Rethinking Mortality: A State-Based Dynamic Probabilistic Modelling Approach Using National-Scale Health Data

Pramo Samarasinghe

A thesis submitted for the degree of
Bachelor of Computing (Honours)
The Australian National University

October 2025

Except where otherwise indicated, this thesis is my own original work.

Pramo Samarasinghe
30 October 2025

To my family; Ammi, Thaththi, Podi, Sudu, and Aththammi, whose love anchored every step.

And to my loving partner, Chamin, your unwavering faith and courage made hard days more bearable.

Acknowledgments

I would like to begin by thanking to *Dr Aaron Bruhn* for being the reason I was able to find such an influential project. Thank you for dedicating the time and effort to help me find a project that aligned with all my interests. Your guidance at the very beginning set the stage for everything that followed.

Next, I would like to sincerely thank my supervisors *Fei Huang*, *Francis Hui*, and *Andres Villegas Ramirez*. I would not have made it through this year without your unwavering support, encouragement, and belief in me. There were so many moments when the workload felt impossible, when the deadlines loomed too close, and when my confidence wavered; but each time, your guidance pulled me through. You never once made me feel like my questions were too small or my worries too trivial. Beyond your incredible intelligence and expertise you brought a level of kindness and humanity to supervision that made the entire experience so much more meaningful. The time you took to explain concepts in detail, to offer constructive feedback, and to celebrate small wins made me feel seen and valued as a student. I am especially grateful for the patience and understanding you showed whenever things became overwhelming, for the way you adapted meetings when I was struggling, and for the genuine interest you showed in both the project and my growth as a researcher. **Working with you has been the most rewarding part of this degree.** I feel so lucky to have had supervisors who were not only brilliant academics but also exceptional mentors and role models. I would also like to thank *Prof. Lexing Xie* for giving me the opportunity to be her student this year, thank you for making the time whenever I reached out.

I would also like to extend my thanks to *Guy Thorburn* for his support on the project. Thank you for making the time to chat about the direction of the project and provide guidance when I was stuck. Your open door (whether for tedious administrative issues or complex project questions) meant so much, especially knowing how busy your schedule was.

Additionally, to the *DataLab team*, thank you for all the effort and time put into my project and for the constant support with the clearance requests.I would also like to extend my gratitude to *Shamim*, whose work on mapping the medical codes into a more interpretable format was invaluable. Your contribution made it possible to produce clear and meaningful results, and I am deeply grateful for your support.

My heartfelt gratitude also goes to the friends who supported me despite my frequent absences, last-minute cancellations, and tendency to rarely reach out. To *Jen*, who unfailingly wished me luck every Friday morning before my meetings with my supervisors, your thoughtfulness never went unnoticed. To *Chamali*, *Nipun*, *Arabella*, and *Chamathka*, thank you for believing in me from afar; your encouragement meant

the world to me. To *Thomas* and *Georgia*, thank you for always making time to catch up, for reaching out even when I went quiet, and for giving me the therapy I did not know I needed over countless lunches and dinners. To *Pamilla* and *Daisy* thank you for staying in touch despite my long silences, I am truly grateful beyond words. To *Carolina*, *Raymond*, *Xylia* and *Andrew* thank you for making the time to study with me, it was so much less tedious with you there. I really appreciate that you made the time to keep me company. To *Tommy*, *Tait*, *Joey*, *Shantha*, and *Sidd* thank you for constant check ins, the fun hangouts and always being ready to listen to me rant.

I would also like to thank my colleagues at the *Australian Government Actuary (AGA)*, who became friends over the years: thank you for the board game nights, coffee catch-ups, and warm messages checking in on me when I disappeared for months at a time. Your kindness and friendship made a huge difference.

To my *uncle and his family; mami, nandhi, loku malli and chooti malli*, thank you for your generosity in hosting me and looking after me; I would not have had the opportunity to study at ANU without your support. Your kindness in opening your home to me made the transition to university life so much easier, and I will always be grateful for the countless ways you helped me feel settled and supported. I am also grateful to *Thilani Aunty, Dulan uncle, Binu, Tehan and achchi and seeya* for always checking in and reaching out, reminding me that there was a caring community around me even when I felt alone. You always make me feel like I make a difference and I feel so loved. To *Nadeeka Akki, Ratnayaka Mama, Loku Nandha*, and *especially Sayuni Nangi*, thank you for your constant love and encouragement. Your faith in me, and your messages and calls, have meant more than I can ever express. It always makes me smile to see your notifications pop up while I'm working through assignments and deadlines.

To *Chamin's parents*, *Amme* and *Thaththe*, thank you for your incredible generosity in letting us live rent-free in your home. Thank you for all the times you saved me from the burden of cooking and for staying on top of all the life admin tasks that made things so much easier for us. Your kindness and support meant more than I can say. I am also grateful to *Evani* for making the time to go out and have fun with us throughout the year, despite how busy you were. I am so lucky to have family like you.

To my parents, *ammi*, *thaththi* it feels like you sacrificed two decades to watch me get where I am today. I know none of this would have been possible if not for you two, thank you for your unconditional love and support. I feel like the debt I owe you continues to grow regardless of my best efforts. You both have been my role models all my life, I love how driven you are in terms of education and your career, and I can only hope to live up to the example you've set. I feel so lucky to have grown up surrounded by your love, guidance, and belief in me. People come and go all the time but the four of you have been my constants You two mean the world to me! To my sisters, *podi* and *sudu*, you have been my free therapists and my biggest cheerleaders for as long as I can remember. *Podi*, thank you for always understanding me better than anyone else ever could, and *sudu*, thank you for always knowing exactly how to make me laugh, no matter how stressed or tired I was. There has never been a dull

moment with you two. I often think about how we don't get to choose the family we're born into, but if I had the choice, it would always be you guys. Thank you for being my best friends and for being with me through it all. To my grandparents, thank you for all the sacrifices you made to look after us especially *aththammi* who retired early to look after me. Thank you for raising me. To my late grandfather, *seeya*, who cried every time I left for university. The time I spent chasing my dreams came at the cost of time with you. I hope you know that I will forever miss you.

Finally, to my husband *Chamin*, I do not know how to begin to thank you. This year would have been so much harder without you by my side. Thank you for always believing in me when I doubted myself, for cheering me on when my motivation ran dry, and for being the person I could always rely on through the highs and lows. I know there were nights when my insomnia meant neither of us got much sleep, but you stayed with me anyway, talking through my worries until they felt a little lighter. You have been there through every draft, every stress craving, and every moment I thought I would never finish this thesis; proofreading my work, making me laugh when I needed it most, and reminding me to take breaks when I refused to, whether its a walk or an episode or a shopping date thank you for being with me through it all. Most of all, thank you for your patience, kindness, and love over these months. I cannot wait to finally enjoy life after university together. We have been talking about it for so long, and I still cannot believe it is almost here. I feel so lucky to have had you beside me through all of this, and I could not have done it without you. More than anything, thank you for making me feel loved and even the hardest days.

Abstract

Australia's retirement income system is undergoing a significant structural shift, from defined benefit schemes to defined contribution schemes. Under defined benefit schemes, longevity risk was borne by employers or the government, guaranteeing retirees a stable lifetime income. In contrast, defined contribution schemes place that risk on individuals, who must manage their savings to last through retirement.

This shift makes determining the retirement planning horizon, the timeframe over which individuals plan for their financial needs in retirement, central to decisions on drawdown strategies, annuity purchases, and the balance between private savings and Age Pension support. Life expectancy is an intuitive planning horizon widely used by households and financial planners, representing the expected duration of financial need.

Most planning horizons, including life expectancy, depend on accurate mortality modelling, which is the focus of this thesis. Recent policy developments have further reinforced the importance of accurate and equitable mortality modelling in retirement planning. This thesis responds to that need by developing a health-informed framework for mortality modelling suited to Australia's evolving retirement system.

In demography and actuarial science, mortality is defined as the age-specific probability of death, specifically the probability that an individual aged x will die before reaching age $x+1$. Traditional approaches to mortality prediction rely predominantly on age and gender. Although straightforward, these models are narrow in scope and overlook key factors such as health status, socioeconomic conditions, and access to healthcare. This is the case for Australia, where mortality rates are estimated primarily using the Australian Life Tables (ALT), which reports mortality for any given age and gender cohort using their observed proportion. Although effective for describing population-level patterns, this leads to oversimplified and inequitable assessments of mortality risk, as individual mortality within each cohort substantially deviates from the average. In particular, when purchasing retirement income products, this causes individuals with poorer health than their cohort's average subsidising for those with better health. This is because they both pay for the same estimated mortality, yet the former typically experience higher mortality and thus do not receive the full value of their retirement income.

In this thesis, we address that gap by incorporating health-related variables into mortality modelling. These variables are used to split age and gender based cohorts into health-based cohorts that are designed to support more accurate and equitable mortality assessments. A key objective is to ensure that the established modelling approach remains interpretable and transparent for practical use in insurance and retirement planning. Thus, the resulting health-based cohorts should also be able to be translated into a small set of health-related questions.

In order to incorporate health-related variables, we use the Personal Level Integrated Data Asset (PLIDA), the Australian Bureau of Statistics' (ABS) secure, de-identified dataset of the Australian population on data relating to public policy. To the best of our knowledge, this is the first work in Australia to integrate detailed health information directly into a mortality improvement framework, making this research a novel contribution to the field.

We investigate many modelling approaches that are standard in the literature, and try to adapt them to determine appropriate health-based cohorts. These include K-means clustering, decision and survival trees, and linear regression. We identify a limiting factor of these standard approaches for our task, namely that they assume static membership to a cohort over time, so the tendency of death records to reflect end-of-life health conditions introduces bias. While this could be mitigated by incorporating a longer historical tail of data, the available training data covers a relatively short period, and it would introduce the limitation that older mortality experience may be less relevant to current or future populations.

Thus, we further propose a non-standard Markov-chain modelling approach, which is able to capture mortality dynamics (transitions between cohorts). By explicitly modelling transitions between health states alongside mortality, the framework provides a more realistic representation of disease progression and survival patterns among older Australians, reducing the bias of end-of-life health conditions from death records. Our approach is scalable to the entire population or any required subset, and is also adaptable to new data.

Incorporating the proposed, transition-based survival estimates has meaningful financial implications both for retirees and for the government. For a cohort of 100,000 retirees planning based on average life expectancy, the proportion expected to outlive that horizon falls from 9.9% with ALT-based predictions to 6.7% with ours, or around 3,200 fewer people. If each avoided case equates to at least one additional year of Age Pension payments estimated at $\approx$ \$500 per week, total pension expenditure would decrease by about \$1.6 million per week ($\approx$ \$83.2 million per year). At the household level, this adjustment reduces weekly withdrawals slightly but extends the period of private income and delays reliance on the Age Pension. Moreover, a stronger predictive performance from our proposed model implies that the drop in the standard of living seen when transitioning to Age Pension becomes significantly lower, thus ensuring that the individual has a guaranteed period of income for longer (provided that Age Pension is income and asset contingent). Using the model life expectancy also implies the expected period in which you depend on Age Pension reduces from 4.31 years to 3.57 years.

Although the primary focus of this study is on retirees, our proposed transition-based framework is flexible enough for broader population contexts, where defined health states would ideally vary by age group to reflect different patterns of morbidity and mortality throughout life. It could also be adapted for other insurance products and policy applications, with states and variables tailored to the specific data available and the health or demographic factors most relevant to the population under consideration.

Contents

List of Figures

List of Tables

Glossary

Age Pension Australia's government-provided income support for eligible retirees, paid subject to means tests on income and assets.

Annuitant The purchaser of an annuity product.

Annuity A financial product that pays a fixed stream of income for a specified period or for the duration of life (an example of a annuity product is a lifetime annuity)..

Co-morbidity The co-occurrence of two or more health conditions in the same individual at the same time.

Cross-subsidy A financial situation in which one group of individuals effectively pays more than the cost of the service they receive, thereby subsidising another group that pays less than the cost of the service they consume. In the context of retirement income and actuarial modelling, cross-subsidies occur when pooled pricing assumptions (such as average mortality rates) lead healthier individuals to subsidise the benefits of less healthy individuals. Reducing cross-subsidies is a key motivation for improving mortality models, as more accurate, health-informed predictions can ensure that individuals are charged premiums and receive benefits more closely aligned with their actual risk.

Defined benefit A retirement plan that promises a specified benefit, typically based on salary and years of service. The plan sponsor bears both investment and longevity risk.

Defined contribution A retirement plan where contributions are made to an individual account. The member bears investment and longevity risk, and the eventual benefit depends on the accumulated balance.

Discount rate The interest rate used to convert future cash flows into present value.

Drawdown The planned withdrawal of money from a retirement account or superannuation balance to provide income, typically following a set schedule and minimum rates.

Inequitable In general inequitable means unfair. It is used to describe a situation or system that is biased or discriminatory.

Life expectancy Life expectancy at age x is the expected number of years a person aged x will live. The relevant calculations are explained under Section 5.3.

Lifetime annuity product An insurance contract that pays a stream of income for as long as the annuitant (or covered lives) survives, often with options such as joint-life, guaranteed periods, escalation, and value-protection riders. Prices reflect survival probabilities, interest rates, expenses, and capital requirements.

Longevity risk The risk that people live longer than expected, causing income streams or reserves to be inadequate (e.g., retirees outliving savings, or insurers/super funds underpricing lifetime benefits).

Morbidity The state of ill health or disease within an individual or population, often expressed as incidence or prevalence over a specified period.

Present value The value today of a future cash flow discounted at the appropriate rate.

Retirement income product An arrangement designed to convert accumulated savings into income during retirement. Examples include lifetime and fixed-term annuities, drawdown/decumulation accounts, collective/pooled arrangements (e.g., CDC or tontine-style products), and hybrids with guarantees or spending rules. Design choices determine the allocation of investment, interest-rate, and longevity risks between the provider and retiree.

Superannuation Australia's compulsory retirement savings system where employers (and sometimes members) contribute to super funds that invest on members' behalf.

Term Insurance Insurance that guarantees payment on death within a predetermined time frame (term).

Time value of money The principle that a sum of money today is worth more than the same sum in the future because of its potential to earn returns over time.

Chapter 1

Introduction

1.1 Motivation and Context

For decades, retirement income in Australia was provided through defined benefit (DB) schemes. Under this structure, employers guaranteed employees a stable income stream from retirement until death, typically calculated based on years of service and final salary. Because employers bore the longevity risk (the financial responsibility if retirees lived longer than expected), they had a strong incentive to model life expectancy with care and precision.

While individual lifespans remain uncertain, for large employee groups the law of large numbers ensures that the average lifespan across the group becomes increasingly predictable as the number of retirees grows. This statistical stability enabled employers and pension funds to distribute individual longevity risk across the cohort, resulting in more reliable and manageable long-term liabilities. Consequently, the current Australian Life Tables (ALT) [Australian Government Actuary, 2024] were considered adequate, as they provided a sound basis for the effective operation of DB schemes.

Recently, Australia has shifted almost entirely to a defined contribution (DC) system. Under this arrangement, the employer's obligation ends once mandatory contributions are made to the employee's superannuation account. Upon retirement, the accumulated balance, comprising employer and employee contributions as well as investment earnings, must sustain the retiree throughout their remaining lifetime. Retirees therefore face the complex task of determining how best to invest and withdraw their savings, whether through account-based pensions, annuities, or other products. They may also choose to withdraw their balance as a lump sum at retirement [Commonwealth Superannuation Corporation, 2025], emphasising the need for accurate life expectancy projections in informing such financial decisions.

The shift from DB to DC fundamentally changes the problem: instead of employers bearing the risk of underestimating life expectancy, individuals now carry the burden of managing their savings amid uncertainty about how long they will live. Sound financial planning therefore relies on accurate mortality predictions at a more fine grained level, especially for retirees.

Recent policy developments have further reinforced the importance of accurate and equitable mortality modelling in retirement planning. The *Retirement Income*

Covenant [Australian Government Treasury, 2022] and the *Treasury Guidance on Best Practice Principles for Superannuation Retirement Income Solutions (2025)* [Australian Government Treasury, 2025] emphasise the need for superannuation trustees to help members balance income adequacy, flexibility, and risk throughout retirement. These reforms coincide with Australia's transition to a mature superannuation system, where individuals increasingly rely on private savings rather than defined benefit guarantees.

Currently, in Australia, life expectancy estimates are largely based on age and gender alone, using population-level mortality tables such as the ALT. These tables provide a useful benchmark, but they ignore individual-level information such as health conditions and medication history, which are especially important in retirement when chronic illness and multi-morbidity become key drivers of mortality.

Global evidence supports this concern. For instance, Swiss Re (a major reinsurance company) has warned that many insurers make longevity predictions without incorporating rich health and clinical data [Swiss Re, 2023]. Their analysis suggests that focusing only on cause of death or demographic factors underestimates the complexity of health trajectories leading to death.

This gap in prediction accuracy has real-world financial consequences. Underestimating life expectancy can leave retirees vulnerable to outliving their savings, increasing dependence on the government's Age Pension. Indeed, despite the growth in superannuation balances, reliance on the Age Pension remains high [Australian Bureau of Statistics, 2024]. Thus, there is an urgent need to improve the mortality models for retirees so as to:

1. Support more accurate and personalised financial planning at the individual level,
2. Enable fairer and more sustainable pricing for longevity-linked products such as annuities, and
3. Reduce reliance on the Age Pension system arising from the use of inaccurate life expectancy assumptions in retirement planning.

This research addresses the gap by integrating administrative health datasets (including the Medicare Benefits Schedule (MBS) and Pharmaceutical Benefits Scheme (PBS)) to build disease and medication centred mortality models for Australian retirees. By linking health conditions, medical interventions, and medication histories to mortality outcomes, this study aims to improve predictive accuracy and support more informed financial decision-making for retirees, policymakers, and insurers. Ultimately, the findings have implications for both individual financial planning and national policy through potential reductions in Age Pension reliance.

1.2 Thesis Statement

This thesis argues that disease-centred mortality models, which incorporate individuals' health conditions, medical procedures, and medication histories, offer significantly enhanced predic-

tive accuracy for retiree mortality compared to conventional models that rely solely on age and gender.

1.2.1 Research Questions and Contributions

This research addresses the current gap in Australian mortality modelling by incorporating health-based predictors for retirees. The main research questions and contributions are summarised in Section 1.2.1.1 and Section 1.2.1.2 respectively.

1.2.1.1 Research Questions

1. Are there clear patterns in annual mortality (q_x) associated with health-related factors such as disease prevalence, medication usage, and medical interventions?

2. Which broad health states explain the most significant deviations in mortality rates around retirement age?

3. To what extent does incorporating health and medication data enhance mortality prediction accuracy compared to the age-sex baseline of the Australian Life Tables?

4. Can cohort-level models (by age, sex, and health state) improve predictive performance, and which modelling approach performs best?

5. How can the information required for the model be translated into a standard questionnaire that can be administered at the time of retirement or when purchasing an insurance product?

1.2.1.2 Key Contributions

1. **Exploratory Data Analysis (EDA):** A detailed examination of the MBS and PBS datasets was carried out to understand how health-related variables influence mortality outcomes. This step provided the empirical foundation on which the modelling work was built.

2. **Cohort-Based Mortality Modelling:** A cohort-based framework for estimating q_x was developed through the integration of administrative health data and the evaluation of multiple modelling approaches. The aim was to improve predictive accuracy while maintaining transparency and interpretability.

3. **Validation Against Benchmarks:** All models were trained on data from 2011–2013 and evaluated against outcomes from 2016. This approach reflects the real-world delay between the availability of health and mortality data and their use in predictive modelling, providing a more realistic test of model performance. All models were tested against the ALT, which acts as the current benchmark for mortality study within Australia. The final model provided in this thesis

utilises 2014-2016 data to ensure that the latest data has been incorporated in the model output.

4. **Financial Impact Assessment:** Modelling improvements were translated into monetary terms by estimating their effect on the expected present value (EPV) of annuities and retirement drawdown strategies. The analysis also shows how projected lifespans differ from those based on current ALT.

5. **Dynamic Health State Modelling:** Static models were extended to incorporate transition probabilities between health states throughout retirement. This allows mortality estimates to adjust over time in response to changes in health status, capturing the dynamic nature of ageing and disease progression.

1.3 Thesis Scope

The scope of this research is to use the Person Level Integrated Data Asset (PLIDA) [Australian Bureau of Statistics, 2025] for the period 2011-2016 to incorporate health related variables into mortality predictions. This study does not delve into demographic factors as they have already been examined in Huang et al. [2023]. Several modelling and data-related limitations arose due to privacy constraints on the dataset, which have been detailed in Section 4.10.

1.4 Ethics Statement

The research adheres to the Australian Bureau of Statistics (ABS) DataLab Principles, ensuring that all data used in the study is de-identified in line with strict data privacy standards.

Additionally, all outputs generated from the analysis were vetted by the ABS DataLab to ensure no sensitive or personally identifiable information was included. This process ensures that data analysis follows ethical guidelines and complies with the National Statement on Ethical Conduct in Human Research (2007) [National Health and Medical Research Council, 2025], issued by the National Health and Medical Research Council.

1.5 Thesis Outline

The thesis is organised as follows:

- **Chapter 1: Introduction** Motivates the problem, sets out research questions, and summarises contributions.

- **Chapter 2: Literature Review and Gap Analysis** Reviews existing approaches to mortality modelling, highlights limitations in current Australian practice, and identifies the research gap addressed in this thesis.

- **Chapter 3: Exploratory Data Analysis** Examines MBS and PBS variables, identifying key factors in health and medication that influence mortality.
- **Chapter 4: Proposed Modelling Approaches** Describes the modelling approaches, including cohort-based methods and dynamic health state models, along with validation strategies.
- **Chapter 5: Model performance** Analyses different models introduced in the previous chapter against selected metrics and benchmarks.
- **Chapter 6: Results and Applications** Presents empirical findings, evaluates model performance against benchmarks, and quantifies financial implications for retirement planning.
- **Chapter 7: Discussion and Conclusion** Summarises findings, and discusses policy implications, limitations, and avenues for future research.

Appendices provide additional technical material, supplementary figures, and proofs.

Chapter 2

Background and Related Work

In this chapter, we provide the necessary background and contextual foundation for this research. Section 2.1 outlines the motivation behind the study, highlighting limitations of traditional mortality prediction models and the need for a more comprehensive approach that incorporates health-related variables. Section 2.2 presents the background information required to understand this thesis, while Section 2.3 reviews existing literature in the field. To further establish the research contributions, Section 2.6 provides a detailed gap analysis, identifying areas where current methodologies fall short and highlighting the potential improvements offered by the proposed approach.

2.1 Motivation

The motivation for this research stems from the need to improve mortality modelling in financial products, particularly annuities and life insurance products for the retirement population. As individuals age, health factors become increasingly significant, making it essential to develop health-driven mortality models aimed at improving predictive accuracy. Traditional mortality models, such as the ALT, often rely on historical population data and demographic factors such as age and gender. However, while these variables remain important, health plays a critical role in determining mortality, especially in later life, and should therefore be a central component of mortality modelling for retirees.

The consequences of inaccurate mortality predictions are significant. Retirees with health conditions may be disadvantaged by annuity products that assume overly optimistic lifespans, leading to lower periodic payouts. Conversely, healthier individuals face longevity risk, where they outlive their expected lifespan, leading to financial instability in old age. These imbalances highlight the need for more precise mortality models that account for individual health characteristics, particularly in the retirement population where health disparities significantly impact financial outcomes.

By incorporating health datasets and modern data-driven techniques, this research aims to advance mortality modelling practices. Improved accuracy and fairness in these models can lead to better financial planning, enhanced consumer pro-

tection, and more sustainable insurance and pension systems. This work will contribute to both actuarial science and data-driven decision-making, emphasising the importance of fairness and accuracy in financial products.

This research lies at the **intersection of comorbidity and mortality studies**, examining how multiple coexisting health conditions influence overall mortality and how their interrelationships shape survival outcomes.

2.2 Background

2.2.1 Explanation of Mortality Tables

A life table, also known as a mortality or actuarial table, is a statistical tool used to describe the mortality characteristics of a population. For each age x, a life table provides the probability q_x that a person aged x will die before reaching age $x+1$.

Key columns typically include:

- x: Age
- q_x: Probability of dying between age x and $x+1$
- l_x: Number of survivors at age x, from an initial cohort of 100,000 individuals
- d_x: Number of deaths between age x and $x+1$
- e_x: Life expectancy at age x,
- p_x: Probability of surviving from age x to age $x+1$
- μ_x: Central (average) death rate at age x
- L_x: Total number of person-years lived between age x and $x+1$
- T_x: Total number of person-years lived by the cohort from age x onwards

For instance the 2020-22 Male mortality table released by the Australian government actuary [Australian Government Actuary, 2024] is shown in Figure 2.1.

Note that a life table is constructed using q_x for a selected initial cohort size (usually 100,000). In other words, empirical data are used to calculate q_x, which then forms the basis to calculate the rest of the mortality table. Thus, the primary objective of this research is to improve the estimation of q_x. As q_x itself is less interpretable, the results of the model were translated into life expectancy at age 65, which we denote throughout this study as e_{65}. Furthermore, the effect of these estimates on annuity pricing was assessed to quantify the direct implications for retirement income.

