to be forgotten by stage of the tumor, we should probably consider the model from Touraine et al. (2020) rather than the one from Esteve et al. (1990). The reason for

preferring the former over the latter is that the corresponding population may not be exactly the same for patients diagnosed at stages I–II and patients diagnosed at stages III–IV. Indeed, patients diagnosed at stages III–IV may be more likely to have comorbidities than patients diagnosed at stages I–II. For example, in lung cancer, it is well know that smokers are diagnosed at a later stage. This happens because the first symptoms (usually a cough) often mask the cancer; smokers usually find out they have cancer quite late (or at least, later than for other cancers) because they cannot easily distinguish whether the symptoms are a consequence of their smoking habits or a consequence of having cancer. Therefore, for lung cancer, it could be that the proportion of smokers is greater in the group of people diagnosed at stages III–IV than in the group of people diagnosed at stages I–II. If this is indeed the case, neither of these two groups is really comparable to the general population because non-smokers are healthier overall, and smokers have many other comorbidities. To sum up, the question is whether we can consider the same general population for both groups of patients. Verification of these assumptions, and whether adjustments are needed to account for this potential bias, are left for future research.

Last but not least, another research idea would be to study other cancer sites. In this thesis, we focused on melanoma, thyroid and female breast cancers to illustrate the proposed methods on a sample of cancer types with clear differences in terms of incidence rates and survival prognosis. However, cancer is not one disease, but a family of many diverse diseases with different outcomes. Therefore, our findings cannot be applied to other cancer types blindly. Note also that *in situ* cancer cases have not been included in the present studies as they are not classified the same way as any other "regular" cancer cases. A natural extension of this work would be to repeat the analyses for all major cancer types, and perhaps even to *in situ* cancers. Moreover, the proposed approaches can obviously be applied to other countries. This would certainly be useful for implementing appropriate market rules, and for defining a tailor-made waiting period for each type of cancer, so that it closely reflects the real risks and patients' vital prognosis. In the same spirit, another intriguing research idea would be to apply these methodologies to other chronic diseases for which the right to be forgotten has been implemented, such as HIV, leukemia or some types of hepatitis. The methodologies could even be applied to chronic diseases for which the right to be forgotten do not yet exist, but for which the excess mortality becomes small or even negligible after some years after diagnosis. Moreover, it is logical to focus primarily on pathologies that affect young people, as they are the ones most likely to take out life insurance. The main difficulty, however, seems to be the lack of nationwide registry for these diseases. An appropriate source of reliable and representative data must thus be identified to perform actuarial calculations assessing the actual costs of these extensions beyond cancer.

Appendix

A

A.1 Development of $e_{11}^{\tau}(t;z)$

This appendix shows how Eq. (4.8) introduced in Chapter 4 is obtained.

Assuming $\tau > t$, and τ and t are integers, we have

$$e_{11}^{\tau}(t;z) = \int_{t}^{\tau} p_{11}(t,u;z)\mathrm{d}u \tag{A.1}$$

$$= \sum_{k=0}^{\tau-t-1} \int_{t+k}^{t+k+1} p_{11}(t,u;z)\mathrm{d}u \tag{A.2}$$

$$= \sum_{k=0}^{\tau-t-1} \int_{t+k}^{t+k+1} p_{11}(t,t+k;z)p_{11}(t+k,u;z+k)\mathrm{d}u \tag{A.3}$$

$$= \sum_{k=0}^{\tau-t-1} \underbrace{p_{11}(t,t+k;z)}_{(1)} \underbrace{\int_{t+k}^{t+k+1} p_{11}(t+k,u;z+k)\mathrm{d}u}_{(2)} \tag{A.4}$$

The terms (1) and (2) in Eq. (A.4) are developed below.

$$(1)\ p_{11}(t,t+k;z) = \exp\left(-\int_{t}^{t+k} \alpha_{12}(u;z+u-t)\mathrm{d}u\right) \tag{A.5}$$

$$= \exp\left(-\sum_{l=0}^{k-1} \int_{t+l}^{t+l+1} \alpha_{12}(u;z+u-t)\mathrm{d}u\right) \tag{A.6}$$

$$= \exp\left(-\sum_{l=0}^{k-1} \int_{t+l}^{t+l+1} \alpha_{12}(t+l;z+l)\mathrm{d}u\right) \tag{A.7}$$

$$= \exp\left(-\sum_{l=0}^{k-1} \alpha_{12}(t+l;z+l)\right) \tag{A.8}$$

$$(2)\ \int_{t+k}^{t+k+1} p_{11}(t+k,u;z+k)\mathrm{d}u = \int_{t+k}^{t+k+1} \exp\left(-\int_{t+k}^{u} \alpha_{12}(u;z+u-t)\mathrm{d}s\right)\mathrm{d}u \tag{A.9}$$

$$= \int_{t+k}^{t+k+1} \exp\left(-\int_{t+k}^{u} \alpha_{12}(t+k;z+k)\mathrm{d}s\right)\mathrm{d}u \tag{A.10}$$

$$= \int_{t+k}^{t+k+1} \exp\left(-\alpha_{12}(t+k;z+k)(u-t-k)\right)\mathrm{d}u \tag{A.11}$$

$$= \left[\frac{\exp\Big(- \alpha_{12}(t+k; z+k)(u-t-k)\Big)}{-\alpha_{12}(t+k; z+k)} \right]_{t+k}^{t+k+1} \tag{A.12}$$

$$= \frac{1 - \exp\Big(- \alpha_{12}(t+k; z+k)\Big)}{\alpha_{12}(t+k; z+k)} \tag{A.13}$$

Hence,

$$e_{11}^{\tau}(t; z) = \sum_{k=0}^{\tau-t-1} \underbrace{\exp\left(- \sum_{l=0}^{k-1} \alpha_{12}(t+l; z+l) \right)}_{(1)} \underbrace{\frac{1 - \exp\Big(- \alpha_{12}(t+k; z+k)\Big)}{\alpha_{12}(t+k; z+k)}}_{(2)}. \tag{A.14}$$

A.2 Conditional one-year observed survival probability

This appendix details the calculation of the conditional one-year observed survival probabilities introduced in Chapter 5.

A.2.1 Observed survival and hazard rate

From the relation between cumulative hazard for all cause death, Λ, follows the observed survival, OS:

$$OS(t) = \exp\left(-\Lambda(t)\right) = \exp\left(- \int_0^t \lambda(u) du \right), \tag{A.15}$$

with $\lambda(u)$ the hazard rate at time u.

Conditional one-year OS at time t can be obtained from Eq. (A.15) by integrating only over the interval $[t, t+1]$:

$$OS(t, t+1) = \exp\left(- \int_t^{t+1} \lambda(u) du \right). \tag{A.16}$$

To calculate this integral in practice, numerical integration can be applied on a set of time values, say with a step of 0.01 year.

A.2.2 Flexible parametric model for the hazard rate

The hazard rate as a continuous function of survival time was obtained from a flexible parametric model (fpm) using the `mexhaz` function from the R package `mexhaz` (Charvat and Belot, 2021).

A.2.3 Predicted observed survival

The observed survival at a given time t, can be obtained from numerical integration of Eq. (A.15):

$$OS(t) = \exp\left(- \sum_{i=0}^{N_t - 1} \lambda(t_i)(t_{i+1} - t_i) \right) = \exp\left(- \sum_{i=0}^{N_t - 1} \lambda(t_i) \Delta t \right) = \exp\left(-\tilde{\Lambda}(t) \right), \tag{A.17}$$

when the $[0, t]$ interval is split in N_t intervals of width Δt ($t_0 = 0, t_1 = \Delta t, t_2 = 2\Delta t, \ldots, t_{N_t} = t$).

To obtain a curve of the observed survival at a set of time values (say from 0 to 10 years in steps of $\Delta t = 0.01$ year, so 1000 data points), the vector of the corresponding cumulative hazards, $\tilde{\mathbf{\Lambda}}$, calculated via numerical integration is needed:

$$\mathbf{OS} = \exp(-\tilde{\mathbf{\Lambda}}). \tag{A.18}$$

The cumulative hazard vector can be calculated from the estimated regression coefficients via matrix multiplication. Let $\mathbf{X}$ be the design matrix (N_t lines, each line corresponds with a time value) for the needed linear combinations of estimated regression coefficients, $\boldsymbol{\beta}$, at the $\log(\lambda(t))$ scale. The estimated $\mathbf{log(\lambda)}$ vector and its covariance matrix at each time points equals:

$$\mathbf{log(\lambda)} = \mathbf{X}\boldsymbol{\beta} \tag{A.19}$$

$$\sum_{\mathbf{log(\lambda)}} = \mathbf{X}\sum_{\boldsymbol{\beta}}\mathbf{X}^T. \tag{A.20}$$

So:

$$\boldsymbol{\lambda} = \exp(\mathbf{X}\boldsymbol{\beta}) \tag{A.21}$$

$$\sum_{\boldsymbol{\lambda}} = \mathbf{J}_{\boldsymbol{\lambda}}\sum_{\boldsymbol{\beta}}\mathbf{J}_{\boldsymbol{\lambda}}{}^T, \tag{A.22}$$

with $\mathbf{J}_{\boldsymbol{\lambda}}$ the Jacobian matrix $\mathbf{J}_{\boldsymbol{\lambda}} = diag(\exp(\mathbf{X}\boldsymbol{\beta}))$.

The cumulative hazard at all time points is easily obtained by multiplying with a upper triangular matrix $\boldsymbol{T}$ (with 1's on the diagonal):

$$\tilde{\mathbf{\Lambda}} = \Delta t \cdot \boldsymbol{\lambda}^T\mathbf{T} \tag{A.23}$$

$$\sum_{\tilde{\mathbf{\Lambda}}} = (\Delta t)^2\,\mathbf{T}^T\sum_{\boldsymbol{\lambda}}\mathbf{T}. \tag{A.24}$$

The variances of the cumulative hazards are the diagonal elements of covariance matrix, which allows to construct an asymptotic normal confidence interval (CI).

From Eq. (A.18), it follows:

$$\mathbf{OS} = \exp\left(-\tilde{\mathbf{\Lambda}}\right) \tag{A.25}$$

$$\sum_{\mathbf{OS}} = \mathbf{J}_{\mathbf{OS}}\sum_{\tilde{\mathbf{\Lambda}}}\mathbf{J}_{\mathbf{OS}}{}^T. \tag{A.26}$$

An asymptotic CI on the obtained survival can be obtained by transforming the CI on the cumulative hazard.