Traditional life tables are built using age and gender data from large populations. While they are widely used in insurance and pension systems, their main limitation lies in the assumption of population homogeneity. They do not account for variations in individual health status, lifestyle choices, or socio-economic factors, all of which

AUSTRALIAN LIFE TABLES 2020–22: MALES

Age	l_x	d_x	p_x	q_x	μ_x	$\mathring{e}_x$	L_x	T_x
0	100,000	339	0.996606	0.003394	0.000000	81.31	99,832	8,131,434
1	99,661	20	0.999801	0.000199	0.000207	80.59	99,651	8,031,602
2	99,641	17	0.999829	0.000171	0.000188	79.61	99,632	7,931,952
3	99,624	12	0.999876	0.000124	0.000148	78.62	99,617	7,832,320
4	99,611	9	0.999908	0.000092	0.000105	77.63	99,607	7,732,703
5	99,602	8	0.999922	0.000078	0.000083	76.64	99,598	7,633,096
6	99,594	7	0.999926	0.000074	0.000075	75.64	99,591	7,533,498
7	99,587	7	0.999926	0.000074	0.000074	74.65	99,583	7,433,907

Figure 2.1: Australian Life Tables 2020–22 for Males, showing mortality data by age group, including the number of deaths, survival probabilities, and life expectancy.

can significantly influence mortality outcomes. This research aims to enhance life table construction by leveraging machine learning methods and health data from the Personal Level Integrated Data Asset (PLIDA), allowing for a more tailored and equitable approach to mortality prediction.

2.2.2 Applications of Mortality Rates

The application of mortality data extends beyond analysing disease-specific trends and supporting clinical research. It plays a key role in a wider range of actuarial practices, including insurance pricing, reserving, policy making, and retirement planning.

2.2.2.1 More Common Mortality Measures

A commonly used measure is hospital mortality [Australian Commission on Safety and Quality in Health Care, 2025], which refers to the rate of in-hospital deaths as a proportion of total admissions. This indicator is often used to assess the quality and performance of in-hospital care.

Disease-specific mortality [van Leeuwen et al., 2010; Cancer Australia, 2025] refers to the proportion of deaths of individuals with a certain diagnosis relative to the total population at risk. Trends in disease-specific mortality are used to identify growing medical risks and to prioritise areas of focus in medical research. A related measure, Case Fatality Ratio (CFR) [Harrington and of Encyclopaedia Britannica, 2020], represents the proportion of diagnosed people with a certain condition who die within a defined period. While disease-specific mortality quantifies the population-level impact of a disease, the CFR measures the risk associated with the disease for an affected individual.

2.2.2.2 Population Mortality and its Use

The mortality referred to in this research is **age and gender specific mortality**, which is the mortality calculated for each combination of age and gender. This denotes the probability that an individual of a given age and gender, will die within a given year. As the main goal of this research is to improve the accuracy of such estimates, most proposed models first partition the data using health-based determinants, before modelling age and gender-based mortality within each subgroup.

The ALT, which use age and gender specific mortality, are widely applied across public policy, insurance, Australian statistics, and actuarial modelling applications, including annuity pricing, reserving and capital modelling. Their uses range from direct applications, such as population forecasting [Australian Bureau of Statistics, 2023], to more *hidden* ones such as estimating default rates for long-term bank loans. In principle, any long-term financial forecast must account for the possibility of death, making the applications of mortality modelling extensive.

Despite these advances, most standard mortality analysis still assumes population homogeneity and focuses on aggregate outcomes. In pricing, this implicitly relies on the Law of Large Numbers: with enough members, pooled experience should converge to the aggregate estimate. However, as detailed in Section 1.1, following the shift from defined benefit to defined contribution, it is important to adopt health-informed models that minimise cross-subsidies and provide fairer, and more accurate outcomes.

2.2.3 Australian Retirement System

Australia's retirement income framework is built upon three pillars [Australian Treasury, 2020b]:

1. a means-tested Age Pension,
2. compulsory superannuation,
3. and voluntary private savings.

A major structural shift in Australia's retirement income system occurred in the late twentieth century, when the dominant model transitioned from **defined benefit (DB)** schemes to **defined contribution (DC)** schemes. Under the DB scheme, the employer guarantees a pension to their employees from retirement until death, bearing both investment and longevity risks. On the other hand, in the DC scheme, employers contribute a fixed proportion of an employee's income throughout their working life, with no guarantee of benefits beyond retirement. One key reason for the transition from a DB to a DC scheme was employers' desire to reduce their exposure to longevity and investment risks, while improving profitability [Australian Treasury, 2020b; Australian Prudential Regulation Authority, 2019]. The introduction of the *Superannuation Guarantee* in 1992 further cemented this shift, as new contributions were directed into DC accounts rather than employer-guaranteed pensions [Parliament of Australia, 1992].

The shift from DB to DC has numerous implications for how mortality is modelled and applied. Under a DB framework, retirement income was largely guaranteed, and pricing relied on population-level mortality estimates that were expected to average out under the Law of Large Numbers. In contrast, the DC system places responsibility for retirement outcomes on individuals, turning mortality estimation into an *individual prediction problem*. As a result, improving mortality modelling, particularly by incorporating health status and co-morbidity, is essential for reducing cross-subsidies, improving retirement income adequacy, and supporting evidence-based policy design [Organisation for Economic Co-operation and Development, 2022; Australian Treasury, 2020a].

Today, Australia's superannuation system is one of the largest in the world, with assets exceeding the national GDP, with the expectation of becoming the second largest superannuation industry globally by 2030 [SMC Australia, 2024]. This growing industry is subject to close regulatory oversight by the Australian Prudential Regulation Authority (APRA) [Australian Prudential Regulation Authority, 2019] and the Australian Securities and Investments Commission (ASIC) [Australian Securities & Investments Commission, 2025]. Recent reforms reflect a shift in policy focus toward improving outcomes during the *retirement phase*, rather than merely emphasising on wealth accumulation [Australian Government, 2022]. The *Retirement Income Covenant*, introduced in July 2022, mandates that superannuation trustees develop and publicly release strategies to assist members in maximising income, managing key risks, and maintaining flexible access to their funds in retirement [Australian Government, 2022; Australian Treasury, 2020a].

This growing emphasis on retirement outcomes reinforces the need for more accurate and personalised mortality modelling, as improving the estimation of life expectancy is fundamental to effective income planning, product pricing, and long-term sustainability within the superannuation system.

2.3 Related Work

The research problem addressed in this study lies at the intersection of the following key areas (Figure 2.2):

1. General mortality modelling,
2. Australian mortality modelling,
3. Health based mortality studies and
4. Dynamic mortality modelling including Markov chains for mortality modelling.

Accordingly, this review examines the literature across each of these domains in more detail. For the selection of the relevant literature, four initial searches were conducted; mortality research, Australian mortality research, health-based mortality

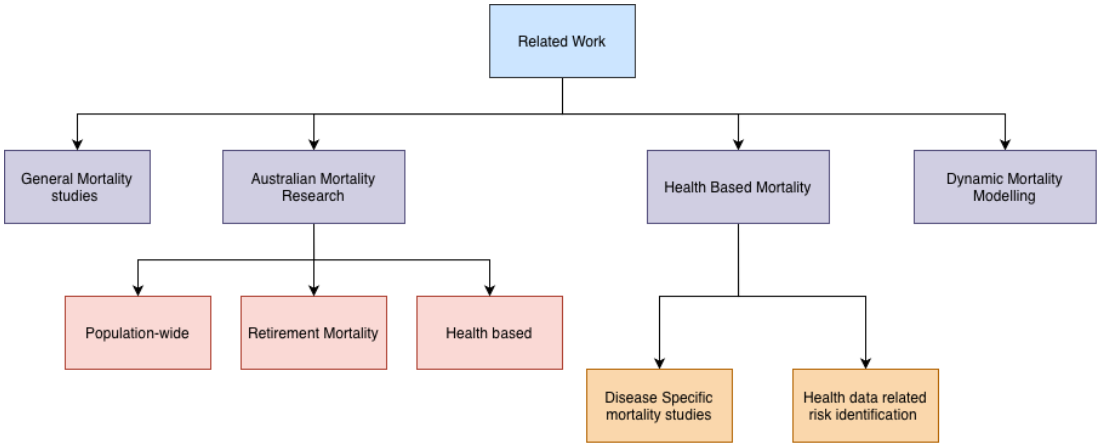


Figure 2.2: Areas of Research.

and dynamic mortality. Each search applied distinct inclusion and exclusion criteria as shown in Figure 2.3.

Originally, this research was aimed to explore the use of machine learning and explainable AI for the modelling task. However, due to the absence of a reasonable target variable (explained in detail in Section 4.7.1), these supervised machine learning models were not applicable to this research.

2.3.1 Mortality Research

General mortality research is a long established field within demography, public health, and actuarial science. The earliest theoretical approach to mortality modelling can be traced back to De Moivre [1725], who proposed a model assuming a uniform distribution of deaths and demonstrated simplified techniques for calculating annuities. This paper released in 1725 is the beginning of the ongoing field of mortality study and modelling.

Following the uniform distribution assumption in 1725, about a century later the actuary Benjamin Gompertz advanced the field further by introducing a mathematical model in which mortality rates increase exponentially with age [Gompertz, 1825], thus laying the foundation for many of the models used today.

The field saw a major shift in the 1990s (about another century later) with the introduction of the stochastic Lee-Carter [Lee and Carter, 1992a] model, which became a standard for mortality projections and sparked further research into extensions and new models such as the Cairns-Blake-Dowd (CBD) [LLP] model. Over time, the scope of mortality research has broadened significantly. The key mortality research conducted, and the relevant model formulas for the models listed in Figure 2.4 are summarised in Table 2.1.

Modern mortality modelling has enhanced understanding and prediction of mortality patterns further. These modelling approaches are not limited to producing basic life tables, but they explore the underlying biological [Li et al., 2023], be-

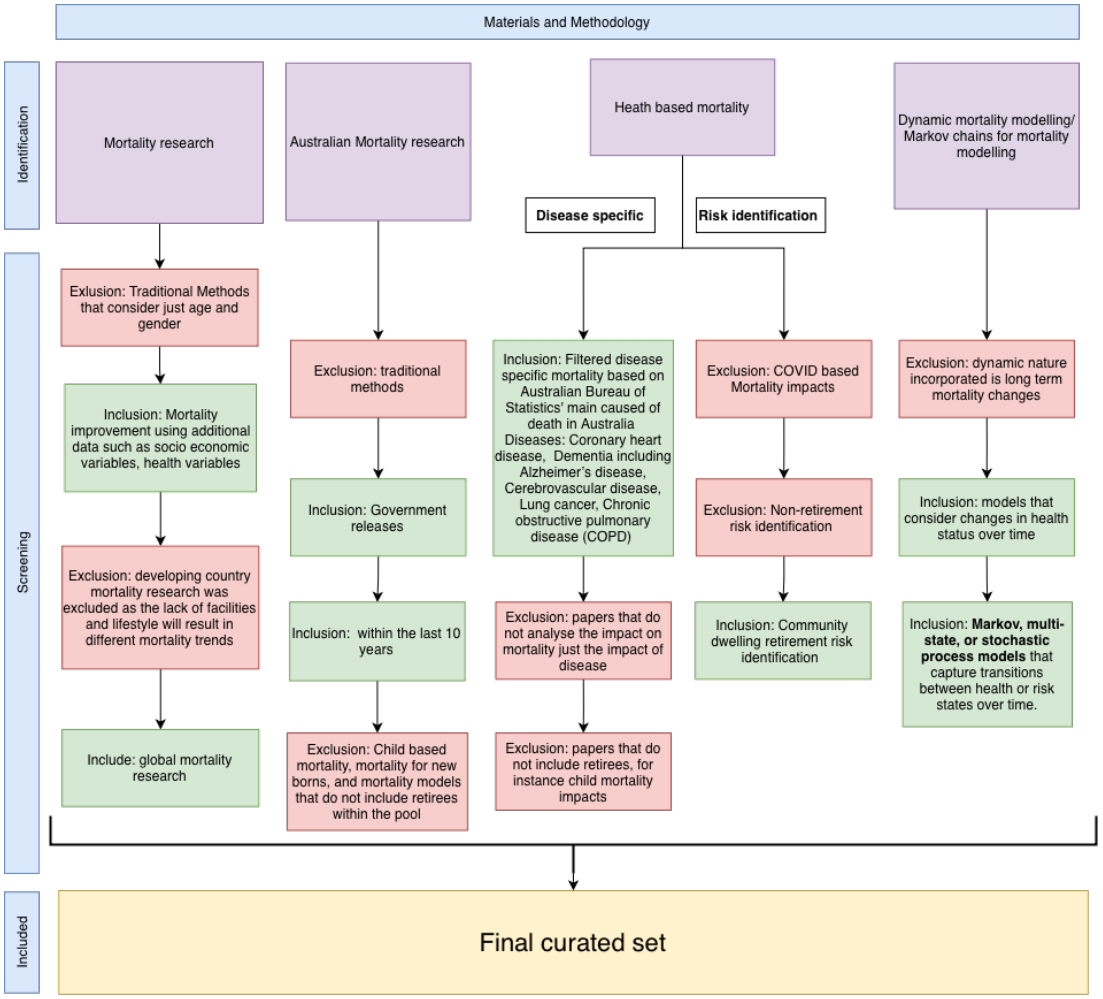


Figure 2.3: A conceptual flowchart outlining the inclusion and exclusion criteria applied during the literature selection process.

Figure 2.4: History of Mortality Modelling [Pascariu et al., 2018].

Table 2.1: Key Historical Mortality Models and Their Formulations [Pascariu et al., 2018].

Author (Year)	Model	Reference
De Moivre (1725)	$\mu(x) = \frac{1}{\omega - x}$	De Moivre [1725]
Gompertz (1825)	$\mu(x) = Ae^{Bx}$	Gompertz [1825]
Makeham (1867)	$\mu(x) = Ae^{Bx} + C$	Makeham [1867]
Thiele (1871)	$\mu(x) = A_1 e^{-B_1 x} + A_2 e^{-B_2 x} + A_3 e^{B_3 x}$	Thiele [1871]
Perks (1932)	$\mu(x) = \frac{A + BC^x}{BC^{-x} + 1 + DC^x}$	Perks [1932]
Weibull (1951)	$\mu(x) = \frac{\beta}{\alpha}\left(\frac{x}{\alpha}\right)^{\beta-1}$	Weibull [1951]
Siler (1979)	$\mu(x) = ae^{-bx} + c + de^{ex}$	Siler [1979]
Heligman–Pollard (1980)	$q_x = A^{(x+B)^C} + D\,e^{-E(\log x - \log F)^2} + \frac{GH^x}{1+GH^x}$	Heligman and Pollard [1980]
Lee–Carter (1992)	$\log m_x(t) = a_x + b_x k_t + \epsilon_{x,t}$	Lee and Carter [1992a]

havioural [Australian Bureau of Statistics, 2023; Salisbury et al., 2014; Fernández-Ballesteros et al., 2022], and social factors [Huang et al., 2023; Zissimopoulos et al., 2021] that shape survival outcomes. Demographers focus on long-term population trends such as ageing, fertility decline, and improvements in healthcare that extend life expectancy. Actuarial science uses mortality modelling to price financial products, manage longevity risk, and design sustainable retirement systems. Public health studies investigate the influence of disease, socioeconomic status, healthcare access, and lifestyle factors on mortality patterns, aiming to identify interventions that reduce premature death.

Recent advancements in mortality prediction over the past decade have been driven by improvements in computational power, availability of higher-quality data, and advanced modelling methodologies. Hunt and Villegas [2017] made an important contribution by shifting the focus from modelling raw mortality rates to modelling improvement or deterioration. This approach reduces forecast bias along longevity trends and provides better coverage of mortality dynamics over time. Modern models increasingly consider *age–period–cohort* dynamics, structural breaks, and cohort effects to more accurately capture mortality differences across generations [Li and Wong, 2018]. The COVID-19 pandemic highlighted the need for methods capable of accommodating major shocks, resulting in the development of prediction models that are resilient to outlier years [Haberman and Millossovich, 2022].

In addition, flexible semi and non-parametric approaches such as *Generalised Additive Models (GAM)* have emerged as powerful tools to capture non-linear mortality trends and variation across populations [Wang and Sherris, 2022]. In order to have more realistic models with explicit adjustments for uncertainty in recent mortality data, an extension to the classic Bayesian framework has been designed. For example, [Wang and Zhou, 2023] embed rare but impactful mortality events, such as epidemics, into established frameworks like the Lee–Carter model or their Bayesian equivalents. Such improvements enable a more nuanced and realistic understanding of the mortality dynamics, with significant implications for forecasting, pension planning, insurance pricing, and public health policy.

While traditional models provide valuable benchmarks for population forecasting and financial planning, they may not capture the complexity and heterogeneity of mortality risk at the individual level. This limitation motivates the move toward more granular, health-informed approaches, which aim to complement demographic methods with richer data and more personalised risk assessment.

2.3.2 Australian Mortality Modelling

Although Australia has not yet incorporated comprehensive health data into national mortality prediction models, numerous studies have explored disease-specific mortality and hospital mortality [Australian Bureau of Statistics, 2023; Australian Institute of Health and Welfare, 2014, 2018; Walsh et al., 2014]. Hospital mortality has also been studied in relation to specific diseases. However, due to the absence of detailed hospitalisation data and the inability to identify individual hospital admis-

sions within available datasets, by comparison, this area has not been researched in depth.

Although this type of research has not yet been undertaken within an Australian context, similar studies in other countries have explored the integration of available health data into mortality modelling. In the United Kingdom, for example, the Clinical Practice Research Datalink (CPRD) [Clinical Practice Research Datalink, 2024] is the most prominent platform for researchers to access health records along with relevant mortality outcomes. Studies using CPRD have explored associations between chronic conditions, treatment protocols, and mortality rates, such as the study on associations between multiple long-term conditions and mortality across diverse ethnic groups [Stafford et al., 2022].

Similarly, New Zealand has developed the Integrated Data Infrastructure (IDI) [Statistics New Zealand, 2022], a comprehensive dataset linking health, social, and demographic data for population studies. Using this, research have analysed the impact of different socio-economic factors (including whether individuals identify as Māori or non-Māori), access to healthcare, and medical conditions such as cancer, cerebrovascular disease, diabetes, and influenza [Ministry of Health].

It is important to acknowledge that mortality patterns very across countries [Rakshit and McGough, 2025; Crimmins et al., 2016; Crimmins and Beltrán-Sánchez, 2018], due to a range of demographic, environmental, and behavioural factors. Some of the key differences include:

- **Population composition:** Variations in age structure, migration patterns, and the proportion of indigenous and First Nations peoples contribute to differences in baseline mortality and improvement rates.

- **Health system and access:** Differences in healthcare systems, such as Medicare settings, insurance coverage, general practitioner gate-keeping, rural accessibility, and screening programs influence survival outcomes and case–fatality rates.

- **Risk factor prevalence:** The distribution of risk factors, such as smoking, obesity, alcohol consumption, heat exposure, and infection histories is country-specific, pushing age–cause profiles in different directions.

- **Climate and geography:** Heat, humidity, bushfire smoke, and remoteness affect morbidity and mortality in ways that may not be comparable to other countries.

- **Health behaviours and adherence:** Variations in primary care engagement, medication adherence, and lifestyle behaviours influence post-diagnosis mortality and overall health outcomes.

In summary, studies conducted elsewhere must be carefully evaluated for their applicability to the Australian context before being adopted locally.

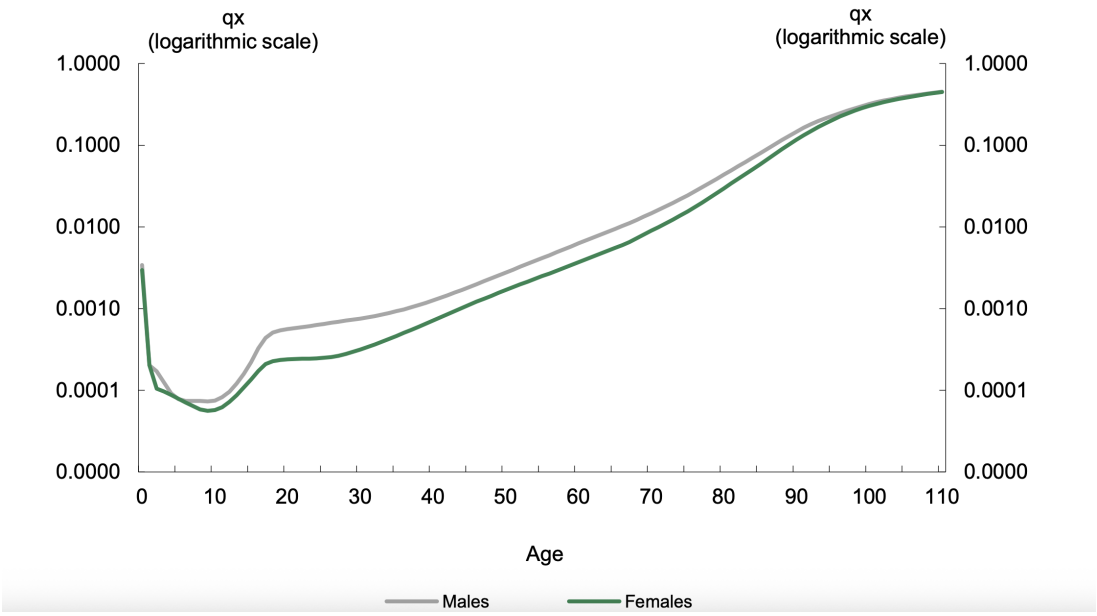


Figure 2.5: Mortality for all ages using Australian Life Tables Australian Government Actuary [2024], shown on a log scale.

Australian mortality research can be divided into three main areas: general population trends, retiree-specific outcomes, and mortality studies focused on particular health conditions.

2.3.2.1 Population-wide Mortality

The ALT is released every five years (the latest one, Australian Government Actuary [2024]), and are used by insurers and the superannuation industry to estimate lifespan and price financial products. These tables summarise how mortality changes across ages and between genders within the Australian population, forming the foundation for many actuarial and demographic analyses

The mortality rate for Australians (Figure 2.5) exhibits the characteristic "U-shaped" pattern with high death rates in infancy, minimal throughout childhood, and a steady increase in adulthood. Male mortality remains consistently higher than female mortality. As this study focuses on retirement mortality (individuals aged 60 and above), no further analysis has been conducted outside this age range. Technically, the Australian Bureau of Statistics (ABS) computes central death rates m_x for single-year age intervals, applies cubic spline graduation with Makeham extrapolation, and then derives probabilities of death q_x, survivorship l_x, and remaining life expectancy e_x using standard life-table formulae [Australian Government Actuary, 2024].

The PLIDA (Australia's national linked administrative dataset) which consists of de-identified micro-data, has been used to study and build subgroup life tables. These tables improve standard population life tables by adding granularity, based on

features identified as significant determinants of mortality.

Using PLIDA, the Australian Government Actuary (AGA) conducted a series of analyses examining the effects of country of birth and socio-economic status on mortality. By linking individual death records to mid-year population exposures, the AGA was able to estimate age and gender specific mortality rates for each subgroup. The findings indicated meaningful differences in life expectancy across Australia: individuals in some jurisdictions lived up to two years longer than others, migrants experienced an average advantage of about 1.5 years, and people from higher socio-economic backgrounds consistently lived longer [Australian Government Actuary, 2021d]. This research series comprises three papers: Paper 1 develops state and territory life tables [Australian Government Actuary, 2021c], Paper 2 explores birthplace-specific life tables [Australian Government Actuary, 2021a], and Paper 3 examines life tables by socio-economic advantage and disadvantage [Australian Government Actuary, 2021b].

Building on this work, Huang et al. [2023] extends the subgroup life–table approach by demonstrating how socio-economic variables can be used to improve mortality predictions for retirees.

2.3.2.2 Retirement Mortality

The Actuaries Institute engaged Rice Warner to investigate retiree mortality in Australia, using UK Continuous Mortality Investigation (CMI) data as a benchmark due to the limited availability of domestic annuitant experience [Actuaries Institute, 2018]. UK annuitant mortality ratios (across both voluntary and compulsory purchase groups) were compared with UK population mortality, and then translated to Australian context by aligning UK–Australia population mortality differentials. From this process, indicative Australian annuitant tables were derived, revealing strong selection effects, clear socio-economic gradients, and notable pension-size impacts. Forward-looking improvement factors were subsequently applied to produce projected mortality rates. The report highlights data-scarcity limitations and advises caution when using these tables for product design and pricing.

Additionally, the study conducted on socio-economic variable impacts to mortality has a specific focus on retirees [Huang et al., 2023]. This is further detailed in Section 2.3.4.

2.3.2.3 Health-Related Mortality

The Australian Institute of Health and Welfare's Deaths in Australia report provides comprehensive mortality data up to 2023 [Australian Institute of Health and Welfare, 2025]. It notes that, between 1907 and 2023, the crude death rate declined by 37%, while the age-standardised rate fell by 75%. According to the ABS's Causes of Death, Australia, 2023, there were 183,131 registered deaths in 2023 with ischaemic heart disease and dementia as the leading causes [Australian Bureau of Statistics, 2024].

However, these reports primarily examine the distribution of causes of death and

do not directly estimate disease-specific mortality rates or incorporate health-based variables into mortality modelling. Consequently, while they identify the leading conditions contributing to mortality, they do not quantify how the presence of a specific disease alters an individual's risk of death compared with those without the condition.