A.2.4 Predicted conditional one-year observed survival

To obtain the conditional one-year observed survival at each time point t_i, the cumulative hazard over only the next 1 year interval $[t_i, t_i + 1]$ is needed. This can be achieved by creating an upper triangular matrix, $\mathbf{T_c}$, for which the number of 1's in each row is limited up to the next $\frac{1}{\Delta t}$ columns. Take as an example $\Delta t = 0.2$ and consider the first six lines and the first 10 columns:

$$\mathbf{T_c} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}. \tag{A.27}$$

The cumulative hazard for the conditional one-year OS becomes:

$$\tilde{\mathbf{\Lambda}}_c = \Delta t \cdot \mathbf{T}_c \boldsymbol{\lambda} \tag{A.28}$$

$$\sum_{\tilde{\mathbf{\Lambda}}} = (\Delta t)^2 \, \mathbf{T}_c \sum_{\boldsymbol{\lambda}} \mathbf{T}_c^T. \tag{A.29}$$

The rest is similar to the observed survival in the previous subsection.

Bibliography

P. D. Allison. *Survival analysis using SAS: a practical guide*. Sas Institute, 2010.

M. Amico and I. Van Keilegom. Cure models in survival analysis. *Annual Review of Statistics and Its Application*, 5:311–342, 2018.

P. K. Andersen. Decomposition of number of life years lost according to causes of death. *Statistics in medicine*, 32(30):5278–5285, 2013.

P. K. Andersen. Life years lost among patients with a given disease. *Statistics in medicine*, 36(22):3573–3582, 2017.

P. K. Andersen and M. Pohar Perme. Inference for outcome probabilities in multi-state models. *Lifetime data analysis*, 14(4):405–431, 2008.

P. K. Andersen, O. Borgan, R. D. Gill, and N. Keiding. *Statistical models based on counting processes*. Springer Science & Business Media, 2012.

P. K. Andersen, V. Canudas-Romo, and N. Keiding. Cause-specific measures of life years lost. *Demographic Research*, 29:1127–1152, 2013.

R. M. Anderson. Discussion: the kermack-mckendrick epidemic threshold theorem. *Bulletin of mathematical biology*, 53(1):1–32, 1991.

T. M. Andersson, P. W. Dickman, S. Eloranta, and P. C. Lambert. Estimating and modelling cure in population-based cancer studies within the framework of flexible parametric survival models. *BMC medical research methodology*, 11(1):96, 2011.

T. M. Andersson, P. W. Dickman, S. Eloranta, A. Sjövall, M. Lambe, and P. C. Lambert. The loss in expectation of life after colon cancer: a population-based study. *BMC cancer*, 15(1):1–10, 2015.

T. M.-L. Andersson, P. W. Dickman, S. Eloranta, M. Lambe, and P. C. Lambert. Estimating the loss in expectation of life due to cancer using flexible parametric survival models. *Statistics in medicine*, 32(30):5286–5300, 2013.

Y. Y. Andrei, B. Asselain, et al. *Stochastic models of tumor latency and their biostatistical applications*, volume 1. World Scientific, 1996.

J. Antero-Jacquemin, M. Pohar-Perme, G. Rey, J.-F. Toussaint, and A. Latouche. The heart of the matter: years-saved from cardiovascular and cancer deaths in an elite athlete cohort with over a century of follow-up. *European journal of epidemiology*, 33:531–543, 2018.

T. J. Aragon, D. Y. Lichtensztajn, B. S. Katcher, R. Reiter, and M. H. Katz. Calculating expected years of life lost for assessing local ethnic disparities in causes of premature death. *BMC public health*, 8(1):116, 2008.

A. Arık, E. Dodd, and G. Streftaris. Cancer morbidity trends and regional differences in england—a bayesian analysis. *PloS one*, 15(5):e0232844, 2020.

A. Arik, A. J. G. Cairns, E. Dodd, A. S. Macdonald, and G. Streftaris. Estimating the impact of the covid-19 pandemic on breast cancer deaths among older women. In *2023 Living to 100 Research Symposium-Asia*, 2023.

P. D. Baade, D. R. Youlden, T. M. Andersson, P. H. Youl, M. G. Kimlin, J. F. Aitken, and R. J. Biggar. Estimating the change in life expectancy after a diagnosis of cancer among the australian population. *BMJ open*, 5(4):e006740, 2015.

P. D. Baade, D. R. Youlden, T. M. Andersson, P. H. Youl, E. T. Walpole, M. G. Kimlin, J. F. Aitken, and R. J. Biggar. Temporal changes in loss of life expectancy due to

cancer in australia: a flexible parametric approach. *Cancer Causes & Control*, 27(8): 955–964, 2016.

J. C. Bailar III and E. M. Smith. Progress against cancer? *New England Journal of Medicine*, 314(19):1226–1232, 1986.

C. M. Balch, S.-J. Soong, J. E. Gershenwald, J. F. Thompson, D. S. Reintgen, N. Cascinelli, M. Urist, K. M. McMasters, M. I. Ross, J. M. Kirkwood, et al. Prognostic factors analysis of 17,600 melanoma patients: validation of the american joint committee on cancer melanoma staging system. *Journal of clinical oncology*, 19(16):3622–3634, 2001.

A. Belot, A. Ndiaye, M.-A. Luque-Fernandez, D.-K. Kipourou, C. Maringe, F. J. Rubio, and B. Rachet. Summarizing and communicating on survival data according to the audience: a tutorial on different measures illustrated with population-based cancer registry data. *Clinical epidemiology*, 11:53–65, 2019.

C. L. Bennett, P. D. Weinberg, and J. J. Lieberman. Cancer insurance policies in japan and the united states. *Western journal of medicine*, 168(1):17, 1998.

J. Berkson and R. P. Gage. Calculation of survival rates for cancer. *Proceedings of the staff meetings. Mayo Clinic*, 25(11):270—286, May 1950. ISSN 0092-699X. URL `http://europepmc.org/abstract/MED/15417650`.

J. Berkson and R. P. Gage. Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, 47(259):501–515, 1952.

F. Berrino, R. Capocaccia, J. Estève, G. Gatta, T. Hakulinen, A. Micheli, M. Sant, and A. Verdecchia. Survival of cancer patients in europe: the eurocare-2 study. In *Survival of cancer patients in Europe: The EUROCARE-2 study*, volume 151, pages 1–CD. IARC Scientific Publication, 1999.

J. W. Boag. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society. Series B (Methodological)*, 11 (1):15–53, 1949.

P. Bolard, C. Quantin, J. Esteve, J. Faivre, and M. Abrahamowicz. Modelling time-dependent hazard ratios in relative survival: application to colon cancer. *Journal of clinical epidemiology*, 54(10):986–996, 2001.

L. Botta, L. Dal Maso, S. Guzzinati, C. Panato, G. Gatta, A. Trama, M. Rugge, G. Tagliabue, C. Casella, B. Caruso, et al. Changes in life expectancy for cancer patients over time since diagnosis. *Journal of advanced research*, 20:153–159, 2019.

O. Boussari, G. Romain, L. Remontet, N. Bossard, M. Mounier, A.-M. Bouvier, C. Binquet, M. Colonna, and V. Jooste. A new approach to estimate time-to-cure from cancer registries data. *Cancer epidemiology*, 53:72–80, 2018.

N. L. Bowers. Actuarial mathematics. *(No Title)*, 1997.

J. Buckley. Additive and multiplicative models for relative survival rates. *Biometrics*, 40:51–62, 1984.

Bureau du suivi de la tarification. Rapport sur l'activité 2017. `https://www.bureaudusuivi.be/images/docs/RapportAnnuel_2017.pdf`, 2018.

R. Capocaccia, G. Gatta, and L. Dal Maso. Life expectancy of colon, breast, and testicular cancer patients: an analysis of us-seer population-based data. *Annals of Oncology*, 26(6):1263–1268, 2015.

Centers for Disease Control and Prevention. Years of potential life lost before age 65–united states, 1990 and 1991. *MMWR. Morbidity and mortality weekly report*, 42 (13):251–253, 1993.

A. E. Chang, P. A. Ganz, D. F. Hayes, T. Kinsella, H. I. Pass, J. H. Schiller, R. M. Stone, and V. Strecher. *Oncology: an evidence-based approach*. Springer Science & Business Media, 2007.

A. Charpentier. *Computational actuarial science with R*. CRC press, 2014.

H. Charvat and A. Belot. mexhaz: An R package for fitting flexible hazard-based regression models for overall and excess mortality with a random effect. *Journal of Statistical Software*, 98(14):1–36, 2021. doi: 10.18637/jss.v098.i14.

M. Chauvenet, C. Lepage, V. Jooste, V. Cottet, J. Faivre, and A.-M. Bouvier. Prevalence of patients with colorectal cancer requiring follow-up or active treatment. *European Journal of Cancer*, 45(8):1460–1465, 2009.

M.-H. Chen, J. G. Ibrahim, and D. Sinha. A new bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association*, 94(447): 909–919, 1999.

C. L. Chiang. Life table and its applications. In *Life table and its applications*, pages 316–316. 1984.

P.-C. Chu, J.-D. Wang, J.-S. Hwang, and Y.-Y. Chang. Estimation of life expectancy and the expected years of life lost in patients with major cancers: extrapolation of survival curves under high-censored rates. *Value in Health*, 11(7):1102–1109, 2008.

M. Clements and X.-R. Liu. *rstpm2: Smooth Survival Models, Including Generalized Survival Models*, 2019. URL `https://CRAN.R-project.org/package=rstpm2`. R package version 1.5.1.

I. Clerc-Urmès and M. Grzebyk. *flexrsurv: Flexible Relative Survival Analysis*, 2023. URL `https://CRAN.R-project.org/package=flexrsurv`. R package version 2.0.17.

I. Clerc-Urmès, M. Grzebyk, G. Hédelin, and CENSUR working survival group. *flexrsurv: An R package for relative survival analysis*, 2020. URL `https://CRAN.R-project.org/package=flexrsurv`. R package version 1.4.5.

W. Cleveland, E. Grosse, M. Shyu, E. Grosse, M. Shyu, and M. Shyu. A package of c and fortran routines for fitting local regression models. *Statistical methods, edited by: Chambers, JM, S. Chapman and Hall Ltd., London, UK*, 1992.

W. S. Cleveland and E. Grosse. Computational methods for local regression. *Statistics and computing*, 1(1):47–62, 1991.

W. S. Cleveland, S. J. Devlin, and E. Grosse. Regression by local fitting: methods, properties, and computational algorithms. *Journal of econometrics*, 37(1):87–114, 1988.

M. P. Coleman, P. Babb, P. Damiecki, P. Grosclaude, S. Honjo, J. Jones, G. Knerer, A. Pitard, M. Quinn, A. Sloggett, et al. *Cancer survival trends in England and Wales, 1971-1995: deprivation and NHS region.* Stationery Office Books, 1999.

D. Collett. *Modelling survival data in medical research.* CRC press, 2023.

D. R. Cox. *Analysis of survival data.* Chapman and Hall/CRC, 2018.

CRUK. Breast cancer survival statistics. `https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer`, Jul 2023a.