2.3.3 Health Variables in Mortality Prediction

Health-related variables, such as the presence of chronic diseases, medication use, and medical procedures undertaken, provide a more comprehensive understanding of an individual's health status. Incorporating these factors can significantly enhance the accuracy of mortality predictions compared to traditional models that primarily focus on age and gender. However, limited research has been conducted on the overall impact of health status on mortality. Instead, most studies in this field have concentrated on the following areas:

1. **Disease-Specific Mortality**: Research in this area focuses on understanding how specific diseases influence an individual's risk of mortality.
2. **Health-Data Driven Risk Identification**: This research aims to facilitate early intervention and improve cost predictions for healthcare providers.

2.3.3.1 Disease-Specific Mortality

Before identifying the data limitations relevant to the current research, a review of the leading causes of death identified was conducted based on findings from the Australian Institute of Health and Welfare [2021]. The principal causes of death can be summarised by age group as follows:

- Ages 65-74: Lung cancer is the leading cause of death, followed by coronary heart disease, dementia (including Alzheimer's disease), cerebrovascular disease, and chronic obstructive pulmonary disease (COPD).
- Ages 75-84: Coronary heart disease is the leading cause, follwed by dementia (including Alzheimer's disease), lung cancer, cerebrovascular disease, and COPD.
- Ages 85-94: Dementia (including Alzheimer's disease) is the primary cause of death, followed by coronary heart disease and cerebrovascular disease.
- Ages 95+: Dementia (including Alzheimer's disease) remains the leading cause, followed by coronary heart disease, cerebrovascular disease, heart failure and ill-defined heart disease.

Although lung cancer remains one of the leading causes of death in Australia [Australian Bureau of Statistics, 2024], data limitations prevented its identification

within the available health datasets. Specifically, the number of chemotherapy procedures was missing for all individuals, making it impossible to identify cancer patients. Similarly, other diseases listed above could not be isolated due to incomplete or inconsistent information in the Medicare Benefits Schedule (MBS) and Pharmaceutical Benefits Scheme (PBS) datasets.

Although the relevant literature on disease-specific mortality was reviewed, the findings could not be directly applied due to the absence of diagnosis information in the available health datasets. Furthermore, the lack of medical expertise required to interpret and map the procedural and prescription codes to specific diagnoses limited the ability to extract meaningful clinical indicators for the above causes of death. To support future research, the relevant literature has been summarised in Appendix C.

As will be discussed in Chapter 3, one distinct trend identified is the impact of pain medication on mortality. Between 2011 and 2016, the data show a significantly high growth in the use of pain-related prescriptions. This pattern aligns with broader public health evidence documenting a rise in opioid prescribing and misuse in Australia during the early 2010's, which contributed to rising rates of opioid dependence and overdose-related deaths [Australian Institute of Health and Welfare (AIHW), 2021].

2.3.3.2 Health-Data Driven Risk Identification

Health-data driven risk identification has gained significant interest in recent years, particularly as more health data becomes accessible and analytical tools evolve. This section reviews key studies in this area, focusing on how electronic health records (EHRs), machine learning, and data-driven approaches are used to identify risk factors for a range of health conditions, including chronic diseases, patient satisfaction, and ageing-related conditions.

One of the foundational studies in this domain is Sun et al. [2012], which combined knowledge-driven and data-driven approaches to identify risk factors using EHRs. Their work emphasises the potential of EHR data to develop predictive models for identifying individuals at risk of developing chronic diseases, thereby enabling early intervention and more targeted treatment strategies. For example, in their study, they found that smoking status, hypertension, and cholesterol levels were well-established knowledge-driven risk factors for cardiovascular disease. These were combined with newly identified data-driven risk factors derived from EHR analysis, including socio-economic variables (such as income and education level) and behavioural patterns (such as physical inactivity and high-fat diets). The integration of these augmented risk factors enabled their model to better predict individuals at risk for developing chronic conditions such as heart disease and diabetes.

Further advancing this concept, Yu et al. [2022] applied a data-driven approach to improve the risk assessment process for medical complications. By integrating various patient data, including clinical history and socio-demographic characteristics, the study demonstrated how predictive models can be refined to more accurately

assess the risk of medical complications, thereby improving patient outcomes and overall healthcare quality.

In terms of ageing, Kuan et al. [2021] explored the identification of ageing-related diseases using large-scale EHR databases. They detected early markers for conditions such as dementia, cardiovascular disease, and diabetes. Their study highlights the power of health data analytics in uncovering early-stage diseases, and enabling timely interventions and personalised care for elderly patients. They also provide a detailed analysis of the prevalence of various diseases and their progression with age.

Additionally they have also analysed the median age at diagnosis across various disease categories. Cardiovascular diseases and cancers showed the highest median ages at diagnosis, around 68-70 years, while conditions like psychiatric disorders and skin diseases are diagnosed much earlier, around 38-50 years. Infections and endocrine diseases were diagnosed at a median age of approximately 56-57 years. This information is highly relevant for improving mortality prediction models, as diseases diagnosed earlier in life (e.g., psychiatric or skin conditions) may have different mortality risks compared to those diagnosed later in life (e.g., cardiovascular diseases). Incorporating such distinctions could support the development of more accurate mortality predictions for ageing populations, particularly retirees.

Finally, in the specific case of chronic kidney disease (CKD), Chiu et al. [2021] applied machine learning algorithms to identify key risk indicators using health data. Their study showed that machine learning models could accurately predict individuals at high risk of developing CKD by incorporating variables such as kidney function markers, demographic characteristics, and pre-existing conditions. The findings highlight the potential for health-data driven models to predict and manage chronic diseases more effectively.

2.3.3.3 Health Index Creation

Another potential approach to mortality modelling using health data is to incorporate health-related variables into a composite health index. which can then be integrated into existing frameworks such as that of Huang et al. [2023]. Hansen et al. [2025] adopted this approach using Danish register data to construct a health index summarising an individual's overall health status. Their method involved selecting the most predictive health-related variables from medical and demographic records and applying machine learning models to combine these into a continuous index representing underlying health risk. The resulting index was validated against mortality outcomes and shown to improve predictive accuracy when included in forecasting models.

However, this approach has several limitations. The construction of the index relies heavily on machine learning techniques, which can make the index less interpretable and difficult to link directly to biological or clinical mechanisms. Furthermore, the assumption that a single index can adequately represent complex health dynamics may oversimplify the diverse ways in which diseases and condi-

tions progress over time.

2.3.4 Socio-Economic Factors and Mortality Prediction

In addition to health variables, socio-economic factors such as income, education, and occupation play a critical role in determining life expectancy. The Australian Retirement Mortality and Longevity study by Huang, Hui, and Villegas [Huang et al., 2023] found that retirees from lower socio-economic backgrounds have significantly higher mortality rates than those from more advantaged groups. This disparity highlights the importance of incorporating socio-economic information into mortality prediction models. Previous studies have shown that the inclusion of socio-economic status improves the accuracy of mortality forecasts, particularly in ageing populations.

The study in Huang et al. [2023] has examined socio-economic mortality differentials in Australia, focusing on the post-retirement population, a critical area for developing more equitable retirement income products.

Key Findings:

- **Socio-economic mortality differentials:** The study identifies significant mortality differences linked to socio-economic factors. For example, the gap in life expectancy between the most disadvantaged and most advantaged males is approximately 11.5 years, and 9.1 years for females.
- **Decreasing gap with age:** These mortality differences tend to narrow with increasing age, becoming negligible by age 100.
- **Longevity and annuity impact:** Longevity differences across socio-economic groups translate directly into variation in annuity income. For example, individuals with the shortest life expectancy (typically those from disadvantaged socio-economic groups) could receive up to 28% more in annuity income than those with the longest life expectancy.
- **Income and education:** Income and education are strong predictors of life expectancy. Individuals with higher incomes and better educational attainment tend to live longer and have better access to higher-quality healthcare. Crimmins et al. [2010] further supports this finding, showing that mortality disparities across socio-economic groups have been widening, particularly among retirees.

The study emphasise the importance of considering socio-economic mortality differentials to ensure that retirement income products are both equitable and sustainable. Their findings provide valuable insights for the superannuation and retirement industries in Australia to support the development of fairer policies and product designs that reflect the diverse longevity risks across socio-economic groups.

The modelling approach adopted by Huang et al. [2023] uses a Hermite spline model, which assumes that individual socio-economic variables remain constant over

time. This assumption is generally reasonable for factors such as education or income level, which are relatively stable throughout life, but less appropriate for health-related variables, which can change significantly, particularly in later life.

While the study offers a broad understanding of how socio-economic variables influence mortality at the population level, their effects may vary considerably across different diseases. Notably, just as the impact of socio-economic factors tends to diminish with age, it may also decrease with increasing disease severity. Therefore, the model proposed by Huang et al. [2023] may not be universally applicable and could require further adjustments to accommodate condition-specific variations identified in Mullany and et al. [2021].

2.3.5 Dynamic Mortality Modelling

Traditional actuarial approaches to mortality modelling have long relied on life tables and parametric models that project future mortality based on historical age-specific rates. Mortality rates tend to remain relatively stable over short time frames, however over longer periods, they can vary substantially due to medical advancements and changes in exposure to different diseases. These long-term shifts are incorporated into the mortality tables through the use of mortality adjustment factors [Australian Government Actuary, 2024].

As mentioned earlier, one of the most influential developments in this field is the Lee–Carter model [Lee and Carter, 1992a], which represents log-mortality rates as a combination of age-specific effects and a time-varying mortality index. This framework laid the groundwork for incorporating mortality improvements directly into the model structure and remains widely used in both industry and academia. However, while it captures the dynamics of population-level mortality trends, it does not account for the evolving, individual-level influence of health-related factors.

Beyond these parametric approaches, multi-state and Markov-based models have emerged as powerful tools for representing the dynamic evolution of health status and mortality risk.

Another form of dynamic mortality modelling focuses not on changes in individual health status over time, but on allowing model parameters to evolve themselves. Aliverti et al. [2022] propose a dynamic Bayesian mixture model for mortality that represents the age-at-death distribution using multiple components, including adjustments for infant deaths and skewed distributions for adult and old-age mortality. This approach captures changes in mortality patterns across both time and populations while maintaining interpretability. By leveraging cross-country information and explicitly modelling temporal dynamics, the model enhances forecasting accuracy relative to traditional approaches and provides a flexible, data-driven framework for analysing heterogeneous mortality patterns. Steinsaltz and Evans [1969] further advanced mortality modelling by allowing for different initial population distributions, thereby improving the representation of heterogeneity in survival patterns.

2.3.5.1 Markov Chains for Mortality Modelling

Markov chain models provide a powerful framework for capturing the dynamic nature of mortality and morbidity processes by representing life as a sequence of transitions between discrete health states. Unlike traditional life tables, which assume static age-specific mortality rates, Markov models explicitly model the probabilities of transitioning between states over time, allowing for a richer representation of ageing, disease progression, and mortality risk.

A standard approach is to define a finite set of states representing the possible health conditions of an individual throughout their lifetime. A common structure includes three primary states: *Active (healthy)*, *Ill*, and *Death*, with the last acting as an absorbing state, meaning that once entered, no further transitions occur [Milan et al., 2021]. Some extensions also incorporate disability as another state [Dudel and Myrskylä, 2020].

When all four states (active, ill, death, and disabled) are incorporated, transitions between these states occur over either discrete or continuous time intervals, governed respectively by transition probability matrices (for discrete-time models) or intensity matrices (for continuous-time models). Let S_t denote the state of an individual at time t. For a discrete-time Markov model with state space $\{A, D, I, \emptyset\}$, where $A, D, I, \emptyset$ denote Active, Disabled, Ill, and absorbing Death states respectively, the transition probability matrix is defined as:

$$P = \begin{bmatrix} p_{AA} & p_{AD} & p_{AI} & p_{A\emptyset} \\ p_{DA} & p_{DD} & p_{DI} & p_{D\emptyset} \\ p_{IA} & p_{ID} & p_{II} & p_{I\emptyset} \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{2.1}$$

where $p_{ij} = \Pr(S_{t+1} = j \mid S_t = i)$ represents the probability of transitioning from state i at time t to state j at time $t + 1$. The Markov property implies that transitions depend only on the current state (and possibly on time or age), rather than on the individual's prior history. In continuous time, the dynamics are represented by a transition intensity matrix $\mathbf{Q}(t)$, and state occupancy probabilities evolve according to the Kolmogorov forward equations [Gallager, 2011].

This modelling approach improves upon the ALT by explicitly distinguishing transitions between health states, such as from Active to Ill or Disabled, and then to Death, allowing for a more realistic representation of mortality processes and more precise estimation of mortality improvements over time.

Further advancements in the application of Markov chains to mortality modelling include the work of Milinovic et al. [2022]. They propose an extension to the three-state model to have "N" different "alive" states, each having a different rate of mortality as shown in Figure 2.6. In this formulation, mortality depends explicitly on the individual's current state. However, the model assumes a single direction of transitions between states, making it unsuitable for health-based mortality prediction, where medical conditions may recur and bidirectional or recurrent state transitions are more realistic.

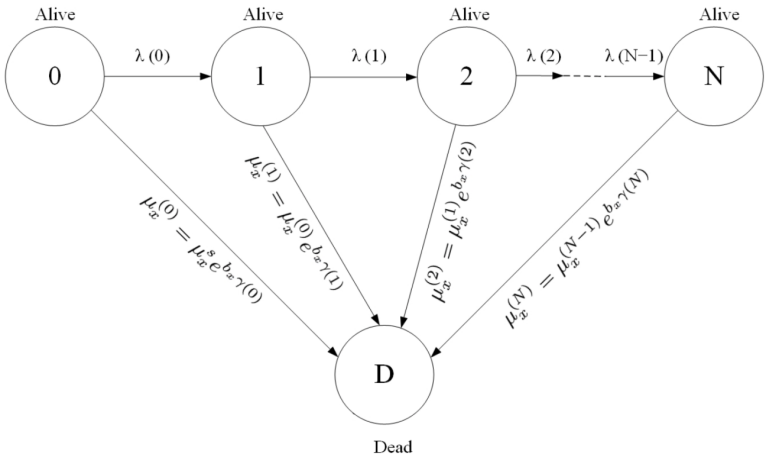


Figure 2.6: An extended Markov mortality model with multiple "alive" states, each with a distinct mortality rate, illustrating how transitions between health states influence overall mortality risk. [Milinovic et al., 2022]

Several studies have applied Markov chains to examine the impact of specific diseases on mortality. For instance, Currie et al. [2023] modelled the impact of pancreatic cancer on mortality using a Markov framework. Their model included multiple states representing the various stages of pancreatic cancer, along with healthy and death states to capture disease progression and its influence on mortality outcomes. However, no research to date has explored the use of general health-based variables on a comparable scale to assess their broader impact on mortality.

The advantages and limitations of disease-based Markov models have been analysed in Siebert et al. [2009]. Their study notes that key assumptions of such models, such as the Markov property, the exclusion of disease duration, and the oversimplification of complex clinical pathways, can reduce predictive accuracy. They also highlight that recovery rates from a disease depend on numerous individual factors and are therefore unlikely to be uniform across patients.

2.4 Key Takeaways

This research aims to incorporate health-related variables into mortality modelling, a task that can be approached in several ways. Section 2.2 reviewed the mortality calculation methods currently in use, highlighting that the age–gender-based models predominantly applied in Australia treat populations as homogeneous and fail to capture the considerable variation in mortality risk arising from health status, disease progression, and lifestyle factors. As a result, such models often lead to significant cross-subsidies in financial products, where healthier individuals subsidise those with higher health risks, and vice versa.

Within the scope of this literature review, the proposed study is novel, representing the first attempt to integrate health information into mortality modelling for

retirees in an Australian context. While previous studies have examined disease-specific mortality, socio-economic differentials, and subgroup life tables, none have linked individual-level health data to mortality outcomes at scale. International research demonstrates the value of incorporating clinical, behavioural, and socio-demographic variables, but differences in healthcare systems, population characteristics, and data availability limit their direct applicability to Australia. For instance, the prescription codes in the PLIDA dataset are uniquely Australian and highly aggregated, making it difficult to map medications to specific disease states or replicate overseas methodologies.

While modelling techniques such as Hermite splines and health indices have provided valuable insights, they often rely on static assumptions or oversimplify the complex, dynamic nature of health trajectories in later life. Multi-state Markov models, which explicitly capture transitions between health states and death, offer a more realistic framework but remain underexplored in the context of broad, health-based mortality prediction. Accordingly, this research explores multiple methods for incorporating health-related information into mortality models and proposes a non-standard dynamic Markov chain that represents both the health states of individuals and the transitions between them. For comparison, it also evaluates the performance of a generalised linear model with cubic splines, designed to replicate a simplified version of the Hermite spline model developed for socio-economic variables by Huang et al. [2023].

To our knowledge, there is currently no published Australian framework that simultaneously (i) models late-life health transitions and (ii) produces annuity-relevant outputs with the level of transparency required for actuarial pricing. Subgroup life tables quantify socio-economic differentials but remain static, while disease-specific Markov models capture dynamic processes yet typically rely on clinical registries not nationally available for retirees. The proposed approach bridges this gap by retaining actuarial interpretability, through life-table outputs such as (q_x) and (e_{65}), while embedding a health state process that reflects individual heterogeneity relevant to retirement products. Compared with overseas studies, this framework emphasises domestic calibration, careful handling of Australian pharmaceutical coding, and a model which can easily perform periodic parameter updates.

To date, no published framework has fully leveraged Australian administrative health data for retiree mortality forecasting. Although international studies have demonstrated feasibility [Steyerberg and Bastiansen, 2019; Alvarez-Garcia et al., 2024], these efforts often overlook key challenges identified in this research:

- The absence of diagnostic information in the MBS and PBS datasets used for modelling.
- Potential misclassification when mapping prescriptions to medical conditions, an issue unaddressed in prior work.
- Contextual differences between healthcare systems that limit the transferability of international models.

- Reliance on historical data without incorporating post-2016 trends.

Nonetheless, the proposed modelling approach remains valid and can be retrained and re-calibrated as more recent data become available. With continued advancements in data access and analytical capability, increasingly personalised and precise mortality predictions will become feasible, improving financial planning for retirees and supporting evidence-based actuarial decision-making.

The broader body of literature primarily examines how other countries have integrated health data and lifestyle factors into mortality prediction models, alongside Australian studies investigating similar determinants. Given that a major goal of this research is to ensure interpretability and transparency, critical for pricing applications in retirement and insurance products, relevant literature on interpretable machine learning methods, such as constrained neural networks, has also been considered.

2.5 Limitations of this Literature Review

The scope and evidence base of this review are subject to several constraints:

1. Breadth of disease-specific research: The body of literature on disease-specific mortality is extensive. Due to the time limitations of this project, it was not feasible to reference all relevant studies in this area.

2. Geographical applicability: Most health data–based mortality research has been conducted outside Australia or on a global scale. As mortality patterns are highly location-specific, the relevance of these findings within an Australian context must be carefully assessed to ensure the model's applicability.

3. Evolving impact of medication: The effect of medications on mortality, including potential drug–drug interactions, represents a key aspect of this research. However, as pharmaceuticals evolve over time, ongoing research and periodic model updates will be necessary to account for medical advancements and changes in treatment patterns.

4. Temporal limitations of data: Health research is continuously evolving, which influences how diseases affect mortality. The data used in this study span 2011–2016; therefore, updating the model with more recent data and literature will be essential to reflect current clinical practices and population health trends.

5. Prescription code aggregation: The prescription codes in the dataset are Australia-specific and highly aggregated, making it impossible to reliably map medications to specific diagnoses. Consequently, disease impacts could not be inferred from prescription data alone, and no detailed literature was reviewed in this area.

2.6 Gap Analysis

Despite extensive research into mortality modelling, several critical gaps remain that limit the accuracy, fairness, and applicability of existing approaches, particularly in the context of retirement planning and annuity pricing.

First, traditional life table methods and widely used demographic models continue to rely almost exclusively on age and gender. While these variables capture broad mortality patterns, they fail to account for the substantial heterogeneity introduced by health status, chronic conditions, and treatment history. This omission leads to biased survival estimates and significant cross-subsidies in financial products, where individuals with higher mortality risk are disadvantaged relative to those in better health when purchasing lifetime income products. As retirement income products increasingly shift from pooled defined-benefit structures to individualised defined-contribution systems, the need for personalised, health-informed mortality models has become more pressing.

Second, although disease-specific mortality models and socio-economic analyses have been conducted, they remain fragmented and narrowly focused. Most disease-based studies examine single conditions or limited subpopulations, restricting their generalisability and preventing a holistic understanding of how multiple coexisting health factors interact to influence mortality. Likewise, socio-economic models, while informative, typically assume static characteristics over time and fail to capture the dynamic evolution of health status as individuals age.

Third, international efforts to integrate health data into mortality modelling cannot be directly applied in Australia due to differences in healthcare systems, data infrastructure, and coding standards. In particular, the prescription codes in Australian administrative datasets are highly aggregated and unique to the local system, making it difficult to map them to specific diseases or treatment pathways. This poses a major challenge for incorporating clinical variables into predictive models without extensive data engineering or external linkage.

Finally, while multi-state and Markov-based models offer a powerful framework for modelling transitions between health states and death, their application has been largely limited to disease-specific models, small-scale clinical studies, or approaches with restrictive assumptions such as unidirectional transitions. To date, no study has combined these dynamic modelling techniques with broad, population-level health data to estimate mortality risk for retirees. Addressing this gap presents an opportunity to develop more accurate, equitable, and policy-relevant mortality predictions that enhance individual retirement planning, product design, and regulatory decision-making.

2.7 Summary

This chapter established the motivation for developing health-informed, dynamic mortality models within Australia's DC-oriented retirement system. It reviewed the evolution from traditional parametric life-table methods to multi-state frameworks,

highlighting how conventional age–gender-based models treat populations as homogeneous and therefore fail to capture the heterogeneity introduced by health status, comorbidities, and lifestyle factors. This limitation contributes to cross-subsidies in retirement products, where healthier individuals effectively subsidise those with poorer health.

The chapter also traced the historical progression of mortality modelling, from classic parametric approaches such as Gompertz [1825] and Lee and Carter [1992b]) to more flexible models, including Alvarez-Garcia et al. [2024]), and finally to multi-state and Markov formulations that explicitly represent transitions between health states Section 2.3.5.1). It further examined how international evidence, while valuable, cannot be directly transferred to the Australian context due to differences in healthcare systems, population structure, and data availability, particularly in relation to aggregated prescription coding.

Building on this, the gap analysis (Section 2.6) identifies four main shortfalls motivating this research:

1. The limited integration of individual health information in Australian mortality studies,
2. The fragmented nature of disease-specific research, which fails to capture **comorbidity**,
3. The low transferability of overseas methods due to **Australia-specific data** and system differences, and
4. The underuse of **scalable, dynamic multi-state models** for retirees.

Together, these gaps underscore the need for a credible, interpretable, and domestically validated framework that embeds health dynamics within mortality modelling while retaining actuarial transparency. The next chapter (Chapter 3) presents the data assets and exploratory analyses that inform the model design choices developed throughout this thesis.

Chapter 3

Exploratory Data Analysis

This chapter presents an overview of the dataset and describes the processes used to prepare it for subsequent modelling. We first introduce the privacy constraints governing data exports (Section 3.1) and explain the dataset naming conventions (Section 3.2). Section 3.3 summarises the scale, completeness, and reliability of the mortality datasets, highlighting their national coverage and suitability for population-level modelling. Preprocessing steps are outlined in Section 3.4, including data filtering, handling of missing or low-variance variables, and then the creation of interpretable clinical indicators are explained in Section 3.5. Section 3.6 goes over the limitations imposed on the model due to the data issues and constrains. Finally, the exploratory analysis in Section 3.7 examines demographic, clinical, and prescription variables to identify factors associated with mortality risk.

Collectively, these steps improve the quality of the dataset, uncover meaningful differences in mortality between subgroups, and inform the modelling approaches introduced later in Section 4.1.

3.1 Privacy Constrains

All outputs of this research abides with the ethics statement Section 1.4. As this research uses sensitive individual level data, any outputs released from the ABS secure environment (DataLab) [Australian Bureau of Statistics, 2021] must comply with strict disclosure controls. Key constrains are:

- **Minimum cell counts:** mortality curves may be released only where both the number of deaths and the exposure are at least 10, to prevent re-identification.

- **Secure execution:** All modelling and analysis were conducted within the ABS DataLab, with only aggregate, disclosure-checked results permitted for export. The research was therefore carried out within the secure environment's computational limits, particularly restricted memory and processing capacity. These constraints affected both the scope and structure of the analytical workflow, making it necessary to adapt the data and apply optimisation techniques to support efficient processing and model construction.

3.2 MBS-PBS Variable Naming Convention

The following convention is used in the MBS and PBS datasets for variable names:

- `_num` denotes the number of a certain type of service provided.
- `_scripts` denotes the number of prescriptions (scripts) dispensed under the Pharmaceutical Benefits Scheme (PBS).
- `_qty` denotes the total quantity of units supplied for a given medication.

All of these variables denote cumulative values for the reporting period.
In the PBS data, variables with the suffix `_scripts` refer specifically to the number of prescription items dispensed, rather than textual content. Each medication is coded according to the Anatomical Therapeutic Chemical (ATC) classification system. The alphanumeric codes shown in the results (e.g. A10, N03) refer to ATC categories, where each prefix represents a therapeutic class, and corresponding variables capture either the number of scripts dispensed or the quantity supplied within that class.