CRUK. Melanoma skin cancer survival statistics. `https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/melanoma-skin-cancer`, Jul 2023b.

CRUK. Thyroid cancer survival statistics. `https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/thyroid-cancer`, Jul 2023c.

C. Czado and I. Van Keilegom. Dependent censoring based on parametric copulas. *Biometrika*, 110(3):721–738, 2023.

L. Dal Maso, S. Guzzinati, C. Buzzoni, R. Capocaccia, D. Serraino, A. Caldarella, A. Dei Tos, F. Falcini, M. Autelitano, G. Masanotti, et al. Long-term survival, prevalence, and cure of cancer: a population-based estimation for 818 902 italian patients and 26 cancer types. *Annals of oncology*, 25(11):2251–2260, 2014.

C. Danieli, L. Remontet, N. Bossard, L. Roche, and A. Belot. Estimating net survival: the importance of allowing for informative censoring. *Statistics in medicine*, 31(8): 775–786, 2012.

L. C. De Wreede, M. Fiocco, and H. Putter. The mstate package for estimation and prediction in non-and semi-parametric multi-state and competing risks models. *Computer methods and programs in biomedicine*, 99(3):261–274, 2010.

J. Dębicka, A. Marciniuk, and B. Zmyślona. Combination reverse annuity contract and critical illness insurance. In *18th AMSE, Applications of Mathematics and Statistics in Economics, Conference Proceedings*, pages 2–6, 2015.

M. Delhelle and I. Van Keilegom. Copula based dependent censoring in cure models. Technical report, Université catholique de Louvain, Institute of Statistics, Biostatistics and …, 2023.

M. Denuit and C. Legrand. Risk classification in life and health insurance: extension to continuous covariates. *European Actuarial Journal*, 8:245–255, 2018.

M. Denuit, N. Lucas, and E. Pitacco. Pricing and reserving in ltc insurance. *Actuarial aspects of long term care*, pages 129–158, 2019.

N. W. Deresa and I. V. Keilegom. Copula based cox proportional hazards models for dependent censoring. *Journal of the American Statistical Association*, pages 1–11, 2023.

N. W. Deresa and I. Van Keilegom. Flexible parametric model for survival data subject to dependent censoring. *Biometrical Journal*, 62(1):136–156, 2020.

P. W. Dickman, A. Sloggett, M. Hills, and T. Hakulinen. Regression models for relative survival. *Statistics in medicine*, 23(1):51–64, 2004.

D. C. Dickson, M. Hardy, M. R. Hardy, and H. R. Waters. *Actuarial mathematics for life contingent risks*. Cambridge University Press, 2013.

D. C. Dickson, M. R. Hardy, and H. R. Waters. *Actuarial mathematics for life contingent risks*. Cambridge University Press, 2019.

E. O. Dodd, G. Streftaris, H. R. Waters, and A. D. Stott. The effect of model uncertainty on the pricing of critical illness insurance. *Annals of Actuarial Science*, 9(1):108–133, 2015.

P. H. Eilers and B. D. Marx. Flexible smoothing with b-splines and penalties. *Statistical science*, 11(2):89–102, 1996.

S. E. Emoto and P. C. Matthews. A weibull model for dependent censoring. *The Annals of Statistics*, 18(4):1556–1577, 1990.

J. E. Enstrom and D. F. Austin. Interpreting cancer survival rates. *Science*, 195(4281): 847–851, 1977.

J. Esteve, E. Benhamou, M. Croasdale, and L. Raymond. Relative survival and the estimation of net survival: elements for further discussion. *Statistics in medicine*, 9 (5):529–538, 1990.

J. Esteve, E. Benhamou, L. Raymond, et al. Statistical methods in cancer research. volume iv. descriptive epidemiology. *IARC Sci publ*, 128(1):302, 1994.

European Initiative on Ending Discrimination against Cancer Survivors. Measures to protect cancer survivors from financial discrimination in the EU/EEA, 2024. URL `https://endingdiscrimination-cancersurvivors.eu/`.

M. Fauvernier, L. Remontet, Z. Uhry, N. Bossard, and L. Roche. survpen: an r package for hazard and excess hazard modelling with multidimensional penalized splines. *Journal of Open Source Software*, 4(40):1434, 2019a. doi: 10.21105/joss.01434.

M. Fauvernier, L. Roche, Z. Uhry, L. Tron, N. Bossard, L. Remontet, and C. in the Estimation of Net Survival Working Survival Group. Multi-dimensional penalized hazard model with continuous covariates: applications for studying trends and social inequalities in cancer survival. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 2019b.

J. W. Gardner and J. S. Sanborn. Years of potential life lost (ypll)—what does it measure? *Epidemiology*, pages 322–329, 1990.

H. U. Gerber. *Life insurance mathematics*. Springer Science & Business Media, 2013.

R. B. Geskus. *Data analysis with competing risks and intermediate states*, volume 82. CRC Press, 2015.

R. B. Geskus. *Data analysis with competing risks and intermediate states*. Chapman and Hall/CRC, 2019.

R. Giorgi, J. Payan, and J. Gouvernet. Rsurv: a function to perform relative survival analysis with s-plus or r. *Computer methods and programs in biomedicine*, 78(2): 175–178, 2005.

S. Haberman and A. Renshaw. Generalised linear models and excess mortality from peptic ulcers. *Insurance: Mathematics and Economics*, 9(1):21–32, 1990.

T. Hakulinen and L. Tenkanen. Regression analysis of relative survival rates. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 36(3):309–317, 1987.

M. J. Hendriks, E. Harju, K. Roser, M. Ienca, and G. Michel. The long shadow of childhood cancer: a qualitative study on insurance hardship among survivors of childhood cancer. *BMC Health Services Research*, 21(1):1–12, 2021.

P. Hougaard. Multi-state models: a review. *Lifetime data analysis*, 5:239–264, 1999.

C. Jackson. Multi-state modelling with r: the msm package. *Cambridge, UK*, pages 1–53, 2007.

L. H. Jakobsen, T. M.-L. Andersson, J. L. Biccler, L. Ø. Poulsen, M. T. Severinsen, T. C. El-Galaly, and M. Bøgsted. On estimating the time to statistical cure. *BMC Medical Research Methodology*, 20:1–13, 2020.

V. Jooste, P. Grosclaude, L. Remontet, G. Launoy, I. Baldi, F. Molinié, P. Arveux, N. Bossard, A.-M. Bouvier, M. Colonna, et al. Unbiased estimates of long-term net survival of solid cancers in france. *International journal of cancer*, 132(10):2370–2377, 2013.

J. D. Kalbfleisch and R. L. Prentice. *The statistical analysis of failure time data*. John Wiley & Sons, 2011.

E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

W. O. Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721, 1927.

N. Keyfitz and H. Caswell. *Applied mathematical demography*, volume 47. Springer, 2005.

J. P. Klein, M. L. Moeschberger, et al. *Survival analysis: techniques for censored and truncated data*, volume 1230. Springer, 2003.

D. G. Kleinbaum and M. Klein. *Survival analysis a self-learning text*. Springer, 1996.

S. W. Lagakos. General right censoring and its impact on the analysis of survival data. *Biometrics*, pages 139–156, 1979.

P. C. Lambert. Modeling of the cure fraction in survival studies. *The Stata Journal*, 7 (3):351–375, 2007.

P. C. Lambert, J. R. Thompson, C. L. Weston, and P. W. Dickman. Estimating and modeling the cure fraction in population-based cancer survival analysis. *Biostatistics*, 8 (3):576–594, 2006.

A. Latouche, A. Allignol, J. Beyersmann, M. Labopin, and J. P. Fine. A competing risks analysis should report results on all cause-specific hazards and cumulative incidence functions. *Journal of clinical epidemiology*, 66(6):648–653, 2013.

A. Latouche, P. K. Andersen, G. Rey, and M. Moreno-Betancur. A note on the measurement of socioeconomic inequalities in life years lost by cause of death. *Epidemiology*, 30(4):569–572, 2019.

M. Lawler, F. De Lorenzo, P. Lagergren, F. S. Mennini, S. Narbutas, G. Scocca, F. Meunier, and E. A. of Cancer Sciences. Challenges and solutions to embed cancer survivorship research and innovation within the eu cancer mission. *Molecular Oncology*, 15(7):1750–1758, 2021.

C. Legrand. *Advanced survival models*. CRC Press, 2021.

C. Legrand and A. Bertrand. Cure models in oncology clinical trials. *Textbook of Clinical Trials in Oncology: A Statistical Perspective*, 1:465–492, 2019.

J. Lemaire, K. Subramanian, K. Armstrong, and D. A. Asch. Pricing term insurance in the presence of a family history of breast or ovarian cancer. *North American Actuarial Journal*, 4(2):75–87, 2000.

P. Lenner. The excess mortality rate: a useful concept in cancer epidemiology. *Acta Oncologica*, 29(5):573–576, 1990.

S. Licher, A. Heshmatollah, K. D. van der Willik, B. H. C. Stricker, R. Ruiter, E. W. de Roos, L. Lahousse, P. J. Koudstaal, A. Hofman, L. Fani, et al. Lifetime risk and multimorbidity of non-communicable diseases and disease-free life expectancy in the general population: a population-based cohort study. *PLoS medicine*, 16(2): e1002741, 2019.

X.-R. Liu, Y. Pawitan, and M. S. Clements. Generalized survival models for correlated time-to-event data. *Statistics in medicine*, 36(29):4743–4762, 2017.

X.-R. Liu, Y. Pawitan, and M. Clements. Parametric and penalized generalized survival models. *Statistical methods in medical research*, 27(5):1531–1546, 2018.

S. Maetani and J. W. Gamel. Parametric cure model versus proportional hazards model in survival analysis of breast cancer and other malignancies. *Advances in Breast Cancer Research*, 2013, 2013.

R. A. Maller and X. Zhou. *Survival analysis with long-term survivors*. John Wiley & Sons, 1996.

D. Manevski, N. Ružić Gorenjec, P. K. Andersen, and M. Pohar Perme. Expected life years compared to the general population. *Biometrical Journal*, 65(4):2200070, 2023.

M. Massart. A long-term survivor's perspective on supportive policy for a better access to insurance, loan and mortgage. *Journal of cancer policy*, 15:70–71, 2018.

L. Meira-Machado, J. de Uña-Álvarez, C. Cadarso-Suárez, and P. K. Andersen. Multi-state models for the analysis of time-to-event data. *Statistical methods in medical research*, 18(2):195–222, 2009.

M. Mesnil. What do we mean by the right to be forgotten? an analysis of the french case study from a lawyer's perspective. *Journal of cancer policy*, 15:122–127, 2018.

D. F. Moore. *Applied survival analysis using R*, volume 473. Springer, 2016.

M. Mounier. *Apport des méthodes de survie nette dans le pronostic des lymphomes malins non hodgkiniens en population générale*. PhD thesis, Université Claude Bernard-Lyon I, 2015.