3.3 Summary of Mortality Data Quality and Coverage

The mortality datasets used in this study provide national-scale coverage, capturing the Australian population aged 55 and above between 2011 and 2016. Across this period, the year-grouped dataset comprises six annual cohorts, encompassing a total exposure of approximately **33.3 million** entries and **576,176** recorded deaths. Exposure and death counts remain stable across years, ensuring consistency for longitudinal analysis.

The gender-grouped dataset includes three demographic categories with total exposures of **10.4 million**, **11.5 million**, and **0.02 million**, corresponding to **285,008**, **290,418**, and **706** deaths, respectively. The first two represent the principal male and female cohorts, while the smaller third group corresponds to records with incomplete or unclassified gender information.

Together, these datasets provide a robust and comprehensive representation of mortality experience across both temporal and demographic dimensions. Their national coverage, stable exposure patterns, and balanced death counts enhance data reliability and support the calibration of transition-based models for population-level longevity analysis. The distribution of exposure and deaths across the six annual cohorts is summarised in Table 3.1, demonstrating consistent coverage and data completeness across the study period.

3.4 Pre-Processing

The current dataset consists of the MBS and PBS data for retired Australians over the years 2011-2016 (inclusive). This was extremely large, and thus modifications were made to the dataset to make it more accessible and aid in the analysis and model

Table 3.1: Total exposure and deaths by year in the mortality dataset.

Year	Exposure	Deaths
2011	5,804,826	88,947
2012	5,707,904	93,958
2013	5,609,753	94,841
2014	5,506,245	98,829
2015	5,392,733	99,973
2016	5,282,410	99,628

building. Although the combined raw data was around 8.2 GB, once loaded into memory it expanded several times in size, often requiring more than 40 GB of RAM to merge due to inefficient data structures, and the overhead of in-memory operations. These factors made direct processing impractical even on a 64 GB machine, so the dataset was streamlined to reduce memory usage and improve computational efficiency.

The modifications are summarised below:

- The dates of diagnosis were reduced to the year only, as the additional detail provided by the month and day offered minimal analytical value and was excluded to optimise memory usage.
- There were less than 100 entries in the 2011 dataset that had a year of death prior to 2011, this is likely due to delays in recording. As those entries spread over a range of 100 years, these entries provide no significant input to the modelling process, and thus have been omitted from the study.
- The MBS dataset has no entries for the given variables: both `_num` and `_date` `dialysis` (Dialysis services) , `chemo_proc` (Chemotherapy procedures), `rad_oncology` (Radiation oncology procedures), `neuro_surg` (Neurosurgery procedures), and `spinal_surg` (Spinal surgery procedures). Thus, these columns were removed.

  Since these variables maybe influential for retiree mortality, incorporating them when more entries are available in future releases will improve the mortality calculation.
- Since this analysis is on those who have retired, the entries were filtered to be for those who are 55 and above.
- Zero variance columns were excluded. This included variables with just NA entries and just zero. This resulted in 76 variables being omitted, including 38 script variables and 38 quantity variables. These can be summarised as;
  - **Anti-diabetics (A10)**; non-insulin glucose-lowering agents. Both quantity and script variables for `A10BA`, `A10BB`, `A10BH`, `A10BJ`, `A10BK`, `A10BG`, `A10BF`, `A10BD`

  - **Anti-epileptics (N03)**. Both quantity and script variables for N03AE, N03AF, N03AG.
  - **Psycholeptics / Antipsychotics, Anxiolytics, Hypnotics (N05)**. Both quantity and script variables for N05AA, N05AD, N05AH, N05AL, N05AX, N05BA, N05BE, N05CD, N05CF.
  - **Psychoanaleptics / Antidepressants & Anti-dementia (N06)**. Both quantity and script variables for N06AA, N06AB, N06DA, N06DX.
  - **Respiratory System; Asthma/COPD Therapies (R03)**. Both quantity and script variables for R03AC, R03BB, R03BA, R03AK, R03AL, R03DC, R03DA.
  - **ENT / Nasal Preparations (R01)**. Both quantity and script variables for R01AD, R01AX.
  - **Addiction Therapies (N07B)**. Both quantity and script variables for N07BB, N07BC. Nicotine dependence N07BA appeared only as quantities in your list.)

  Due to limited data availability for these variables, they were not included. If later datasets provide adequate detail, incorporating these will aid in further improvement of the model. Note that, these consist of the most common causes of death mentioned under Section 2.3.3.1.

- If the number of observations for the data for a given variable was less than 100, these have also been omitted from the study. Though this can potentially be significant indicators of mortality, the relevant mortality curve will not be able to meet the requirements set by DataLab with fewer than 100 entries.

After performing the listed modifications, the size of the dataset had reduced enough to fit within RAM. The resulting variables are summarised under Table 3.2.

Table 3.2: Variables in the dataset, grouped by category after the pre-processing steps were completed. All these variables are values over the reporting period and not aggregated over the lifespan.

Variable	Description	Source
Identifiers and Demographics		
mort_id	Unique de-identified patient identifier	
gender_y	Gender (from Medicare/Census linkage)	Demographic
year,year_of_death, year_of_birth	year of data, death and birth	Demographic
age	current year - year of birth	Derived
Service Counts (MBS)		

Continued on next page

Table 3.2 (continued)

Variable	Description	Source
total_services	Total number of Medicare services used within that year	MBS
Clinical Procedures (MBS)		
mental_health_num	Number of mental health services	MBS
diab_mellitus_num	Diabetes-related services	MBS
pain_med_num	Pain management services	MBS
addict_med_num	Addiction treatment services	MBS
amp_hip_num	Hip amputation or replacement	MBS
hind_qtr_num	Hindquarter amputation	MBS
laryngectomy_num	Laryngectomy (removal of larynx)	MBS
cor_bypass_num	Coronary bypass surgery	MBS
gen_surgery_num	General surgery procedures	MBS
pancreas_proc_num	Pancreatic surgery procedures	MBS
oesoph_surg_num	Oesophageal surgery procedures	MBS
abdom_repair_num	Abdominal repair surgery	MBS
Prescriptions (PBS): scripts-scripts dispensed, qty- number of units supplied		
total_scripts, total_qty	All PBS prescriptions	PBS
C03_scripts, C03_qty	Diuretics	PBS
C07_scripts, C07_qty	Beta-blockers	PBS
C08_scripts, C08_qty	Calcium channel blockers	PBS
C09_scripts, C09_qty	Renin–angiotensin system agents	PBS
C02A_scripts, C02A_qty	Antihypertensives	PBS
C04_scripts, C04_qty	Peripheral vasodilators	PBS
C05A_scripts, C05A_qty	Varicose therapy agents	PBS
C01A_scripts, C01A_qty	Cardiac glycosides	PBS
C01B_scripts, C01B_qty	Antiarrhythmics, class I	PBS
C01C_scripts, C01C_qty	Antiarrhythmics, class III	PBS
C01D_scripts, C01D_qty	Vasodilators for cardiac diseases	PBS
C01E_scripts, C01E_qty	Other cardiac therapy	PBS
C10_scripts, C10_qty	Lipid-modifying agents (statins)	PBS
B01AE_scripts, B01AE_qty	Direct thrombin inhibitors	PBS
B01AF_scripts, B01AF_qty	Factor Xa inhibitors	PBS
B01A_scripts, B01A_qty	Other antithrombotic agents	PBS
A10AD_scripts, A10AD_qty	Insulins	PBS

Continued on next page

Table 3.2 (continued)

Variable	Description	Source
N04_scripts, N04_qty	Anti-Parkinson drugs	PBS
N07B_scripts, N07B_qty	Drugs for addictive disorders	PBS
N07BA_scripts, N07BA_qty	Nicotine-dependence drugs	PBS

Note: The variable age is a simplified approximation calculated as current year - year of birth. It does not consider the exact month or day of birth or death, and therefore may differ slightly from the ALT age or true age at death.

3.5 Derived Clinical Indicators (MBS/PBS)

To make the dataset more interpretable, the raw MBS and PBS variables, which indicate prescriptions and services, were summarised to indicator variables that represent the relevant diagnosis. As the variables of the data are cumulative for the entire reporting period, the indicator variables denote whether or not they had the condition over the past year. We will refer to the initial dataset with raw MBS and PBS variables as the *raw dataset*, and the dataset with summarised indicator variables per disease as the *disease dataset*.

Due to this definition missing counts are treated as zero. For each indicator Section 3.5.1 we set the variable to 1 if *any* relevant count is > 0 during the relevant year, and 0 otherwise. This creates reproducible, memory-light features and avoids instability from sparse counts. This encodes whether the person ever received the service or drug class during the reporting period, which is suitable for mortality modelling at population level. After filtering for variables with at least 100,000 (a minimum of 0.3%) non-missing observations from the original 33 million records, the following variables were retained for analysis.

3.5.1 Indicators Constructed

When _num is used _date is also considered if available.

- **Mental health**: mental_health_num.
- **Diabetes**: diab_mellitus_num or PBS insulin (A10AD_scripts/qty).
- **Pain medication**: pain_med_num.
- **Addiction treatment**: addict_med_num, N07B_scripts/qty, N07BA_scripts/qty.
- **Cardiac therapy proxy (PBS)**: any of C07, C08, C09, C02A, C04, C05A, C01A-E (_scripts or _qty).
- **Lipid disorder therapy**: C10_scripts/qty (statins).
- **Antithrombotic therapy**: B01AE, B01AF, or B01A (_scripts/qty).

- **Parkinson's therapy**: `N04_scripts/qty`.
- **Coronary bypass**: `cor_bypass_num`.

From these indicators, `coronary bypass` did not have a sufficient distribution of observations across age to model the mortality (Figure 3.9), and thus was omitted for some models.

To reduce dimensionality and improve stability, variables with insufficient information were removed. Specifically, a column was retained only if it had at least *100,000 non-zero, non-missing* entries across the full cohort. Very low-frequency procedures were therefore excluded from downstream analysis.

The approach assumes that the absence of a claim implies no exposure in the period and that PBS/MBS classes are reasonable proxies for clinical conditions (for example statins for lipid disorder, `B01A*` for anti-thrombotic therapy). This may under-record conditions treated outside PBS/MBS and should be considered when interpreting results.

3.5.2 Benefits and Limitations

Where possible, variables are expressed as broad disease indicators (or clinically interpretable service groups) rather than raw MBS/PBS ATC codes. This choice supports three practical goals:

- **Interpretability.** Disease-level variables make model outputs and their implications for mortality clearer to non-technical stakeholders.
- **Questionnaire alignment.** Disease framing translates directly into answerable items, improving accuracy and response rates. For example, *"Have you been diagnosed with a cardiac condition in the past five years?"* is more accessible than *"Have you taken any C10-classified medicines in the past five years?"*
- **Privacy and proportionality.** Broader questions reduce the need to disclose specific medicines or dosages, lowering privacy and confidentiality risks while still capturing the risk signal required by the model.

The latter two considerations are particularly important as the results of this thesis could guide annuity providers in developing fair and practical underwriting questionnaires. They ensure that data collection is not only accurate and accessible for respondents but also ethically sound and compliant with privacy standards. By grounding the model's application in these principles, the research supports a responsible translation of technical insights into real-world actuarial practice.

However it is important to note that these indicator variables may be an over simplification of the original raw variables of the MBS and PBS datasets. For instance, the disease dataset considers the presence of a cardiac condition: if they have ever received at least one prescription from beta-blockers (C07), calcium channel blockers (C08), agents acting on the renin-angiotensin system (C09), or related groups.

Though this is a good indicator it is not a perfect proxy for true disease status. Certain drugs in these categories may be prescribed for non-cardiac conditions (e.g. beta-blockers for migraines [Migraine Australia, 2025]), creating false positives. It also misses untreated or undiagnosed cases, producing false negatives, and a single short-term prescription may not reflect chronic disease. In short, this indicator is *"a reasonable proxy for"* to capture exposure to cardiac-related treatment and suitable for broad research use, but it cannot precisely measure the prevalence of cardiac disease without mis-classification.

3.6 Limitations due to Data

Model limitations for individual models have been provided under each model but general limitations of this research are:

1. The MBS and PBS datasets used for this research **do not include diagnosis data**. The data consists of the doctor visit details and the drug prescriptions. One justification for not including the diagnostics for the model is to prevent the model from being impacted by different perspectives of the doctors. Though this is the case, in terms of the modelling this means that the details of the disease specific mortality can not be directly mapped to the data without further study on what prescriptions relate to which drug.

   Though this study attempts to summarise the variables MBS and PBS to disease related variables for the interpretability of the questionnaire, there are limitations of the approach (Section 3.5).

2. The death dataset in PLIDA [Australian Bureau of Statistics, 2025], which contain detailed information about death (including cause of death and time of death), had no ID column and hence was not incorporated in this study. In the absence of cause-of-death codes, we assume the available health variables capture most of the variation in mortality risk that those codes would explain, noting this as a limitation to be revisited if a linkable file becomes available.

3. Due to the sensitive nature of the data, the analysis and modelling were performed within the ABS DataLab with limitations in the accessible packages and the export criteria (Section 3.1).

3.7 Exploratory Data Analysis

The exploratory analysis provides valuable insights into the data, highlighting which variables influence mortality. This section examines the impact of these key variables on mortality and explores the trends that emerge.

3.7.1 Mortality

Mortality (q_x) defined for an age x, is calculated as follows

$$q_x = \frac{\text{Total number of individuals that die between ages } x \text{ and } x+1}{\text{Total number of individuals alive at age } x}. \tag{3.1}$$

This is referred to as the crude mortality as it is purely calculated from observed data.

Note that due to exporting constraints set on the dataset (Section 3.1), only values where both the numerator and the denominator are at least 10 are provided.

3.7.1.1 Mortality Trends Against Non-Health Features

Before investigating how mortality varies by health-related characteristics, it is useful to examine broader demographic patterns. This section explores how mortality evolves with age, calendar year, and gender, providing context for how these non-health factors interact with mortality dynamics.

Mortality Against Age. Mortality rates (q_x) generally increase with age, reflecting the underlying biological ageing process. Figure 3.1 shows this pattern clearly, with mortality rising steeply from mid-life onwards. The slight decline observed at the oldest ages is unlikely to reflect biological reality; rather, it typically arises from small sample sizes (few survivors, leading to noisy estimates), age misreporting, and survivorship selection effects [Crimmins and Beltrán-Sánchez, 2018]. In this analysis, the upper age range has been grouped into a single open-ended category (100+) to reduce instability and improve model performance.

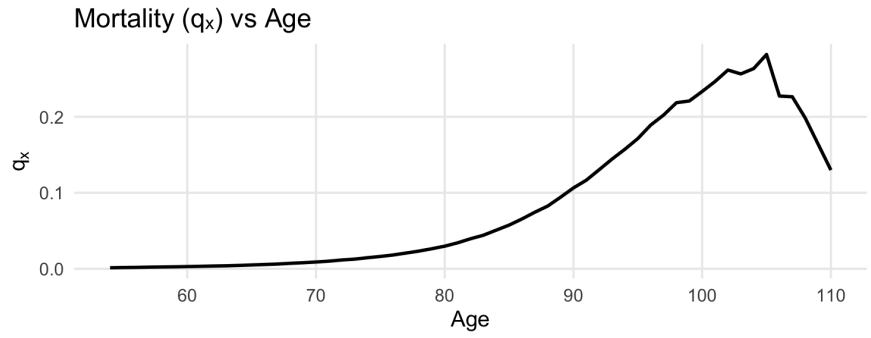


Figure 3.1: Mortality (q_x) against age using the complete dataset (2011–2016). Mortality increases steeply with age, consistent with expected demographic patterns.

Mortality Against Year. Mortality curves were broadly consistent across calendar years, as shown in Figure 3.2. This stability is expected given the relatively short observation period (2011–2016) and is valuable for modelling, as it ensures that short-term variations do not bias mortality estimates. Over longer time horizons, mortality

typically improves due to medical, technological, and public health advances. The ALT account for these long-term trends using *mortality improvement factors*, which adjust projected mortality rates beyond the observed period.

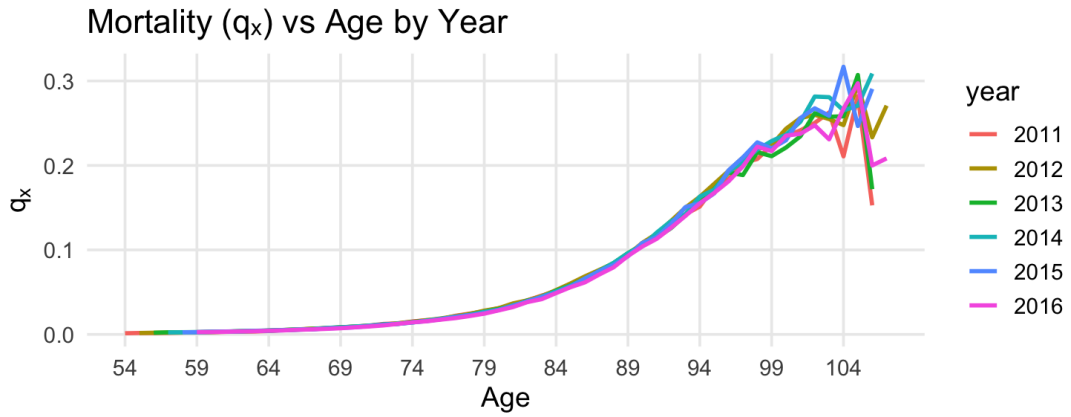


Figure 3.2: Mortality against calendar year (2011–2016). The short-term stability in mortality rates supports consistent model calibration.

Mortality Across Gender. Gender differences in mortality are well-established and are reflected in this dataset. As shown in Figure 3.3, males exhibit consistently higher mortality rates than females across most ages, a pattern attributable to a combination of biological, behavioural, and socio-economic factors. A third category, representing individuals with inconsistent gender records across linked datasets, was present but contained too few observations for reliable modelling. This group was therefore excluded from further analysis.

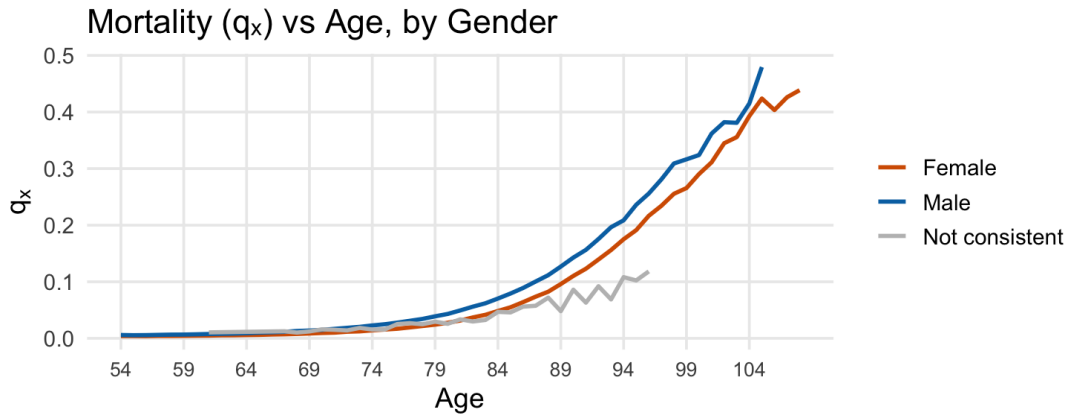


Figure 3.3: Mortality against age by gender. Male mortality is consistently higher than female mortality across most ages. A small group with inconsistent gender entries was excluded from modelling due to insufficient sample size.

3.7.1.2 Mortality Trends Against Health Related Variables

While demographic features such as age, year, and gender explain much of the broad structure of mortality, they cannot capture the substantial variation driven by health status, disease history, and patterns of medical care. Health-related variables provide critical additional context, as they directly reflect underlying morbidity profiles and healthcare interactions that influence longevity outcomes.

The following sections examine how mortality varies across key health-related dimensions, including diagnosed conditions, pharmaceutical usage, and the intensity of healthcare interventions. Together, these analyses provide a more nuanced understanding of the impact of the variables offered in the MBS and PBS datasets.

Mortality Against Quantity of Prescription Units Supplied. Figure 3.4 shows the reverse of the expected trend, the mortality decreases as the total quantity of prescription units supplied (`total_qty`) increases. There are a few potential reasons for this:

1. Firstly, there might be a impact from death on the total quantity. Given someone died within a given year, they will likely have a lower time period to buy units.

2. Another possible explanation is under-diagnosis of diseases, resulting in those with lower total quantity to be under treated. Higher quantity often means more contact with healthcare and better monitoring.

3. Finally, this can be the result of ceasing medication at end of life though still contributing to high mortality.

The first reason, which seems the most intuitive, is also problematic as it implies that the total quantity is directly affected by mortality, and thus should not be used to predict mortality. This is due to the information leakage that occurs when using a variable impacted by the target variable of the model. Furthermore, this would imply that all MBS/PBS variables are also impacted by the same problem, as they are also cumulative. To further ensure this is the case, the trend was checked for total script variables too. Though not as prominent the total scripts also showed a negative corelation between variables.

There are two main approaches to deal with the identified information leakage in using the raw script, number and quantity variables.

1. Using the previous year's MBS and PBS variables

2. Converting the MBS and PBS variables into indicator variables

Each approach has trade-offs, lagged features allow for us to analyse in detail the impact of the number of diagnosis, the number of scripts dispensed and the number of units supplied. Since we only have a short tail of data, which will further reduce with the use of a lagged variable this approach is not ideal. Alternatively, if we convert the variables to indicator variables, this allows us to use more recent information

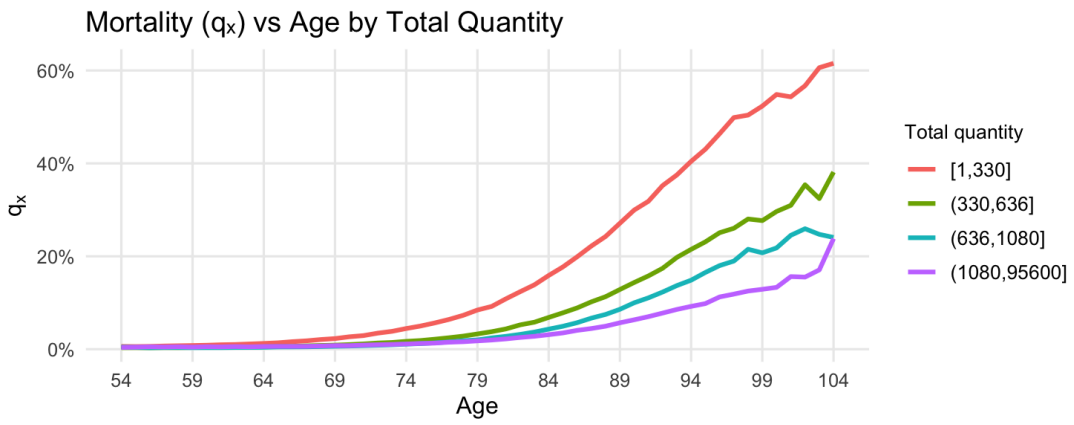


Figure 3.4: Mortality against total quantity of prescriptions supplied. Total quantity is split into quantiles.

and create a questionnaire that is more answerable from retirees. Furthermore, as analysed in Section 3.7.1.2, using raw counts does not provide a significant improvement over using a binary indicator of whether the condition is present. Thus, the raw dataset was converted into binary variables (this will be referred to as the *indicator dataset*).

Number-of-Services vs Indicators. Figure 3.6 examines the `num` variables for mental health and pain medication to assess whether summarising them as binary indicators ($\geq$ 1 vs 0) sacrifices important information. In both plots of Figure 3.6, the largest mortality gap is between individuals with no recorded condition and those with at least one. Beyond one service, the incremental difference is modest, with only small separations between one and multiple conditions. These number variables are also long-tailed: a small subset accumulates high values, stretching the scale and rendering models with many distinct count levels impractical and hard to interpret. Retaining all values would increase variance and model complexity without clear gains in fit or clarity. Given the sharp zero-to-one contrast and the flattening thereafter, we summarise these variables as binary indicators. Using the approach described in Section 3.5.

Mortality Against the Standard MBS and PBS Dataset. Figure 3.7 illustrates mortality trends by condition, based on the original diagnostic variables from the MBS and PBS datasets. Because individuals may satisfy multiple diagnostic definitions, they can appear in more than one condition curve this overlap is intentional, as it reflects the condition-specific mortality profile.

The most prominent finding is that mortality among those prescribed pain medication is substantially higher than in all other groups. Mortality among individuals with diabetes or mental health conditions follows a broadly similar trajectory but remains noticeably lower than other condition-specific curves. These differences high-

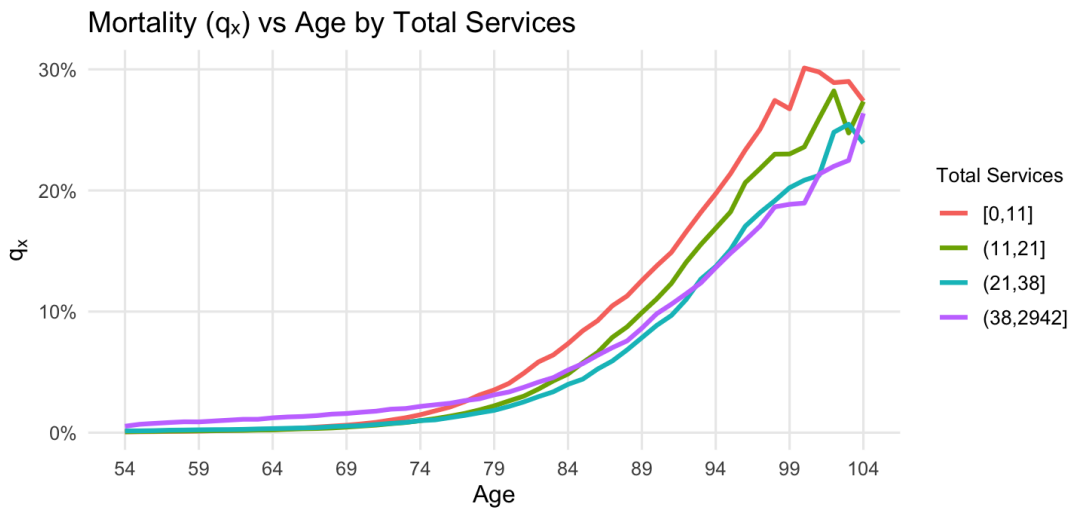


Figure 3.5: Mortality against the total scripts dispensed. Where the total scripts number is binned into four quantiles.

light the heterogeneity in mortality experience across conditions. To further examine whether a condition itself significantly influences mortality risk, we compare mortality for individuals with and without each condition in Figure 3.8. This comparison shows that diabetes and pain medication are associated with particularly elevated mortality risk, whereas conditions such as mental health diagnoses and coronary bypass procedures appear to have a smaller relative impact.

3.7.1.3 Mortality Against Summarised Health Variables

This section examines how mortality differs between individuals with and without major health conditions, using simplified diagnostic categories derived from the MBS and PBS data. As described in Section 3.5, the MBS and PBS variables were subsequently summarised into broader diagnostic categories to improve interpretability and support downstream modelling. Using these summarised variables, Figure 3.9 compares mortality between individuals with and without each condition.

The results show that individuals with diabetes or mental health conditions exhibit slightly lower observed mortality. This counterintuitive result is likely driven by under-diagnosis and treatment heterogeneity within these groups rather than reflecting a genuine survival advantage. Lipid disorders show no significant association with mortality, and coronary bypass procedures (though common) are not well distributed across age groups, limiting their suitability for inclusion in the final models.

Use of pain medication is associated with approximately a threefold increase in mortality, indicating a strong relationship between prescription pain management and death rates. One possible explanation is that pain medication is frequently prescribed for palliative care at the end of life. However, this pattern appears to be driven primarily by individuals who were prescribed pain medication in isolation,

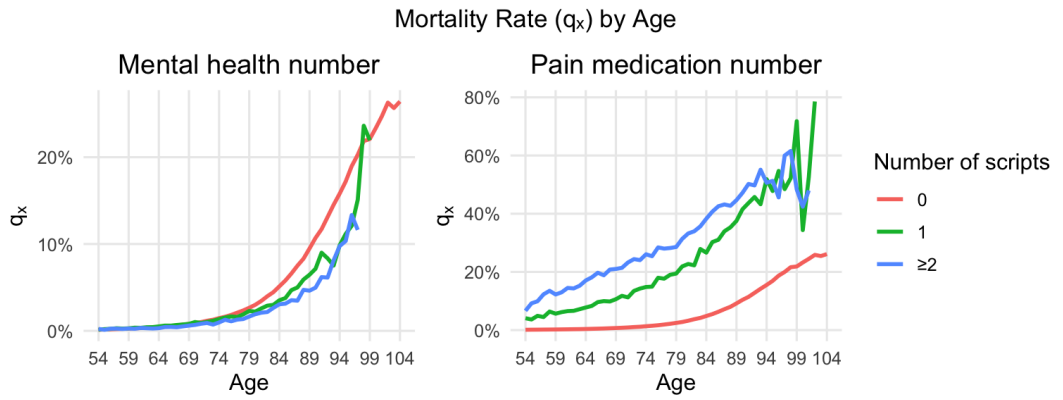


Figure 3.6: Plotting the number for mental health services and pain medication services. For those with 0,1, and at least 2.

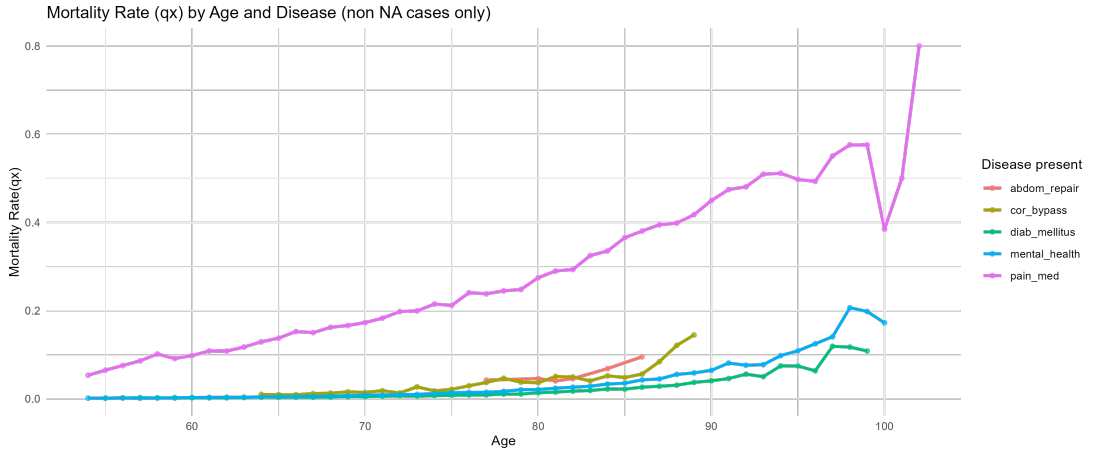


Figure 3.7: Mortality against selected conditions in the original dataset, using MBS-based disease variables.

without accompanying diagnoses or other prescriptions. Further analysis was undertaken to assess whether this elevated mortality aligns with the higher mortality observed among more socially disadvantaged groups and to investigate additional potential explanations for this trend (see Section 3.7.1.4).

3.7.1.4 Pain Medication Mortality Analysis

Demographic Profile of Pain-Medication Users. We first describe the distribution of individuals with pain medication across demographic features (Figure 3.10). As seen in Figure 3.10, a higher proportion of those who take pain medication have years 10 and above education, are female, are homeowners, are married with low income. Because these patterns partly mirror the overall population, we compared them to the distribution among non-users. To mitigate class imbalance, we randomly

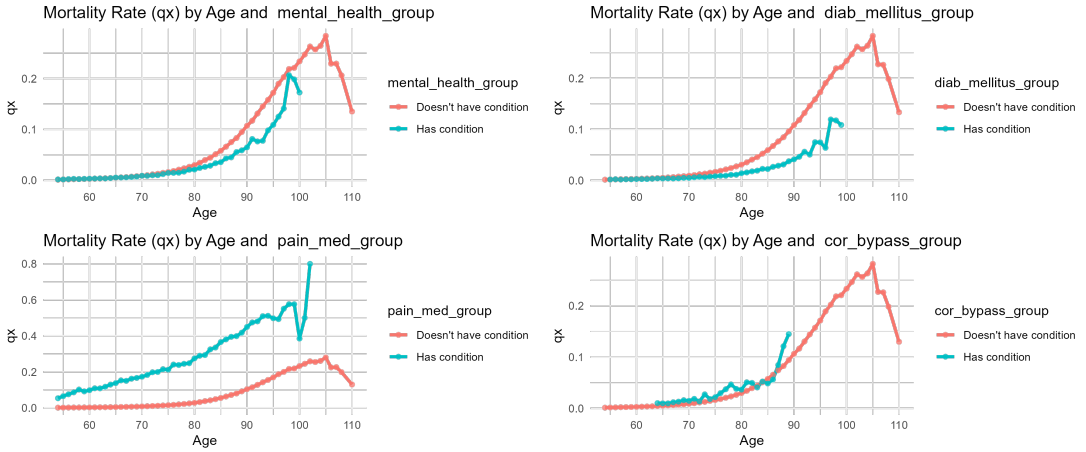


Figure 3.8: Mortality comparison for individuals with (1) and without (0) selected conditions, using summarised diagnostic variables.

drew 12,000 observations from each group, producing a balanced evaluation sample. For each demographic category, we calculated the conditional probability of pain medication use as, $P(\text{pain} = 1 \mid \text{variable} = v)$ where v represents a specific category of the demographic variable. We then computed the difference in proportions of pain medication use between users (pain $= 1$) and non-users (pain $= 0$) within each category.

The raw distribution in Figure 3.10 suggests that pain-medication users are more common among people with $\geq$10 years of schooling, females, homeowners, and married low-income groups. However, once we compare $P(\text{pain} = 1 \mid \text{variable} = v)$ across categories (Figure 3.11), these apparent differences largely disappear: the estimated probabilities cluster within a narrow range and no single demographic attribute stands out as a strong discriminator. In other words, much of the pattern in the first plot reflects overall population composition rather than a large effect of any one demographic factor. We therefore treat demographics as control/stratification variables and place emphasis on clinical signals (diagnoses, medication classes) for modelling pain-medication use and, downstream, mortality risk.

Another potential explanation of the increase in mortality seen with pain medication is due to the increase in opioid addictions seen in Australia in the same time frame [Australian Institute of Health and Welfare (AIHW), 2021]. According to the Australian Institute of Health and Welfare, the number of deaths due to opioid has increased from 439 in 2006 to around thrice the value (1119) by 2016.

3.8 Key Empirical Findings and Modelling Implications

1. **Age gradient and gender gap.** q_x rises monotonically with age, with right-tail noise due to small exposures. Moreover, male mortality exceeds female mortality at most ages.

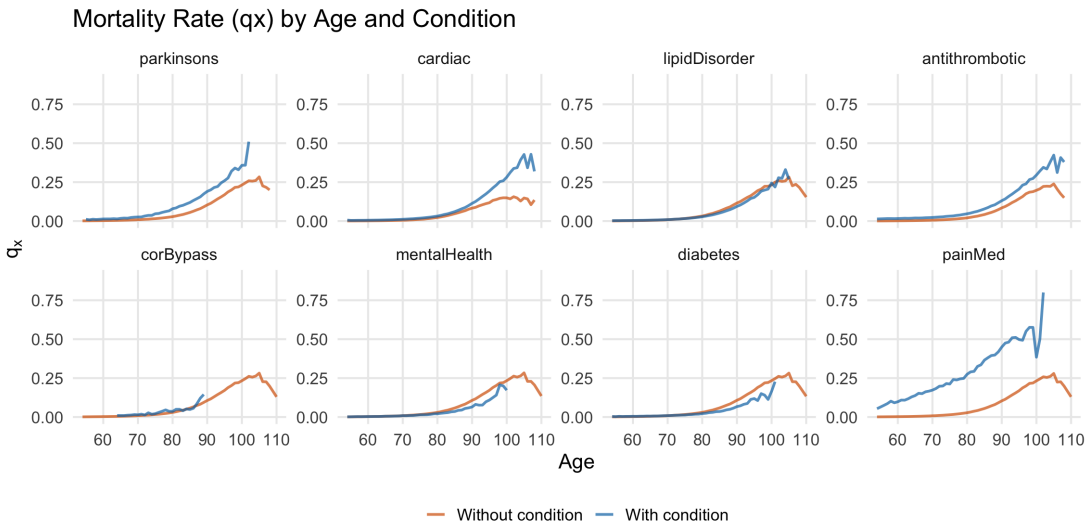


Figure 3.9: Mortality with and without selected conditions for those most prominent in the summarised dataset.

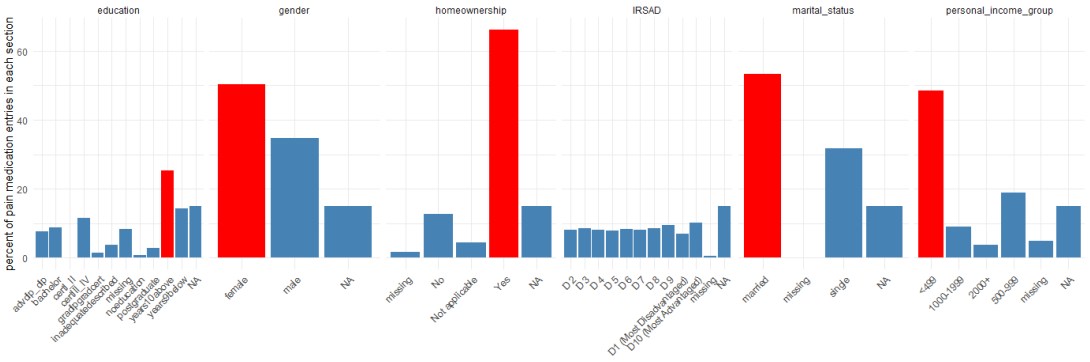


Figure 3.10: Distribution of those who take pain medication across demographic features.

Note: **NA** indicates values missing or unavailable in the source data after linkage or disclosure controls. **Not applicable** means the variable does not apply to the individual, for example where a category is irrelevant to that person's circumstances.

Implication: Age and gender remain key drivers of mortality and were incorporated into most models developed in this thesis. Where they were not used for grouping, mortality within each cluster was calculated separately by age and gender.

2. **Year-wise stability (2011–2016).** Curves are similar across years, consistent with a short observation window.

   Implication: we don't need to make any further adjustments to the observed mortality to account for changes in mortality over the short period.

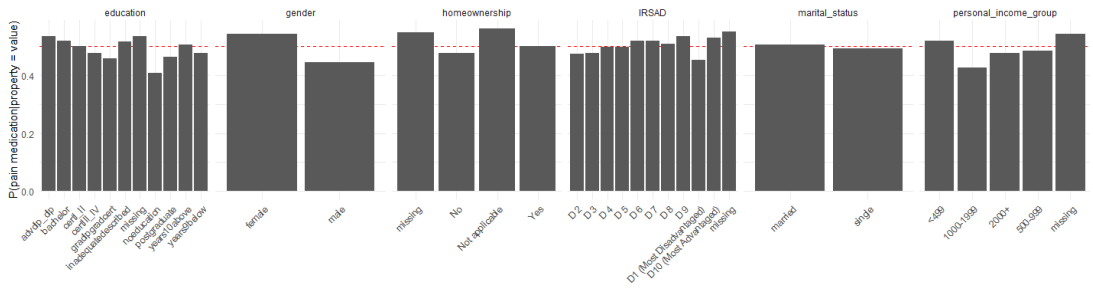


Figure 3.11: Given individuals from a certain property, the proportion of those which take pain medication.

3. **Counts vs indicators.** There is likely information leakage that occurs with the raw `scripts/qty/num` resulting in the seen negative corelation. The largest risk separation is between "none" and "at least one" service, with diminishing marginal signal thereafter (Figure 3.6).

   Implication: prefer binary indicators over raw counts found in the MBS and PBS datasets.

4. **Condition signals.** Pain-medication exposure is associated with a *threefold* increase in mortality relative to non-exposed peers, diabetes and mental-health indicators show smaller effects.

   Implication: The extent to which the model can capture mortality trends linked to factors like pain medication use serves as a useful indicator of its effectiveness in identifying key drivers.

These observations directly inform the group-based modelling framework described in Chapter 5, guiding the inclusion of gender and age specific calculations in all models, as well as the use of right-tail smoothing and feature construction methods that prevent information leakage.

3.9 Conclusion

This chapter examined the dataset in depth, detailing the preprocessing steps undertaken, describing the derived variables, and analysing key trends. The next chapter introduces the modelling objectives and outlines the approach used to construct the models. A central focus of that discussion is how limitations in the data, particularly information leakage within the raw MBS and PBS variables, informed the modelling decisions.

Chapter 4

Proposed Modelling Approaches

This chapter introduces the model objective (Section 4.1), describes the adjustments from crude annual mortality needed to align the results with the ALT (Section 5.2), and sets out the constraints on the modelling process (Section 4.2, Section 4.2.1). This chapter also performs an analysis of model limitations caused due to technical constraints (Section 4.10). It then outlines the PCA analysis on the variables used for dimensionality reduction and variable selection. Due to the lack of prior studies performed on this data, this chapter provides a detailed analysis of potential models and their implementation (Sections 4.4–4.9). The analysis is continued in Chapter 5 with performance metrics to evaluate the models.

4.1 Model Overview

4.1.1 Objective

The main goal of the proposed model is to incorporate the impact of health variables to better predict mortality. There are a few constraints (Section 4.2) on how this model should be constructed to ensure its application in the superannuation industry. Overall, as no previous models have been built to incorporate health variables, the goal is to outperform the ALT in providing more granular predictions based on health data.

Initially, the goal identified was to have a good prediction accuracy, but this was insufficient as the ALT is very accurate due to consistency within the mortality curve seen amongst retirees (Figure 3.2). The more granular we get with the mortality prediction the more deviation we would notice, thus there is a clear trade-off between the following three goals:

1. identifying a binning mechanism that captures differences in mortality across clusters as large as possible,

2. predicting the mortality of a certain cluster accurately, and

3. having large enough bins to ensure mortality prediction is not personalised (further discussed in Section 4.2).

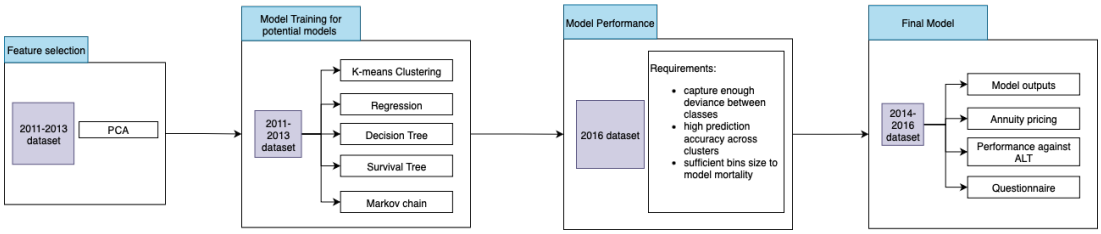


Figure 4.1: A summary of the methodology approach followed, including the data used in each stage. Split by feature selection, model training, model performance, and final model construction. The models used in the training stage have been summarised under Table 4.1.

Evaluation and selection of a model that best meets these requirements will be discussed in Section 5.4. Note that this task is not simple, as mortality rates, the target variable, is not defined prior to knowing the binning mechanism.

4.1.2 Modelling Pipeline

The modelling pipeline is summarised in Figure 4.1. Since, in practice, there is often a delay of around three years between when mortality events occur and when the corresponding life table data is released, we incorporate this lag into our modelling pipeline. This lag typically reflects the time required for death certification, registration of death, cause of death identification, and quality assurance processes. Thus, model training was done with 2011-2013 data while model performance testing was done on 2016 data. This ensures that the model performs reasonably, even in the absence of the most recent data, as required. After model selection, it is refitted using most recent data (in this case 2014-2016) so that the latest information is utilised in the final model. The final model results are provided in Chapter 6.

Note that testing against expected lifespan would be ideal as it accounts for an aggregated impact of predicted mortality. However, as the training data we have only spans the three year window for modelling, we are not able to test the expected lifespan of individuals against their predictions. Instead, model tests and performance metrics have been done for mortality rate only.

4.1.3 Model Selection

Model selection used two criteria; an accuracy metric and a deviation (calibration) metric defined in Chapter 5. Accuracy is assessed directly on predicted mortality rates.

To capture real-world implications, we also derive life expectancy from each model: as an aggregate of the mortality curve, it serves as our deviation/impact measure. The resulting life expectancies are reported in Section 5.3 and compared directly with the ALT to understand the monetary impact of this research.

There are multiple modelling approaches available for the mortality prediction task, each with its own strengths and limitations. Consequently, no single approach

can be deemed universally optimal. The chosen methods were first examined from a theoretical perspective, then implemented and evaluated, with their performance compared to the performance of the ALT as the benchmark to determine the final selected model. As the trade off between granularity and predictability is not clear, in order to identify the best approach this study aims to implement a range of potential models to perform this task. Given the lack of prior work, a comprehensive analysis of the model is required to identify the best model.

The model uses the MBS and PBS datasets to generate the crude mortality (the model outputs will denote crude mortality as q_x) defined in Eq. (3.1). As the ALT does adjustments to the crude mortality, these are replicated for the model output to calculate the overall final mortality, denoted as q_x^*. This is further detailed under Section 5.2.

4.2 Constraints on the Modelling Process

Given the intended practical use of this research, two constraints were imposed on the modelling approach: (i) safeguards against provider risk selection and (ii) interpretability.

4.2.1 Provider Risk Selection and Design Safeguards

Provider risk selection refers to the behaviour that occurs when insurers or superannuation providers use detailed risk information to shape the type of people they serve. If a model can predict an individual's likelihood of dying earlier or living much longer, a provider could use that information to attract customers who are cheaper and more profitable to serve, while discouraging or avoiding those who are likely to be more expensive. This can happen in subtle ways. For example, a life insurance company offering term insurance might offer lower fees or better benefits to people predicted to live longer, design products that mainly appeal to low-risk customers, target marketing campaigns to healthier groups, or create extra barriers for people with higher predicted risks.

Implication for this study. Because individual-level predictions could be misused in this way, this study deliberately avoids producing or releasing mortality scores for individual people. Instead, all results are presented only at an aggregated level, such as by age, gender, or broad health group. This ensures the model still provides useful insights for designing policies and retirement income strategies but cannot be exploited to screen or exclude individuals based on their personal risk profile.

4.2.2 Interpretability and Communication.

The model must be thoroughly audited, clearly understood, and effectively communicated to stakeholders, including members and regulators in the insurance and superannuation industries, to ensure they are aware of the key pricing drivers and

that transparency is maintained. This is important as fair pricing within retirement products can drastically change the living standards of retirees. Thus the model prioritises:

1. **Simple, explainable structure:** modelling at the cohort level (age, sex, broad health states) with documented feature construction and minimal transformations.

2. **Transparent, single output modelling process:** the transparent process applied to the modelling of mortality implies that the model can be audited with no ambiguities.

3. **Member-facing explanations:** plain-language summaries of what drives cohort rates and how these differ from standard age–sex tables. This includes a mapping between the answers to the questionnaire and the resultant pricing model output.

4.2.3 Design Decisions

To avoid the challenges associated with individualised mortality predictions, as outlined in Section 4.2.1, the proposed models adopt two alternative approaches.

1. Group individuals into clusters based on features that are expected to exhibit similar mortality patterns.

2. Discretise (bin) the input variables and estimate mortality rates for each binned combination of variables.

Both approaches avoid individual-level predictions and are therefore well suited for use in the retirement industry.

Additionally, to ensure model interpretability the potential models fitted in this research were initially selected based on their interpretability. These are further detailed under Section 4.4.

4.2.4 Model Supervision

The advantages and disadvantages for both supervised and unsupervised learning approaches for this model are key to deciding which approach to take.

Supervised learning (regression modelling, decision tress and survival trees) offers several benefits, including direct optimisation for the chosen outcome, the ability to validate performance using standard metrics, and the ability to provide variable importance measures and fairness checks. However, it also has notable drawbacks: it requires labelled outcomes, and if age at death or years until death are used as targets, the training set is limited to individuals who have died, leading to a significant reduction in available data. There is also a risk of information leakage, over fitting, and capturing patterns in data rather than genuine health-related trends. Moreover, the lack of a robust target variable is a key challenge, as mortality is dependent on

the model's partitioning and is not directly available in the dataset beforehand. The obtained results for the models along with advantages and disadvantages have been summarised under Table 4.1.

In contrast, unsupervised learning (k-means and Markov chain) does not require labelled outcomes, allowing the entire dataset to be used for training, and can reveal latent structures and subgroups within the data. Nonetheless, its connection to outcomes can be uncertain; for example, when disease-related or raw data are used in k-means clustering, the resulting clusters are based on latent structures without a guarantee that mortality trends differ as intended. Additionally, unsupervised methods can be sensitive to scaling and methodological choices, suffer from unstable clusters or factors, and are generally harder to validate and justify operationally.

4.3 Summary of Datasets Used

For clarity, three related datasets were constructed and used at different stages of the analysis:

1. **Raw MBS–PBS Dataset:** The original linked data with detailed service counts (`_num`), prescription items (`_scripts`), and quantities (`_qty`). Used primarily for exploration and variable screening. A survival–tree experiment was also run on this dataset to demonstrate the risk of information leakage and the resulting over-optimism; it was not used for final modelling due to leakage risk and computational constraints.

2. **Indicator Dataset:** (Section 3.7.1.2) A fully binarised version of the raw data in which each variable was converted to an indicator (present/absent). This format improves interpretability and aligns with questionnaire-style inputs. It was analysed using k-means clustering and survival trees to assess robustness and to verify whether summarisation reduces leakage relative to the raw dataset.

3. **Disease Dataset (Interpretable):** (Section 3.5) A clinically grouped summary of the Indicator Dataset, aggregating related indicators into common disease categories (e.g., diabetes, mental health, pain medication, lipid disorders, cardiac therapy). Variables were retained only if they met a minimum support threshold (e.g., at least 100,000 non-missing observations) to ensure stability. This was the primary dataset for downstream modelling (regression, survival trees, and Markov process chains) and for deriving cohort life expectancy.

4.4 Analysis of Potential Models

Before the modelling process a principal component analysis (PCA) was performed for variable selection. Though the principal components identified were not used within models due to lack of transparency, a variable importance measure was generated, allowing us to select a subset of variables. This is further detailed under Section 4.5.

Table 4.1: Summary of candidate approaches (inputs, q_x production, strengths/limits). Notes: *Lrn* denotes how the learning occurred, whether or not its supervised (*S*) or unsupervised (*U*). Additionally, note that the k-means and survival tree models were trained on the raw datasets, due to information leakage the models have been provided in Appendix B and not in the main body.