NHS Digital. Cancer survival in england, cancers diagnosed 2016 to 2020, followed up to 2021. `https://digital.nhs.uk/data-and-information/publications/statistical/cancer-survival-in-england/cancers-diagnosed-2016-to-2020-followed-up-to-2021`, Feb 2023.

R. B. Nielsen and R. N. Mayer. Why do people buy cancer insurance? an exploratory study. *Advancing the Consumer Interest*, pages 16–22, 2000.

M. Noordzij, F. W. Dekker, C. Zoccali, and K. J. Jager. Measures of disease frequency: prevalence and incidence. *Nephron Clinical Practice*, 115(1):c17–c20, 2010.

H. Oksanen. *Modelling the survival of prostate cancer patients*. University of Tampere, 1998.

M. Othus, B. Barlogie, M. L. LeBlanc, and J. J. Crowley. Cure models as a useful statistical tool for analyzing survival. *Clinical Cancer Research*, 18(14):3731–3736, 2012.

S. Pasquali, A. V. Hadjinicolaou, V. C. Sileni, C. R. Rossi, and S. Mocellin. Systemic treatments for metastatic cutaneous melanoma. *Cochrane Database of Systematic Reviews*, (2), 2018.

K. Pavlič and M. Pohar Perme. Using pseudo-observations for estimation in relative survival. *Biostatistics*, 20(3):384–399, 2019.

S. Pedersen, R. B. Holmstroem, A. von Heymann, L. K. Tolstrup, K. Madsen, M. A. Petersen, C. A. Haslund, C. H. Ruhlmann, H. Schmidt, C. Johansen, et al. Quality of life and mental health in real-world patients with resected stage iii/iv melanoma receiving adjuvant immunotherapy. *Acta Oncologica*, 62(1):62–69, 2023.

C. Percy, E. Stanek 3rd, and L. Gloeckler. Accuracy of cancer death certificates and its effect on cancer mortality statistics. *American Journal of Public Health*, 71(3): 242–250, 1981.

M. P. Perme and K. Pavlič. Nonparametric relative survival analysis with the R package relsurv. *Journal of Statistical Software*, 87(8):1–27, 2018. doi: 10.18637/jss.v087.i08.

M. P. Perme, J. Stare, and J. Estève. On estimation in relative survival. *Biometrics*, 68 (1):113–120, 2012.

M. P. Perme, J. Estève, and B. Rachet. Analysing population-based cancer survival–settling the controversies. *BMC cancer*, 16(1):1–8, 2016.

E. Pitacco. Health insurance. *Basic Actuarial Models, Cham, Switzerland: Springer Verlag*, 2014.

M. Pohar and J. Stare. Relative survival analysis in r. *Computer methods and programs in biomedicine*, 81(3):272–278, 2006.

S. Preston, P. Heuveline, M. Guillot, et al. Measuring and modeling population processes. *Wiley-Blackwell*, 2001.

S. D. Promislow. *Fundamentals of actuarial mathematics*. John Wiley & Sons, 2014.

H. Putter, M. Fiocco, and R. B. Geskus. Tutorial in biostatistics: competing risks and multi-state models. *Statistics in medicine*, 26(11):2389–2430, 2007.

C. Quantin, M. Abrahamowicz, T. Moreau, G. Bartlett, T. MacKenzie, M. Adnane Tazi, L. Lalonde, and J. Faivre. Variation over time of the effects of prognostic factors in a population-based study of colon cancer: comparison of statistical models. *American journal of epidemiology*, 150(11):1188–1200, 1999.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL `https://www.R-project.org`.

L. Remontet, Z. Uhry, N. Bossard, J. Iwaz, A. Belot, C. Danieli, H. Charvat, L. Roche, and C. W. S. Group. Flexible and structured survival model for a simultaneous estimation of non-linear and non-proportional effects and complex interactions between continuous variables: Performance of this multidimensional penalized spline approach in net survival trend analysis. *Statistical methods in medical research*, 28(8):2368–2384, 2019.

A. E. Renshaw. Modelling excess mortality using glim. *Journal of the Institute of Actuaries*, 115(2):299–315, 1988.

L. Ries, M. Eisner, C. Kosary, B. Hankey, B. Miller, L. Clegg, and B. Edwards. Seer cancer statistics review, 1973–1999. bethesda, md: National cancer institute. *Also available from: URL: http://seer. cancer. gov/csr/1973–1999*, 2002.

L.-P. Rivest and M. T. Wells. A martingale approach to the copula-graphic estimator for the survival function under dependent censoring. *Journal of Multivariate Analysis*, 79(1):138–155, 2001.

C. Ruckman and J. Francis. *Financial Mathematics: A Practical Guide for Actuaries and Other Business Professionals*. BPP Professional Education, 2005.

R. Schaffar, B. Rachet, A. Belot, and L. M. Woods. Estimation of net survival for cancer patients: relative survival setting more robust to some assumption violations than

cause-specific setting, a sensitivity analysis on empirical data. *European Journal of Cancer*, 72:78–83, 2017.

G. Schvartsman, P. Taranto, I. C. Glitza, S. S. Agarwala, M. B. Atkins, and A. C. Buzaid. Management of metastatic cutaneous melanoma: updates in clinical practice. *Therapeutic Advances in Medical Oncology*, 11:1758835919851663, 2019.

G. Scocca and F. Meunier. A right to be forgotten for cancer survivors: a legal development expected to reflect the medical progress in the fight against cancer. *Journal of Cancer Policy*, 25:100246, 2020.