Model	Lrn.	Primary inputs (years)	Target / Output	How q_x is obtained	Key strengths	Key limits / assumptions
k-means	U	Indicator dataset, and the disease dataset	Cluster ID	Observed q_x by age/sex within cluster	Simple to explain, interpretable questions for a questionnaire	Not outcome-guided; possible weak mortality separation; static health state
Decision tree	S	Raw & disease (train 2011–2013)	Age at death; years to death; yearly death flag	Leaf death rate as cohort q_x (or map leaves to cohorts)	Transparent, auditable rules	Leakage if age included; without age weak signal; deep trees reduce interpretability; static state **noncomparable performance**
Survival trees	S	Disease and Indicator datasets; censoring-aware	Node survival curves	From node $\widehat{S}(t)$ via $q_x = 1 - \widehat{S}(x+1)/\widehat{S}(x)$	Handles censoring; non-linearity/ interactions; segment curves	Needs enough events; size vs. interpretability trade-off; static
Regression	S	Disease indicators (6), sex, age spline; test 2016	$\log q_x$ by age, sex, disease combo, target variable q_x is calculated for each combination of diseases using 2011–2013	Age spline + disease main effects & age–disease interactions; predict q_x per combo	Strong vs. ALT, esp. high morbidity; smooth; coefficients auditable	Assumes stable trends; potential form mis-specification; static health state
Markov chains	U	15 interpretable disease states; transitions 2011–2013 (by sex; coarse age bins)	State transition matrix (yearly)	State-specific q_x; project paths with transitions for life-table/LE; compare with/without transitions	Captures health dynamics; quantifies transition impact on LE; complements static models	Coarse/"other" states mix heterogeneity; limited years ⇒ coarse age bins; no explicit death state

The subset of models to be trained were selected based on the theoretical interpretability of the model, which resulted in the following subset of models:

1. Clustering using k-means (Section 4.6)
2. Survival trees (Section 4.7)
3. Regression (Section 4.8)
4. Markov process chains (Section 4.9)

In addition to these, a decision tree approach (Section 4.7.1) was also explored early in the modelling process. Although not included in the final shortlist above, this attempt was valuable in testing the feasibility of supervised learning techniques. However, the absence of a clearly defined target variable for optimisation significantly limited its usefulness. Without a strong response variable, the decision tree produced results with low interpretability and poor predictive performance (further detailed in Section B.2). This limitation also reduced the potential to extend the analysis using neural networks with explainable AI, which had been part of the initial research plan. As a result, the focus shifted toward models better suited to the available data structure and research objectives.

In principle, neural networks combined with post-hoc interpretation methods such as SHAP [DataCamp, 2023] or LIME [C3.ai, 2025] could have provided a flexible framework for capturing complex, non-linear relationships in the data while retaining some degree of interpretability. However, without a robust and meaningful target variable to guide model training, any predictions produced by such models would have been unreliable and potentially misleading. As a result, we instead turned to survival analysis methods, and specifically survival trees, which are naturally suited to censored time-to-event data and better aligned with the underlying structure of the problem.

This subset of models ensured that the results were interpretable and thus the final model selection was performed based on the performance of the models (Section 5.4). All of these approaches can be incorporated into the preceding work done, Huang et al. [2023], on incorporating demographic variables.

4.5 PCA Analysis on the Variables

One main limitation of principal component analysis (PCA) directly within models is the lack of interpretability, as each principal component is a linear combination of many original variables. This makes it difficult to translate the results into questionnaire items.

However, PCA offers several potential benefits for the dataset:

- **Dimensionality reduction:** The dataset contains a large number of variables for conditions, procedures, and services. Many of these variables are correlated (e.g. there are certain medications, prescriptions, procedures and visits that occur together). PCA can transform these into a smaller set of uncorrelated components that capture most of the variation in the data, reducing noise and computation time in downstream models.

- **Handling multi-co-linearity:** Many of the health indicators are correlated, which can reduce the performance of certain models. PCA produces orthogonal components, mitigating multi co-linearity and improving stability for regression-based or distance-based methods.

- **Pre-processing for state definition:** When defining states for Markov process modelling, using the raw variables can result in a very large number of rare states. Clustering in PCA space can group similar health profiles into fewer, more stable states while retaining the main patterns of variation. This is further detailed under Section 4.9.

4.5.1 Correlation Between Variables

The analysis found three pairs of variables `N07B_scripts` and `N07BA_scripts`, `C05A_scripts` and `C05A_qty`, and `N07B_qty` and `N07BA_qty` that are perfectly correlated (correlation coefficient of 1), indicating redundancy and suggesting that one variable from each pair could be removed to avoid duplication in PCA.

4.5.1.1 Variable Importance

Methodology. We use a subset of ten million entries due to computational limits of DataLab (Section 3.1). Let $x_1, x_2, \ldots, x_p$ be the variables in the dataset. We first standardise the variables to ensure that the variance in each variable is consistent. PCA determines p principal components, which are the directions that best explain the variance of the full (standardised) dataset in order, as well as exactly how much of that variance is explained by each principal component.

Using this, we can determine the amount of variable j's variance explained by principal component k, which we denote $\sigma^2_{j,k}$. Note that the variance of variable j is given by $\sum_{k=1}^{p} \sigma^2_{j,k}$, and the total variance of the dataset is $\sum_{j=1}^{p} \sum_{k=1}^{p} \sigma^2_{j,k}$.

We assume that the true variance in the data, i.e., variance that does not arise from noise, is well modelled by the first n ($n < p$) principal components. A total of $n = 33$ principal components was identified as sufficient by plotting the explained variance against the number of components and selecting the point where 95% of the total variance in the dataset was captured.

Then our measure of the importance of variable j is the proportion of variance of the variable j explained by the top n principal components relative to the total variance captured by those top n components:

$$\text{Importance}_j^{(n)} = \frac{\sum_{k=1}^{n} \sigma^2_{j,k}}{\sum_{x=1}^{p} \sum_{k=1}^{n} \sigma^2_{x,k}}.$$

A larger $\text{Importance}_j^{(n)}$ indicates that the variable j plays a major role in shaping the dominant variance patterns captured in the reduced-dimensional space, whereas smaller values indicate that it mostly influences residual variance in latter components.

Mathematical Interpretation. Given a dataset of m data points over p variables x_i, PCA solves for the directions that maximise the total dataset variance (sum of the variance of each variable) when the dataset is projected onto that direction. These

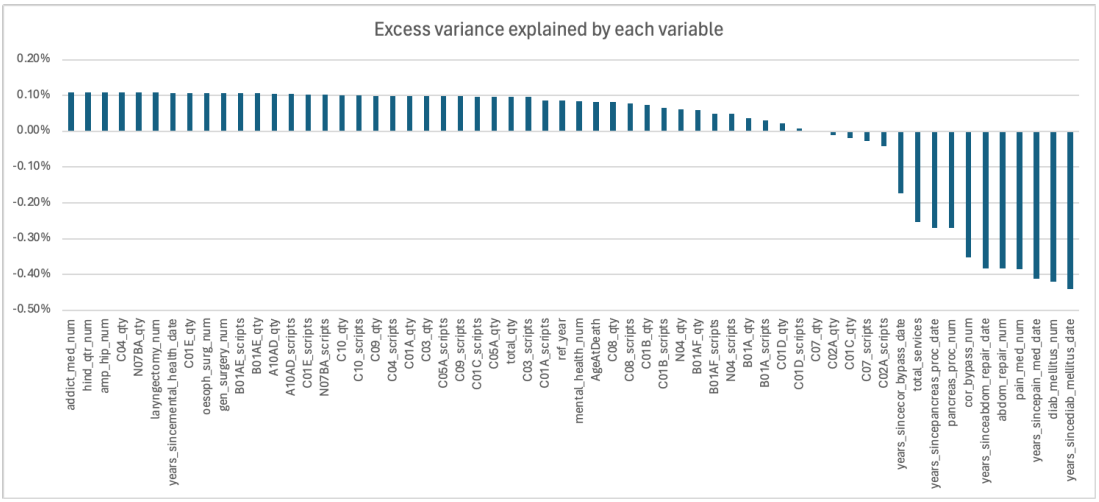


Figure 4.2: Excess variance explained by the variable compared to a benchmark if all variables explain equal variance.

directions are called the principal components, the first principal component is the direction which maximises dataset variance when projected onto it, and the k^{th} principal component is the direction orthogonal to the first $k-1$ principal directions that maximises dataset variance when projected onto it. PCA not only returns principal components v_k, but also the amount of variance explained by each component λ_k.

Let v_{jk} denote the j^{th} element of v_k, then as v_k is orthonormal $\sum_j v_{jk}^2 = 1$, and v_{jk}^2 denotes the relative proportion of variable j that is in the direction of component v_k. As a result, the amount of variable j's variance explained by principal component k is given by $\sigma_{j,k}^2 = v_{jk}^2 \lambda_k$.

Results. Variable importance was calculated, and the equal-contribution baseline was subtracted to reveal the additional variance each variable explained beyond uniform allocation (Figure 4.2). This was used to select subsets of variables for models that needed a smaller dataset for interpretability.

Ideally the difference between the positive weights would be higher allowing us to select a smaller subset. Given the difference is minor no variable explaining above the equal variance benchmark was removed. These results were used to omit the variables that explained less than the benchmark of equal weights with the exception of pain medication as it shows clear implications for mortality (Section 3.7.1.4). Additionally, variables `pancreas_proc_num`, `cor_bypass_num`, `abdom_repair_num`, `diab_mellitus_num` `total_services` were not removed as they were more interpretable when compared to the script and quantity variables that may corelate to this. i.e., `C02A_qty`, `C01C_qty`, `C07_scripts`, `C02A_scripts`, and the years since variables were omitted from the models.

4.6 K-means Clustering

K-means clustering is an unsupervised algorithm that partitions data into k clusters by choosing centroids that minimise the sum of squared distances from observations to their nearest centroid.

A k-means clustering algorithm was trained on the indicator dataset and the disease dataset. Advantages and disadvantages specific to this model have been summarised under Section 4.6.1. This section introduces both models, the mortality distributions for each cluster, the performance in comparison to the ALT and an analysis of the output.

4.6.1 Advantages and Disadvantages

As discussed earlier, k-means is trained on both the indicator and the disease dataset Section 4.3, both with their own advantages and disadvantages discussed under Section 3.5.

The model is static and when used for pricing of retirement products it will assume that individuals do not change clusters over time. Additionally, as the model is unsupervised the clusters may not have any real implication for the actual mortality but instead just group based on the input variables.

The available memory (within the secure virtual environment (DataLab)) was insufficient to run batch k-means on the full dataset, and in practice was also inadequate for a single-year when using the standard algorithm (multiple full passes over X with large working buffers). To mitigate this, we employed *mini-batch* k-means, which performs centroid updates on small, randomly sampled batches and reduces per-iteration RAM and I/O. Even so, within the ABS DataLab constraints we could only train on a single-year slice (2011) rather than the intended 2011–2013 window. This restriction is a primary disadvantage of the approach: the learned clusters reflect 2011 composition only and cannot capture temporal variation across subsequent years, limiting generalisability and downstream comparability.

For k-means, we used Euclidean distance to assign points to clusters. However, for the raw dataset (61 variables), the reliability of this distance measure decreases as dimensionality grows. In high-dimensional spaces, most data points become almost equally distant from one another, a phenomenon known as *distance concentration* [Aggarwal et al., 2001]. This happens because many weak or irrelevant features add noise to the distance calculation, making truly similar and dissimilar points appear more alike. As a result, Euclidean distance struggles to capture meaningful structure, and the algorithm may form clusters that do not reflect real patterns in the data. The effect is particularly pronounced when features vary widely in scale, or when many dimensions contribute little information. Consequently, cluster assignments in the raw and interpretable datasets may be less robust. By contrast, the disease dataset contains only six variables, so the distance metric retains more discriminatory power and produces more meaningful clusters.

4.6.2 Indicator Dataset

The raw variables in the MBS and PBS datasets were fed into a model, allowing for the clustering to account for individual script variables, quantity variables and num variables. The resulted model was used to run predictions for the entire available dataset.

The results are promising,

- They align with key trends noticed during EDA
- The results are consistent throughout the years indicating that as expected the trend patterns remain similar within groups

Though the model was completed, unfortunately the model was not vetted in time and hence numerical results for this section can not be included in this thesis.

4.6.3 Disease Dataset

The k-means algorithm was rerun to analyse the interpretability-performance trade-off caused due to the disease dataset (Section 5.4.3). Using this dataset allows for more understandable clusters and better questionnaires.

4.6.3.1 Model

As all the disease variables were indicator variables, there were no scaling required. The elbow plot identified 8 clusters to be optimal, and thus trained on. The resultant centroid coordinates are shown in Table 4.2

Table 4.2: Centroid coordinates for the K-means model trained on the disease features (dominant features for each centroid in **bold**).

Cluster	Parkinsons	Cardiac	Anti-throm	Cor-Bypass	mental health	Diabetes	Pain med
1	0.00	0.00	**0.05**	0.00	0.00	0.01	0.00
2	0.00	**1.00**	0.00	0.00	0.00	0.00	0.00
3	0.01	**0.96**	0.44	0.00	0.01	**1.00**	0.00
4	0.00	**1.00**	**1.00**	0.01	0.03	0.00	0.01
5	0.02	0.00	0.06	0.00	**1.00**	0.02	0.00
6	**1.00**	0.45	0.00	0.00	0.00	0.00	0.00
7	0.02	**1.00**	0.00	0.00	**1.00**	0.06	0.00
8	**1.00**	0.76	**1.00**	0.00	0.08	0.03	0.00

Cluster Interpretation. Because k-means was applied to indicator variables, the resulting clusters should be interpreted as patterns of co-occurrence rather than distinct clinical subgroups. Each cluster groups individuals with similar combinations of disease indicators, highlighting common profiles present in the data:

- **Cluster 1:** A small proportion of individuals with anti-thrombotic treatment indicators, with few other recorded conditions.
- **Cluster 2:** Individuals primarily associated with cardiac-related indicators, showing minimal presence of other health factors.
- **Cluster 3:** A group with frequent co-occurrence of cardiac and diabetes indicators, suggesting a metabolic–cardiovascular pattern.
- **Cluster 4:** Individuals with both cardiac and anti-thrombotic indicators present, indicating a pattern of more complex cardiovascular management.
- **Cluster 5:** A cluster where mental health indicators are dominant, with limited co-occurrence of other conditions.
- **Cluster 6:** Individuals with Parkinson's-related indicators, often appearing without strong associations to other disease categories.
- **Cluster 7:** A pattern characterised by co-occurrence of cardiac and mental health indicators.
- **Cluster 8:** A more complex pattern where Parkinson's, cardiac, and anti-thrombotic indicators frequently appear together.

These clusters do not represent predefined disease groups but rather reflect how certain health indicators tend to occur together in the population.

Pain Medication not Isolated by Clustering. None of the clusters isolate the painMed indicator. Although painMed is associated with visibly different mortality trajectories, the k-means procedure is unsupervised and optimises within-cluster similarity of the input features, not separation by outcomes. Because painMed is a single, relatively sparse binary feature with few correlated companions, its contribution to the Euclidean distance is small compared other features. As a result, k-means prioritises clusters driven by higher-prevalence and co-occurring conditions rather than isolating painMed.

4.6.3.2 Cluster Predictions

The mortality trend for each cluster in the training data is shown in Figure 4.3. Gaps at some ages arise because either observed deaths or exposure were fewer than 10, which prevents release under the ABS DataLab export rules (Section 3.1) so those points are therefore omitted. There are several ways to handle these sparse ages (e.g., logit–smoothing, partial pooling across adjacent ages, or backstopping with ALT rates), but for this section we report only the raw cluster curves; refinements are applied before scoring in Chapter 5. Consistent with the training sparsity, the corresponding test set has very few observations at those same ages (Figure 4.4).

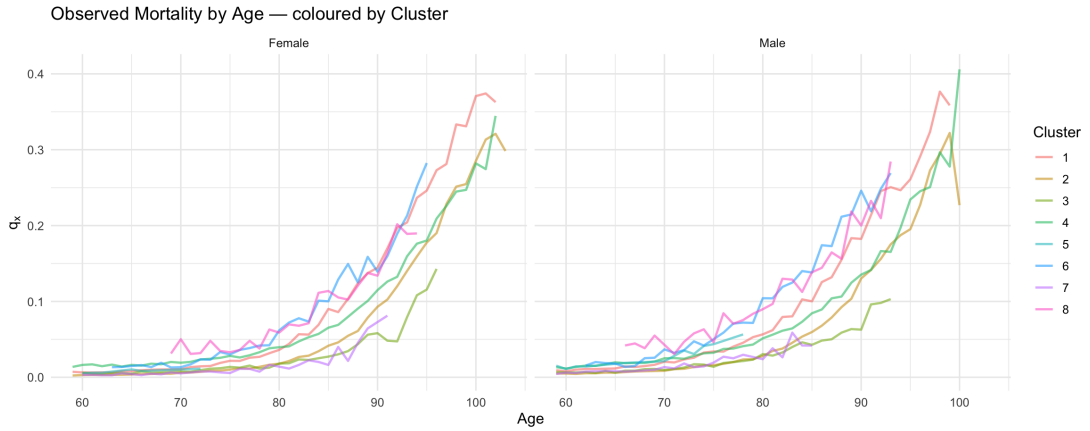


Figure 4.3: Mortality trends for clusters training on the disease only dataset.

This sparsity is not evenly distributed across clusters. Cluster 8, which captures individuals with more complex co-morbidity profiles and overlapping conditions, shows the largest gaps at younger retirement ages (typically below age 70), reflecting the very small number of such individuals present earlier in retirement. By contrast, Clusters 3, 5, and 7; which represent groups dominated by diabetes and cardiac conditions, mental health–related conditions, and mental health and cardiac, respectively; display more pronounced gaps at the oldest ages (late 90s and above). These groups tend to have higher mortality risk, meaning fewer individuals survive into very old age, resulting in insufficient observations to reliably estimate mortality rates in those age ranges.

Within each sex, the rank ordering of clusters is reasonably stable with age, and separation between clusters widens at older ages as q_x increases. As expected, male q_x exceeds female q_x at the same age for most clusters. Irregularities at the right tail (especially $x > 95$) reflect small exposures rather than structural effects; these are smoothed and aligned to the ALT conventions in Section 5.2 before model evaluation.

4.6.3.3 Results

As shown in Figure 4.4, the disease-based k-means model provides a substantially closer fit to the observed mortality rates when compared to the ALT benchmark across most clusters and age ranges. The predicted mortality curves (blue) consistently track the observed data (red) more accurately than the ALT life table estimates (green), capturing both the level and shape of mortality dynamics within each health-defined cohort. This demonstrates that even a simple unsupervised segmentation approach based on disease indicators can meaningfully improve mortality prediction over the population-level baseline provided by the ALT. But the variability in the mortality between clusters is not as significant as other models, this is likely due to the unsupervised nature of the algorithm.

There are also clusters (5 and 7) that do not span the full age range.Age was not

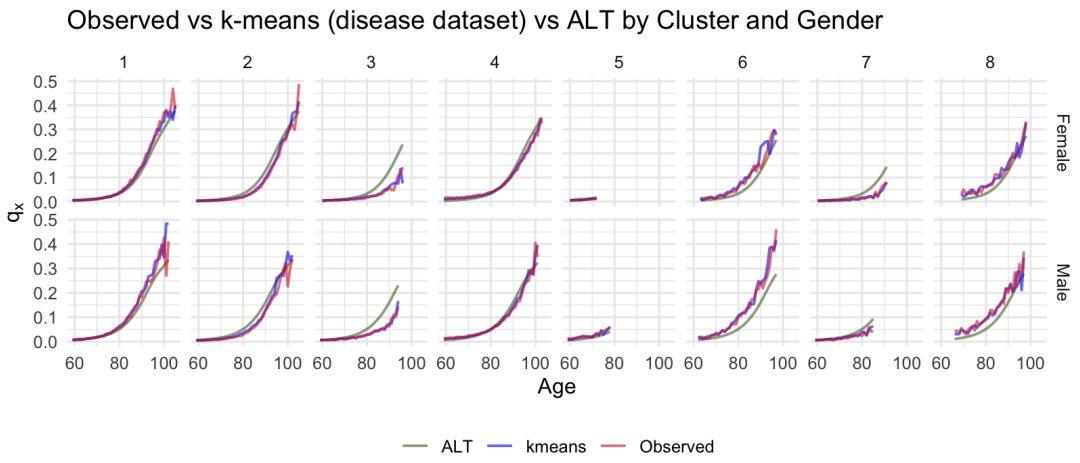


Figure 4.4: Mortality against clusters identified using the disease only dataset. Mortality predicted by the ALT is provided as a reference.

used as an input variable in the clustering process, so this truncation is a natural outcome of how the data groups. Because these clusters only include individuals present at the beginning of the retirement period, they most likely represent groups with shorter life expectancy and higher mortality risk. This aligns with their steep observed mortality curves and the absence of observations at older ages. Performance metrics for the k-means clustering using the disease-only dataset are provided in Section 5.4.1.

4.7 Survival Trees

A standard decision tree was initially trialled for mortality prediction but performed poorly due to the lack of a clear target variable (Section 4.7.1). Consequently, a survival tree model was fitted instead, which adapts decision tree methods for time-to-event data with censoring (Section 4.7.2). Survival trees are a common modelling approach in actuarial science and bio-statistics for analysing mortality and time-to-event outcomes [Hothorn et al., 2006].

The model is trained on all three datasets, due to the risk of information leakage on the raw dataset it was not provided in the main body of this paper (Section B.5.2).

4.7.1 Decision Trees vs Survival Trees

A **decision tree** is a supervised learning algorithm that partitions the dataset into subgroups where the target variable is relatively homogeneous. It does this by repeatedly splitting the data based on input features to minimise prediction error for a specified outcome, such as a binary indicator for death within a year. Standard decision trees assume that the outcome is fully observed for every individual, which

makes them unsuitable when the time of death is unknown for some individuals within the observation period (a situation known as *censoring*).

A **survival tree**, by contrast, is an extension of the decision tree designed for time-to-event data where censoring is present. Instead of predicting a single outcome, it partitions the co-variate space into groups with similar *survival patterns* and estimates a survival curve $\hat{S}(t)$ for each terminal node. This curve represents the probability of surviving beyond time t, allowing the model to incorporate individuals whose event time is not fully observed. Survival trees therefore provide richer information, capturing both the probability and timing of events, and are particularly well suited for mortality modelling where not all deaths occur within the study window.

4.7.2 Survival Trees: Concept and Mechanics

Time-to-event data and right-censoring. In the context of mortality modelling, the *event* of interest is death, and the time-to-event T represents an individual's age at death. However, in many cases, not all individuals die during the observation period (2011–2016 in this study). For these individuals, we do not observe their true time of death; we only know that they survived beyond the last point of observation. This situation is known as *right-censoring*.

Formally, let C denote the censoring time (i.e., the end of the observation window). What we observe is

$$\tilde{T} = \min(T, C), \qquad \delta = 1\{T \leq C\},$$

where $\tilde{T}$ is the recorded time (either the actual death time if the event occurred, or the last observed age if not), and δ is an indicator variable equal to 1 if the death occurred during observation and 0 if the observation was censored. For censored individuals ($\delta = 0$), we only know that their death occurs at some point beyond C ($T > C$).

Survival analysis uses this censored information to estimate the *survival function* $S(t) = \Pr(T > t)$, which represents the probability that an individual survives beyond age t. This approach allows us to incorporate all available data including individuals still alive at the end of the study rather than discarding incomplete cases, which is crucial for accurate mortality modelling.

Why use Survival Trees (and Caveats). They are non-parametric, capture non-linear effects and interactions automatically, and produce rule-based segments that are easy to audit and explain.

Predictions. Each terminal node yields an estimated survival curve $\widehat{S}(t)$ for its subgroup. From $\widehat{S}(t)$ you can derive the probability of death for a given age as,

$$q_x \;=\; 1 - \frac{\widehat{S}(x+1)}{\widehat{S}(x)},$$

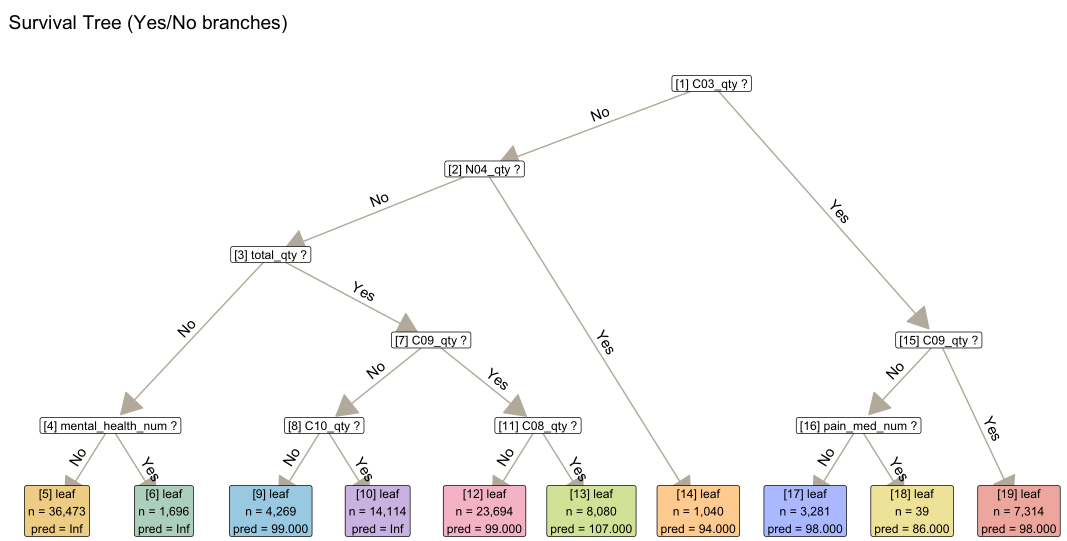


Figure 4.5: Splitting criteria for a survival tree fitted on the indicator dataset. n denotes the expected proportion in each cluster for a sample of 100k observations.