G. Scocca and F. Meunier. Towards an EU legislation on the right to be forgotten to access to financial services for cancer survivors. *European Journal of Cancer*, 162: 133–137, 2022.

K. Shang. *Individual Cancer Mortality Prediction.* Fundaciòn Mapfre, sep 2019. ISBN 8498446651. URL `https://www.xarg.org/ref/a/8498446651/`.

A. Shrestha, C. Martin, M. Burton, S. Walters, K. Collins, and L. Wyld. Quality of life versus length of life considerations in cancer patients: a systematic literature review. *Psycho-oncology*, 28(7):1367–1380, 2019.

G. Silversmit, D. Jegou, E. Vaes, E. Van Hoof, E. Goetghebeur, and L. Van Eycken. Cure of cancer for seven cancer sites in the Flemish region. *International journal of cancer*, 140(5):1102–1110, 2017a.

G. Silversmit, E. Vaes, and L. van Eycken. Estimation of population-based cancer-specific potential years of life lost in belgium. *European Journal of Cancer Prevention*, 26:157–163, 2017b.

A. Soetewey, C. Legrand, M. Denuit, and G. Silversmit. Waiting period from diagnosis for mortgage insurance issued to cancer survivors. *European Actuarial Journal*, 11 (1):135–160, 2021.

A. Soetewey, C. Legrand, M. Denuit, and G. Silversmit. Semi-markov modeling for cancer insurance. *European Actuarial Journal*, 12(2):813–837, 2022.

A. Soetewey, C. Legrand, M. Denuit, and G. Silversmit. Right to be forgotten for mortgage insurance issued to cancer survivors: critical assessment and new proposal. Technical report, Université catholique de Louvain, Institute of Statistics, Biostatistics and Actuarial Sciences, 2023.

A. Soetewey, C. Legrand, M. Denuit, and G. Silversmit. Health indices for disease incidence and duration in the semi-markov setting. Technical report, Université catholique de Louvain, Institute of Statistics, Biostatistics and Actuarial Sciences, 2024.

G. A. Spedicato et al. The lifecontingencies package: Performing financial and actuarial mathematics calculations in r. *Journal of Statistical Software*, 55(10):1–36, 2013.

E. Syriopoulou, H. Bower, T. M. Andersson, P. C. Lambert, and M. J. Rutherford. Estimating the impact of a cancer diagnosis on life expectancy by socio-economic group for a range of cancer types in england. *British journal of cancer*, 117(9): 1419–1426, 2017.

Terry M. Therneau and Patricia M. Grambsch. *Modeling Survival Data: Extending the Cox Model.* Springer, New York, 2000. ISBN 0-387-98784-3.

The Belgian Cancer Registry. Cancer survival in belgium. `https://kankerregister.org/media/docs/publications/CancerSurvivalinBelgium.PDF`, 2012.

T. M. Therneau. Extending the cox model. In *Proceedings of the first Seattle symposium in biostatistics: survival analysis*, pages 51–84. Springer, 1997.

F. Tichanek, A. Försti, A. Hemminki, O. Hemminki, and K. Hemminki. Survival in melanoma in the nordic countries into the era of targeted and immunological therapies. *European Journal of Cancer*, 186:133–141, 2023.

C. Touraine, C. Helmer, and P. Joly. Predictions in an illness-death model. *Statistical methods in medical research*, 25(4):1452–1470, 2016.

C. Touraine, N. Grafféo, R. Giorgi, and C. W. S. Group. More accurate cancer-related excess mortality through correcting background mortality for extra variables. *Statistical Methods in Medical Research*, 29(1):122–136, 2020.

P. Tralongo, M. S. McCabe, and A. Surbone. Challenge for cancer survivorship: improving care through categorization by risk. *J Clin Oncol*, 35(30):3516–7, 2017.

I. Tromme, C. Legrand, B. Devleesschauwer, U. Leiter, S. Suciu, A. Eggermont, J. Francart, F. Calay, J. A. Haagsma, J.-F. Baurain, et al. Melanoma burden by melanoma stage: Assessment through a disease transition model. *European Journal of Cancer*, 53:33–41, 2016.

A. Tsodikov, J. Ibrahim, and A. Yakovlev. Estimating cure rates from survival data: an alternative to two-component mixture models. *Journal of the American Statistical Association*, 98(464):1063–1078, 2003.

A. Van Ginckel, G. Silversmit, B. Van Gool, N. Van Damme, and P. Jonckheer. The right to be forgotten in breast cancer: new propositions. *Belgian Health Care Knowledge Centre (KCE)*, KCE Reports 351, 2022. doi: 10.57598/R351C.

Y. E. Yilmaz, J. F. Lawless, I. L. Andrulis, and S. B. Bull. Insights from mixture cure modeling of molecular markers for prognosis in breast cancer. *Journal of clinical oncology*, 31(16):2047–2054, 2013.

J. C. Yue, H.-C. Wang, Y.-Y. Leong, and W.-P. Su. Using taiwan national health insurance database to model cancer incidence and mortality rates. *Insurance: Mathematics and Economics*, 78:316–324, 2018.

Y. Zhan, X.-R. Liu, C. A. Reynolds, N. L. Pedersen, S. Hägg, and M. S. Clements. Leukocyte telomere length and all-cause mortality: a between-within twin study with time-dependent effects using generalized survival models. *American Journal of Epidemiology*, 187(10):2186–2191, 2018.