.

This relationship follows directly from the definition of the survival function. $\widehat{S}(x)$ is the probability of surviving beyond age x, and $\widehat{S}(x+1)$ is the probability of surviving beyond age $x+1$. The ratio $\widehat{S}(x+1)/\widehat{S}(x)$ therefore gives the probability of surviving *one more year* given that an individual has already survived to age x. Taking one minus this conditional survival probability gives the complementary probability of *dying within that year*, which is precisely q_x. In other words, q_x measures the proportion of people alive at age x who do not make it to age $x+1$.

4.7.3 Potential Future Improvements to the Model

Several extensions could further enhance the survival tree modelling framework. For example, separate trees could be trained within narrower age bands (e.g., five-year intervals) to examine whether the most influential conditions shift across different stages of later life. This would provide a more granular view of how risk factors evolve with age and could highlight condition-specific impacts that are otherwise masked when modelling the entire population together. These have not been explored within the scope of this thesis due to time constraints.

4.7.4 Indicator Dataset

The indicator dataset was used to train a survival tree for time-to-event (mortality) modelling. Using a 5% split-significance level, the fitted tree produced **10 terminal**

nodes (leaves) (see Figure 4.5). The model incorporated service, prescription, and quantity indicators.

Out of the identified 10 terminal nodes, minimal (almost zero) deaths were observed for leaves 5 and 6. Hence, no curve has been drawn for them. This highlights a broader limitation of the modelling framework: because the tree is trained on a fixed historical observation window and assumes that mortality risk remains constant within each terminal node over that period, it fails to account for changes in future exposure or cohort composition. As a result, terminal nodes with very low event counts are treated as if mortality risk is negligible, even though this may simply reflect insufficient follow-up time or incomplete observation rather than genuinely low risk. This "static snapshot" assumption can therefore lead to misleading conclusions, particularly for groups with long survival horizons or delayed mortality onset. Especially in this case, the mortality was extremely low resulting in extreme issues with pricing.

The fitted survival tree also provides useful insight into which variables have the greatest influence on mortality outcomes. The most prominent splits were driven by prescription quantity variables, particularly `C03_qty`, `N04_qty`, `total_qty`, and `C09_qty`, indicating that medication use patterns and overall prescription volume are strong differentiators of mortality risk. These variables correspond directly to key clinical categories defined in Table 3.2, such as cardiac therapy (C03, C09), Parkinson's treatment (N04), and overall treatment intensity (total prescription volume). Secondary splits involved more specific treatment indicators such as `C10_qty` and `C08_qty`, which map to lipid management and cardiovascular therapy, as well as `pain_med_num` and `mental_health_num`, representing chronic pain and mental health conditions respectively. Importantly, several of these variables; particularly `total_qty`, `C08_qty`, and `C10_qty` are not explicitly included in the final disease dataset though `total_qty` is implicitly included with the other qty variables.

4.7.5 Results

As expected, the modelled survival tree is able to capture more variation in mortality than the unsupervised k-means algorithm (Figure 4.6). This algorithm was able to identify individuals with `C03_qty`, `C09_qty`, and `pain_med_num` to have significantly higher mortality. However, the cluster isolated a small proportion, for 100k individuals there will be only 36 observations satisfying this condition.A reasonable amount of variance was captured through the other clusters.

Figure 4.7 shows the predicted and observed mortality curves for each terminal node of the survival tree, alongside the corresponding ALT mortality curve as a benchmark. The plots are stratified by gender and clearly illustrate how the fitted model captures heterogeneity in mortality patterns across the identified subgroups. While the predicted curves generally align closely with observed experience, there are notable deviations from the ALT, particularly in nodes with either very low mortality (e.g. Node 10) or elevated risk (e.g. Node 13 and Node 18). These differences highlight the value of incorporating health and treatment indicators into survival

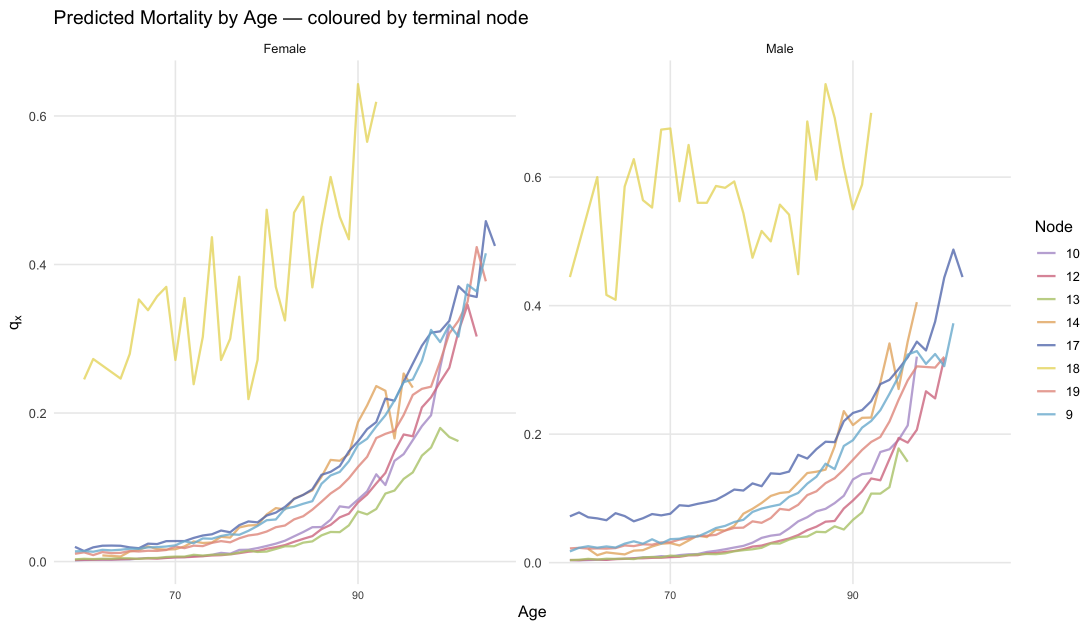


Figure 4.6: Mortality for each gender and cluster, showing the difference in mortality trends captured by the model. Colours match the colours used for nodes in Figure 4.5.

modelling, as they enable the model to capture risk stratification not reflected in standard population-level life tables.

4.7.6 Disease Only Dataset

The disease dataset was used to train a survival tree, the results were significantly more easier to interpret when compared to that training on the entire dataset. The resulting node breakdown can be summarised in Figure 4.8.

Since the survival tree produces a clear sequence of decisions leading to each terminal node, its structure and decision-making process are transparent and unambiguous. The mortality for the selected nodes are plotted in Figure 4.9. The model clearly captures the trend in mortality associated with pain medication usage identified during the data analysis. The model was able to identify that those with antithrombotic conditions, no cardiac conditions and with pain medication have high mortality. However the trend seen between other nodes is not as significant. It is also worth noting that the observations for nodes 11 and 12 fit in the same range (both nodes have lower number of observations when compared to the others)

4.7.7 Results

The survival tree captures substantially more variance than the k-means approach (Section 4.6), particularly through its ability to isolate the influence of pain medication. In this case, the model identifies a subgroup with significantly higher mortality

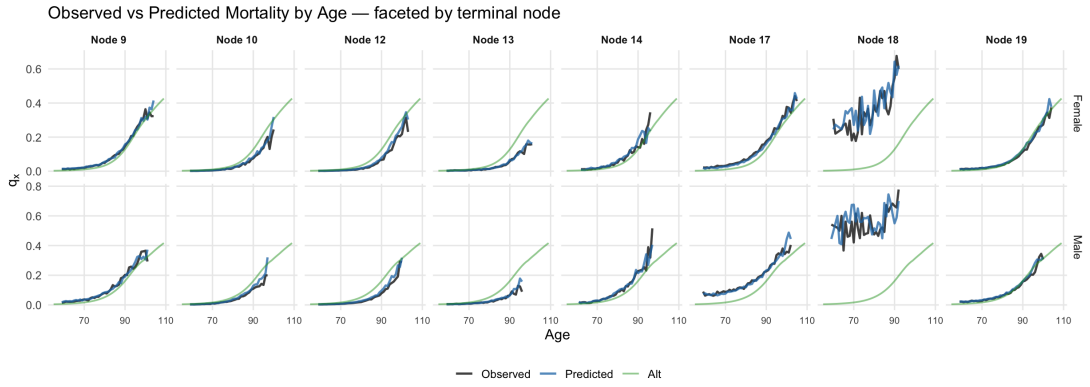


Figure 4.7: Survival tree performance on the indicator dataset compared with ALT as a benchmark.

risk than what was observed when using pain medication alone as a predictor. However, the proportion of individuals falling into this high-risk subgroup remains very small, similar to the survival tree fitted to the indicator dataset (Section 4.7.4). Moreover, the distribution of observations across terminal nodes is highly uneven: nodes 4 and 5 account for the majority of individuals, while most other nodes contain relatively few observations. This concentration limits the granularity of insights that can be drawn from some branches of the tree, despite its improved ability to differentiate risk patterns.

Overall, the predictive performance is strong across most nodes, with observed and predicted mortality closely aligned. The primary exception is node 12, where high volatility in the observed data leads to greater prediction error. The curves presented here represent raw mortality predictions prior to smoothing. Smoothed curves, provided in Chapter 5, illustrate how smoothing improves the stability and predictive power of node 12 in particular.

4.8 Regression

4.8.1 Objective.

After attempting to run a decision tree on the model, it is clear that there is no clear target variable that predicts mortality. Thus, the summarised disease based dataset (Section 3.5) was used to build a regression model. Here, as the dataset has been significantly reduced in dimension to 6 conditions, we were able to use these to calculate the mortality for each combination of the conditions, age and gender. This allowed us to calculate q_x against age and gender for every combination of diseases, allowing for q_x to be used as the target variable for the regression model.

As a result of the high dimensionality of the raw and indicator datasets, calculating mortality based on age, gender, and all combinations of MBS/PBS variables will result in sparse mortality observations, which will further be impacted by the

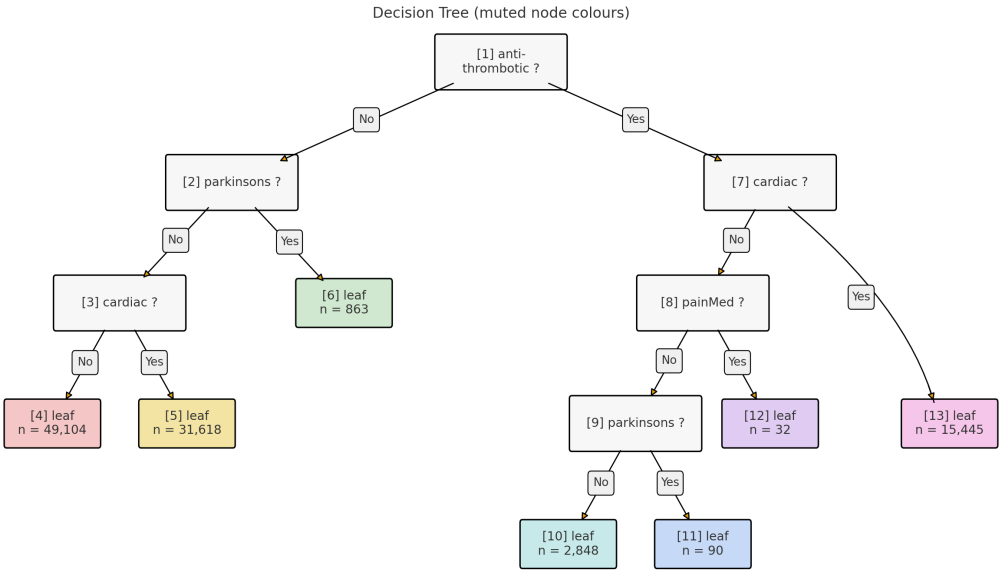


Figure 4.8: The survival tree denoting the conditions leading out to the selected nodes. Here n denotes the number of observations per 100,000 entries.

DataLab constrains (Section 3.1). Hence, this model was not trained using either dataset.

In particular, we use linear regression to model the log mortality, because as discussed in Section 3.7.1, mortality generally increases exponentially with age (the Gompertz–Makeham pattern). We also confirmed that this was better than a regression.

4.8.2 Model Assumptions

1. The log mortality follows a linear trend against the predictor variables used for the model.

2. There is no trend in mortality seen across time. This is a simplifying assumption and should be revisited when a longer tail of data is available.

3. In cases where certain combinations of health status, age, and gender lack observed mortality values, the model estimates mortality through fitted values derived from the regression relationship.

4.8.3 Model

The model fitted used a log transformation due to the exponential slope seen in the mortality curve. The model consists of the following components:

- a natural cubic spline in age (3 degrees of freedom),

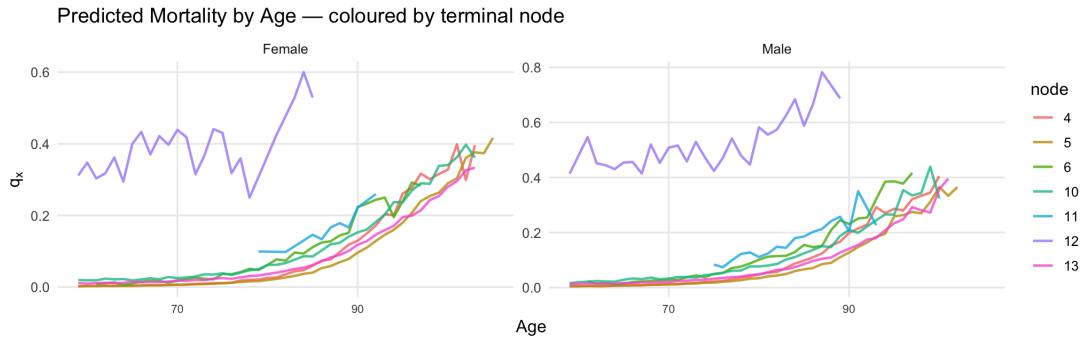


Figure 4.9: The distribution of the mortality for each survival tree leaf (using disease dataset). Using the same colour convention as Figure 4.8.

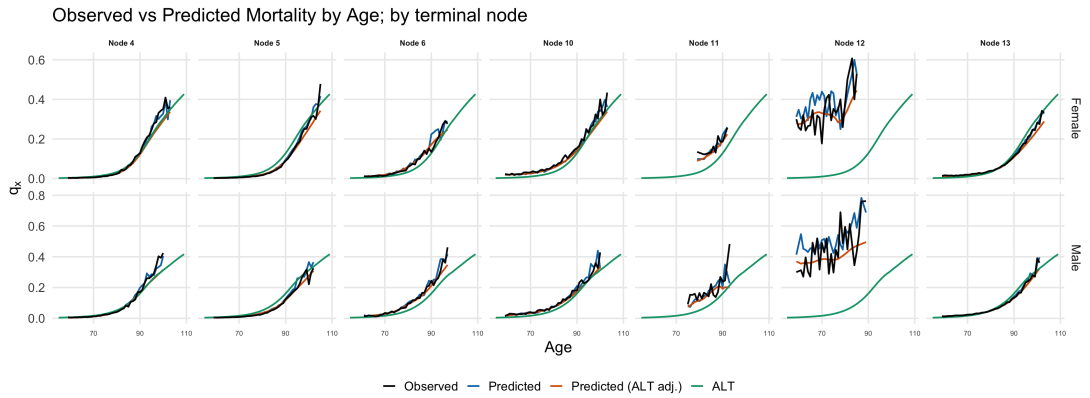


Figure 4.10: Comparison of predictive performance across survival tree nodes on disease dataset for mortality outcomes in the test dataset.

- main effects for the six conditions and gender, and
- interactions between the age spline and each condition (allowing disease specific age curves).

4.8.3.1 Target Variable Creation

The target variable for the training dataset was created by filtering out the observations for each combination of diseases. Using this we calculate the mortality based on the disease combination, gender and age.

4.8.3.2 Formal Description

There are six diseases in the summarised disease dataset, as all of these conditions are indicator variables, this results in 2^6 disease combinations, which we enumerate by $i = 1, 2...n$ ($n = 2^6$). Consider the combination of diseases for each of the groups

as $\mathbf{m}_i = (m_{i1}, \ldots, m_{ip})^\top \in \{0,1\}^p$ (here $p = 6$). For each of these combinations we can find the mortality based on age and gender using the 2011-2013 dataset.

Let us denote the mortality rate of group i at age x and gender $g \in \{0,1\}$ (0=Female, 1=Male) as $q_{x,g,i}$. Then for the ith group, regression model then predicts

$$y_i = \log(q_{x,g,i}). \tag{4.1}$$

We model age for each group x_i as a natural cubic spline with three basis functions, which we will denote $x = z_{ik}$, $k = 1,2,3$. We model interaction terms between these age spline bases and the other disease indicators (except for mental health, as the interactions were identified as insignificant). Let us assume that $p = 6$ denotes mental health, and thus only the remaining conditions were considered for interactions ($p = 1, 2..5$). We also do not model interaction terms against gender.

Then our model is:

$$y_i = \beta_0 + \beta_g\, g + \sum_{p=1}^{5} \beta_m\, m_{ip} + \beta_6\, m_{i6} + \sum_{k=1}^{3} \alpha_k\, z_{ik} + \sum_{k=1}^{3}\sum_{p=1}^{5} \gamma_{pk} z_{ik} m_{ip} + \varepsilon_i \tag{4.2}$$

where β and α are the coefficients of the main effects, and γ are the coefficients of the interaction terms. The error term ε_i is assumed to follow a Gaussian distribution, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, consistent with a standard linear regression framework on $\log(q_{x,g,i})$.

Prediction. Using this model, the log mortality rate is predicted for each entry in the 2016 dataset (used as the test set). After grouping by age and gender, the predicted mortality for each medical combination is compared with the observed mortality. The results are presented in Figure 4.11.

4.8.4 Results

For each combination of diseases, age and gender, the predicted q_x (using the model, trained on 2011-2013 data) against actual q_x (using 2016 data) are shown in Figure 4.11. Here the scatter points close to the main diagonal denote good predictions, while those that are further away denote poor predictions. It is clear that the regression significantly outperforms the ALT. The ALT seems to fail in identifying high mortality values seen in some disease combinations (as seen by the absence of red scatter plots on the top the main diagonal). This is further seen from the metrics shown under Table 5.10.

Furthermore, the performance of the model against each combination of disease has been plotted to ensure that the prediction for a certain combination is not much less than others Figure 4.12. There is no significant under-performance seen in any specific disease combination. However, there is a high number of just pain medication entries with deviation from the actual. This is expected as it shows a significant increase in mortality over the years. Thus, since the model is trained on 2011-2013 data the 2016 data will have a higher mortality than predicted by the training data.

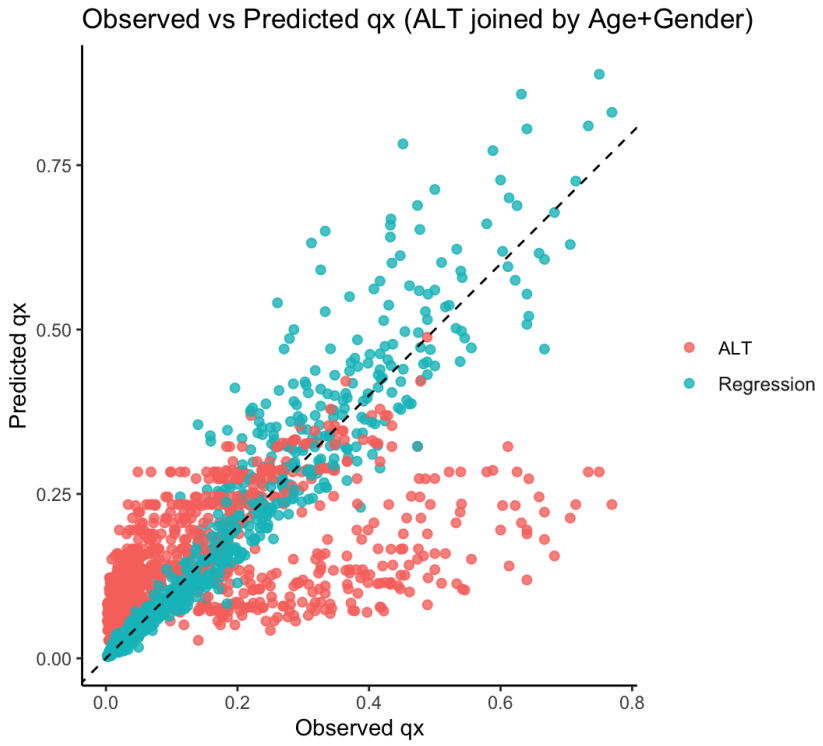


Figure 4.11: The performance of the regression against the performance on the ALT.

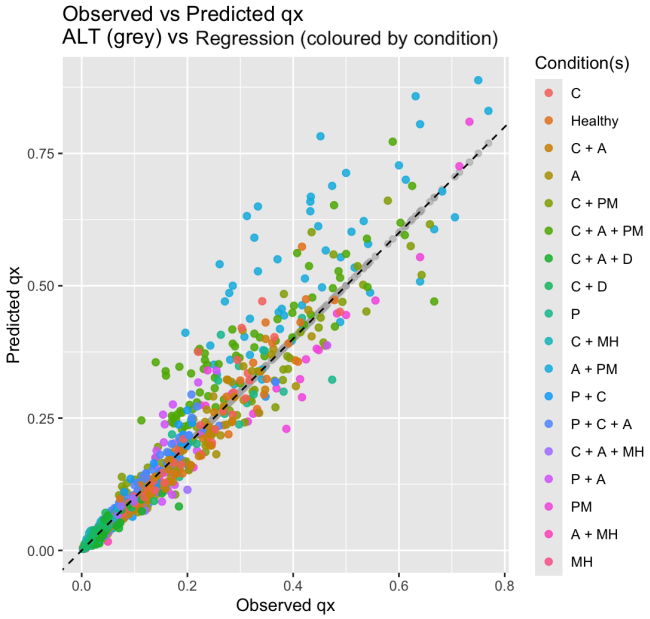
The mortality plots in Figure 4.13 compare observed mortality rates, the ALT predictions, and the regression model predictions across age, gender, and combinations of health conditions. The dashed line represents observed rates, the dotted line the ALT predictions, and the solid line the regression model. Across most disease combinations, the regression closely tracks both the trend and the absolute mortality level along with a good prediction of the trend.

The ALT baseline performs adequately for individuals with no or few conditions but not for higher number of combinations. This allows us to conclude that the average mortality across all entries within an age-gender combination does not capture the mortality changes within the group. Hence, the regression better captures heterogeneity in risk.

Overall, the regression model provides a closer fit to the observed mortality than the ALT baseline across most disease combinations, particularly for groups with higher morbidity. Thus, the regression model better predicts the level and shape of age specific mortality under varying health profiles.

4.9 Markov Process

One key assumption in the above models being used for pricing is that we assume that the health status of an individual does not change from the point of purchase un-



Note: The diseases present are denoted with abbreviations, P:Parkinson's, C: cardiac, A: anti-thrombotic, MH: mental health, D: diabetes, PM: pain medication. Though theoretically there are meant to be 2^6 combinations the number of exported conditions are significantly lower due to export constrains (Section 3.1) and low exposure.

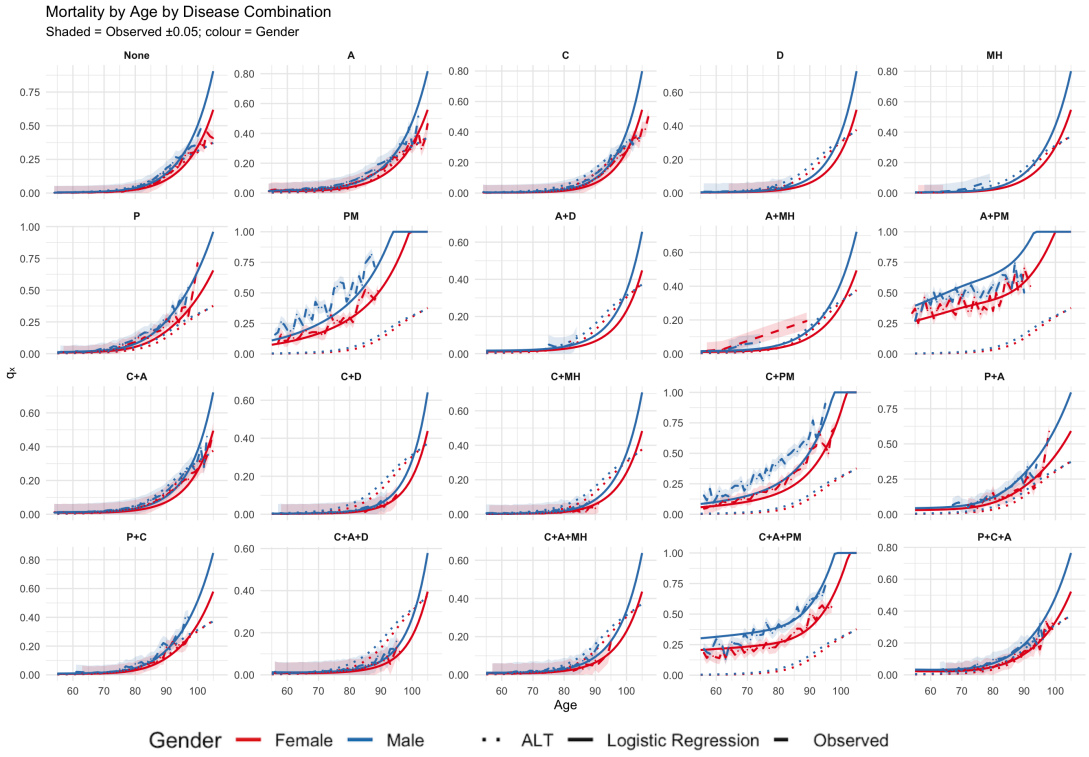
Figure 4.12: The observed vs predicted from the regression, split by disease combination.

til death. This is a strong assumption, as older cohorts' health often deteriorates and state transitions are informative about near-term mortality risk. Given the limited observation window in our data, treating health as static likely understates mortality for initially healthy states and overstates for more unhealthy states, resulting in biased annuity values. Therefore, results from the earlier discussed models should be interpreted as prices under a no-transition approximation rather than full multi-state risk.

4.9.1 Static Model Assumptions

Clustering using k-means (Section 4.6), regression (Section 4.8), decision trees (Section 4.7.1), and survival trees (Section 4.7) are all static models, each making the key assumption that *an individual's health status remains fixed over time*. This is a strong assumption, given that the dataset spans only six years and therefore reflects only the most recent health data for individuals who have died. In contrast, this assumption is not critical in the original ALT, where mortality is modelled solely against age and gender. In ALT, the effect of age is explicitly incorporated into the model. This limitation is not immediately reflected in the overall model performance, as the available data on expected mortality is limited. With such a small observation window, the model cannot fully capture the long-term effects of changing health statuses, even if short-term performance metrics appear acceptable.

Markov chain (Section 4.9) is a dynamic model accounting for the transitions



P: Parkinson's, A:Anti-thrombotic, C: Cardiac, D: Diabetes, MH: Mental Health conditions and PM: pain medication

Figure 4.13: Mortality against disease combinations, for the regression model, ALT and the 2016 observed values. The hue around the observed denote a ± 0.05 change in the mortality.

between certain disease types. Thus, the above limitation does not apply to Markov chains model.

4.9.2 Markov Chain Model Overview

The Markov chains model was built with the main disease combinations using the summarised disease variables from the MBS and PBS dataset (Section 3.5). The model does not consider death as a state as the transition probability will depend on gender and age and thus making the transition probabilities more complicated, and not applicable to other models.
Diseases were selected by filtering combinations that had more than 450,000 recorded observations. This is a reasonable lower bound for the most common diseases, given the dataset had over 33 million observations. This reduced the number of clusters to a reasonable amount, resulting in the following disease combinations:

1. Healthy (none of the following; diabetes, mental health, anti thrombotic, cardiac, Parkinson's)

2. Diabetes only

3. Mental health conditions only
4. Anti-thrombotic conditions only
5. Cardiac conditions only
6. Cardiac conditions with diabetes
7. Cardiac conditions with mental health conditions
8. Cardiac conditions with anti-thrombotic conditions
9. Cardiac conditions with anti-thrombotic conditions and diabetes
10. Cardiac conditions with anti-thrombotic conditions and mental health conditions
11. Parkinson's disease
12. Parkinson's disease with cardiac conditions
13. Parkinson's disease with cardiac and anti-thrombotic conditions
14. Pain medication with any condition
15. Other conditions

Note that all the provided states are mutually independent and there is no ambiguity on which state an individual belongs to.

The transition probabilities for the model have been calculated, firstly using the entire population and then split by gender (provided in the Appendix, Appendix B). To analyse the impact for male and female, the difference in transition probabilities have been calculated.

4.9.2.1 Transition Probabilities

The transition probabilities between states for the 15 identified health states are shown in Figure 4.14. This shows that for most states, individuals are most likely to remain in the same state over the course of a year. However, mental health (MH) is an exception: for example, the probability of transitioning from a MH condition to a healthy state (57.4%). exceeds the probability of remaining in the MH state (33.9%). This suggests that MH conditions are more likely to resolve within a year compared to other disease states. It is important to note that the transitions where they remain in the same state remain to be in the 60-75% range, thus indicating that though most common a significant portion transitions over a year. Further, these transition probabilities make a simplifying assumption that transitions do not depend on the age or gender of the individuals. Therefore, transitions were analysed by age and gender to identify whether these factors had a significant impact.

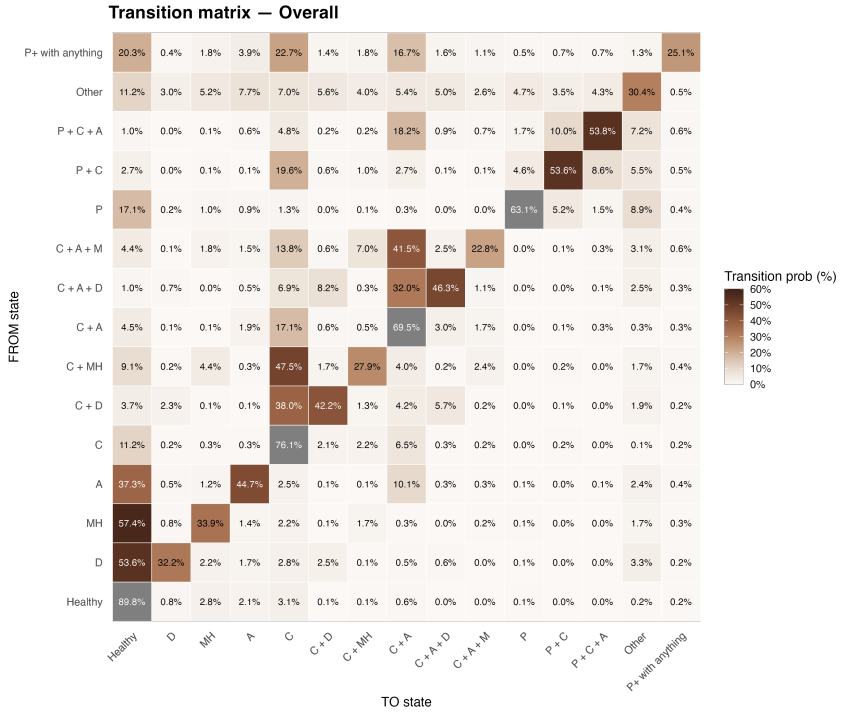


Figure 4.14: Transition probabilities for the entire population from one state to another within a year.

Gender Separated Transitions. The one-year transition probabilities were analysed by gender (Figure B.3 and Figure B.2 visualise the transition probabilities for females and males respectively). The differences in transition probabilities by gender (Figure 4.15) shows that the probability to return to a Healthy state from another state and remain Healthy was higher for females. For males, the most significant transitions indicate that their probability of recovery from diabetes, Parkinson's, and anti thrombotic conditions is significantly lower compared to females. Other than these deviations, most transitions appeared similar. This analysis is based on indicator variables and may not perfectly correspond to clinical variables.

Age Bin Separated Transitions. Age bins of 50-70, 70-90 and 90+ were considered. Transition probabilities for each bin are provided in Figures 6.1–6.3. To understand the changes observed between these age groups, the differences in transition probabilities are shown in Figure 4.16. Between ages 50–70 and 70–90, transitions back to Healthy decline markedly, while disease persistence and co-morbidity increase. Returns to Healthy from single-condition states become notably rarer, and the Healthy state itself becomes less stable. Individuals are more likely to remain within an existing condition. Transitions from single cardio conditions toward combined cardio–metabolic states also increases.

From 70–90 to 90+, these patterns intensify. Healthy persistence declines further,

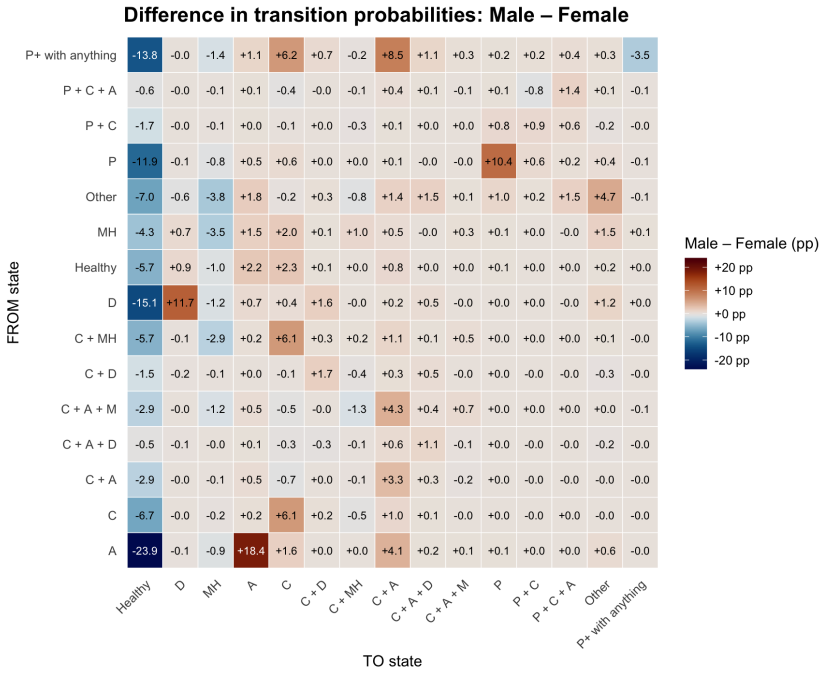


Figure 4.15: The transition probabilities denote the difference between male and female transition probabilities each from-to state combination. Warmer tones: transitions made more prominently by males, cooler tones: transitions made more prominently by females.

recovery from disease states becomes uncommon, and persistence in cardio/anti-thrombotic conditions strengthens. Flows into co-morbidity combinations continue to grow, so by 90+ the dynamics are dominated by staying diseased and accumulating cardio-adjacent conditions rather than reverting to Healthy.

Significance Testing. To understand whether patterns of movement between health states differed by age or gender, we compared how people transitioned out of each starting state across these groups. We did this by looking at the distribution of all possible next states and checking whether these distributions were statistically different. Using this two p-tests were carried checking for statistical significance between transitions spilt by gender and between transitions split by age bins.

For gender, there were no meaningful differences: after accounting for multiple testing, none of the comparisons were statistically significant and the effect sizes were very small. This suggests that, once a person's starting state is considered, men and women transition to other states in broadly similar ways over a one–year period.

For age, however, the results were quite different. All comparisons showed clear differences between age groups, and these differences were statistically significant even after adjustment. Although the overall effect sizes were modest, they were still meaningful. The biggest differences were seen for transitions involving *P+ with anything*, *Other*, and *A* states, while differences were less noticeable for *MH* and more complex multi–morbidity states like *C+A+D*. Overall, this indicates that age plays a

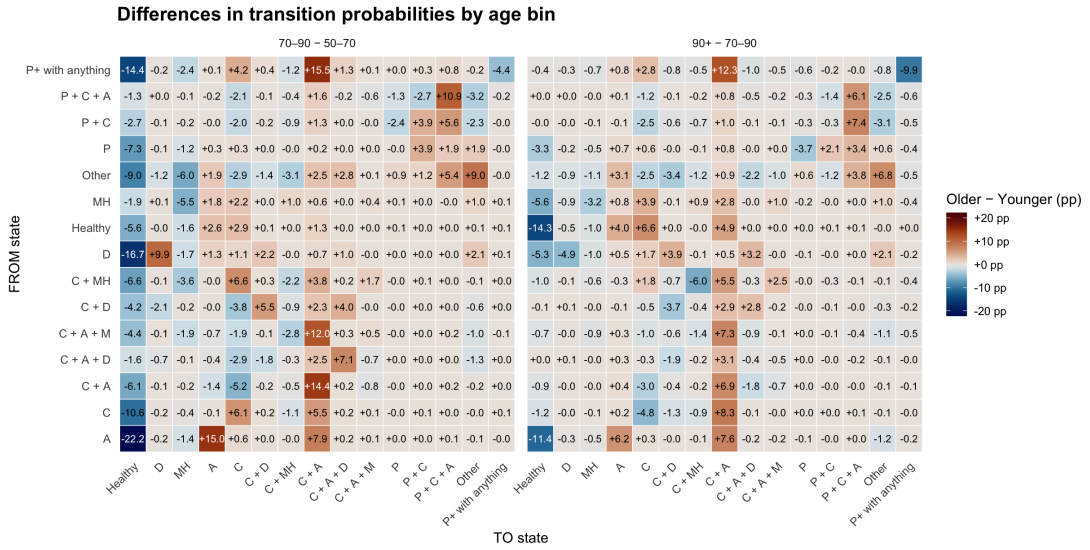


Figure 4.16: Transition probability differences between age bins. Left: the difference in transition probabilities between 50-70 and 70-90, Right: the difference in transition probabilities between 70-90 and 90+. Warmer tones: transitions more prominent among the older age bin, cooler tones: transitions more prominent among the younger bin.

key role in shaping health state transitions, even if the size of these differences is not always large. In contrast, gender does not appear to have a significant impact on these transition patterns.

Conclusion. Age-based transitions will be incorporated into the final model, as age clearly influences how individuals move between health states. In contrast, gender based transition splits will be excluded from the model due to its lack of significant impacts.

4.9.3 Model Results.

The definition of the states were done using most common medical condition combinations, thus is equivalent to an unsupervised clustering approach. The resultant clusters showed different trends in mortality across the training dataset of 2011 (Figure 4.17).

The results of the model trained using 2011-2013 data were used to predict 2016 mortality for each subgroup. The model predictions together with the corresponding ALT prediction for each subgroup is shown in Figure 4.18. The hue seen in the plot denotes a 5% window from the actual value. It is clear that the Markov chain model predictions fall under the provided window for all conditions.

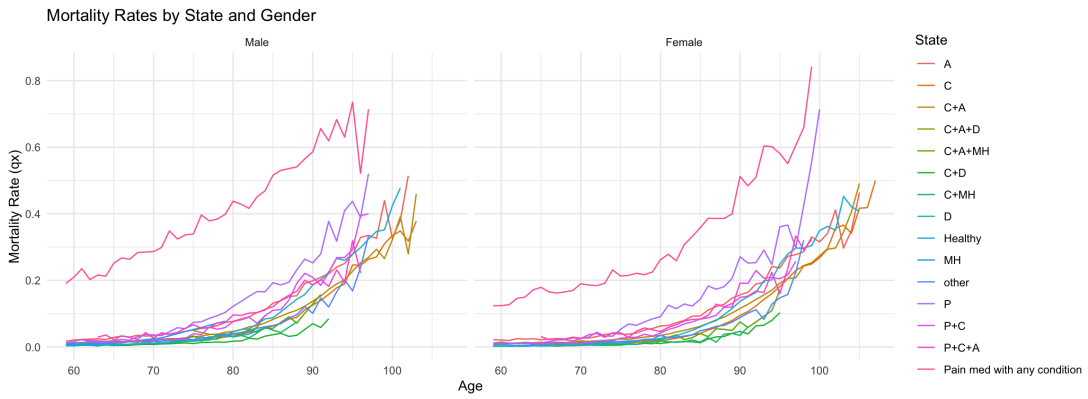


Figure 4.17: Plot of the mortality for each gender, split by state.
Note: The diseases present are denoted with abbreviations, P:Parkinson's, C: cardiac, A: anti-thrombotic, MH: mental health, D: diabetes.

4.9.3.1 Graduation for Sparse Cells.

The condition combinations highlighted in Section 4.9.2 are the most common in the dataset; hence any censoring in q_x at younger ages is primarily due to the lack of deaths rather than low exposure (see Section 3.1. To avoid downward bias resulting from zero counts, we graduate the fitted curves for *C+A+D*, *MH*, and *D* (Figure 4.19).

For these conditions, deaths occur after transitioning out of the state. As a guide, if a state records at most 10 deaths over ages $60-110$ with a total exposure of $\approx$ 45,000 person–years, the average exposure per age is $45{,}000/(110-60) \approx 900$, yielding an expected average mortality of

$$\frac{10}{900} \approx 1.1\%\,.$$

We therefore anchor the curve near 1% at mid–late working ages and impose an exponential increase with age on the logit scale,

$$\operatorname{logit}\big(\hat{q}_x\big) = \alpha_s + \beta_s\,(x - x^*),$$

where x^* is an anchor age (we use $x^* = 75$) and $\alpha_s = \operatorname{logit}(0.01)$ to encode the 1% anchor. The slope β_s is obtained by shrinking the ALT slope β_{ALT} (estimated from $\operatorname{logit}(q_x^{\mathrm{ALT}})$ over ages 60–95) toward the sparse empirical slope: we choose $\beta_s = \kappa\,\beta_{\mathrm{ALT}}$ with κ selected to minimise squared error on the observed ages (grid search). The resulting $\hat{q}_x = \operatorname{logit}^{-1}(\alpha_s + \beta_s(x - x^*))$ is then truncated to $(0,1)$ and made non–decreasing in x. This graduation is applied only to *C+A+D*, *MH* and *D*; all other states use the un-smoothed estimates.

Detailed Approach. For each gender–state pair we:

1. build a complete age grid $x \in \{\,\min\ \text{age}, \ldots, 110\,\}$;

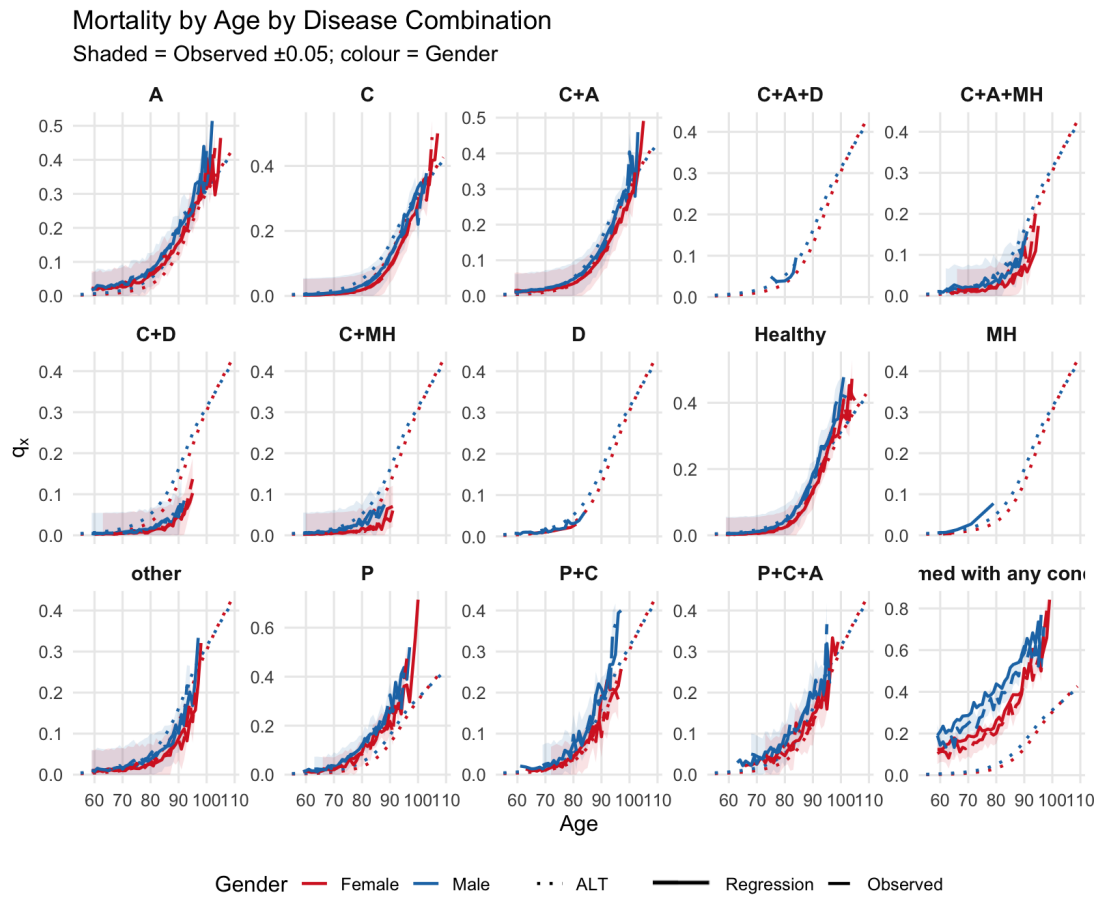


Figure 4.18: Plot against mortality prediction from the proposed Markov chain model and the ALT for each node.

2. clamp reported values to $(\varepsilon, 1-\varepsilon)$ with $\varepsilon = 10^{-6}$;
3. enforce non–decreasing mortality with age via the running maximum, i.e. $q_x \leftarrow \max(q_{x-1}, q_x)$.

We use a hybrid approach:

- **Anchored logit graduation** for sparse states *C+A+D*, *MH* and *D*;
- **Normal graduation** (logit–spline) for all remaining states.

Anchored logit graduation, explained under Section 4.9.3.1 was performed for all states with sparse observations. For non sparse states, each gender-state series is smoothened on the logit scale, $y_x = \text{logit}(q_x)$, using the simplest method that the data allow:

1. ≥ 3 ages: fit a shape-preserving spline $f(x)$ to y_x; if that fails, use linear interpolation.

2. exactly 2 ages: linearly interpolate y_x.
3. ≤ 1 age: use the ALT curve for that gender; if there is one point, rescale ALT to match it; if none, use ALT as is.

Finally, set $\hat{q}_x = \text{logit}^{-1}(f(x))$, clip to $(\varepsilon, 1-\varepsilon)$, enforce non-decreasing values with age (running maximum), and extend the series to age 110.

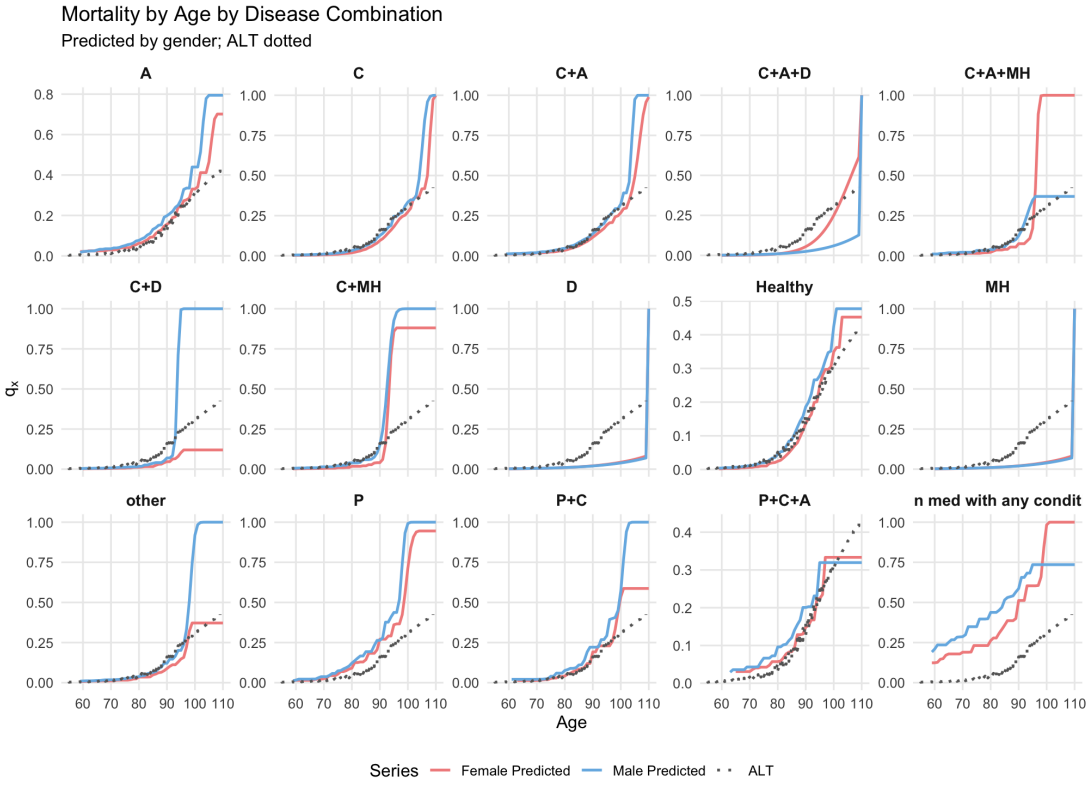


Figure 4.19: Smoothened mortality predictions for each of the Markov states (along with the ALT predictions).

4.9.3.2 Life Expectancy and Impact of Transitions

The life expectancy of a 60 year old in each of the states were calculated with and without allowing for transitions to understand the impact of the transitions on the overall model. The calculation of life expectancy deviates significantly with the transitions, thus implying the static assumption (Section 4.9.1) is a strong assumption for modelling health-based mortality. Further more, the noticed trend with pain medication (shown in Section 3.7.1.4) is not as prominent when considering transitions to and from the state.

Interpretation. Figure 4.20 (see Section B.4.1 for raw values) compares remaining life expectancy at age 65 across disease states, split by gender, under four mortali-

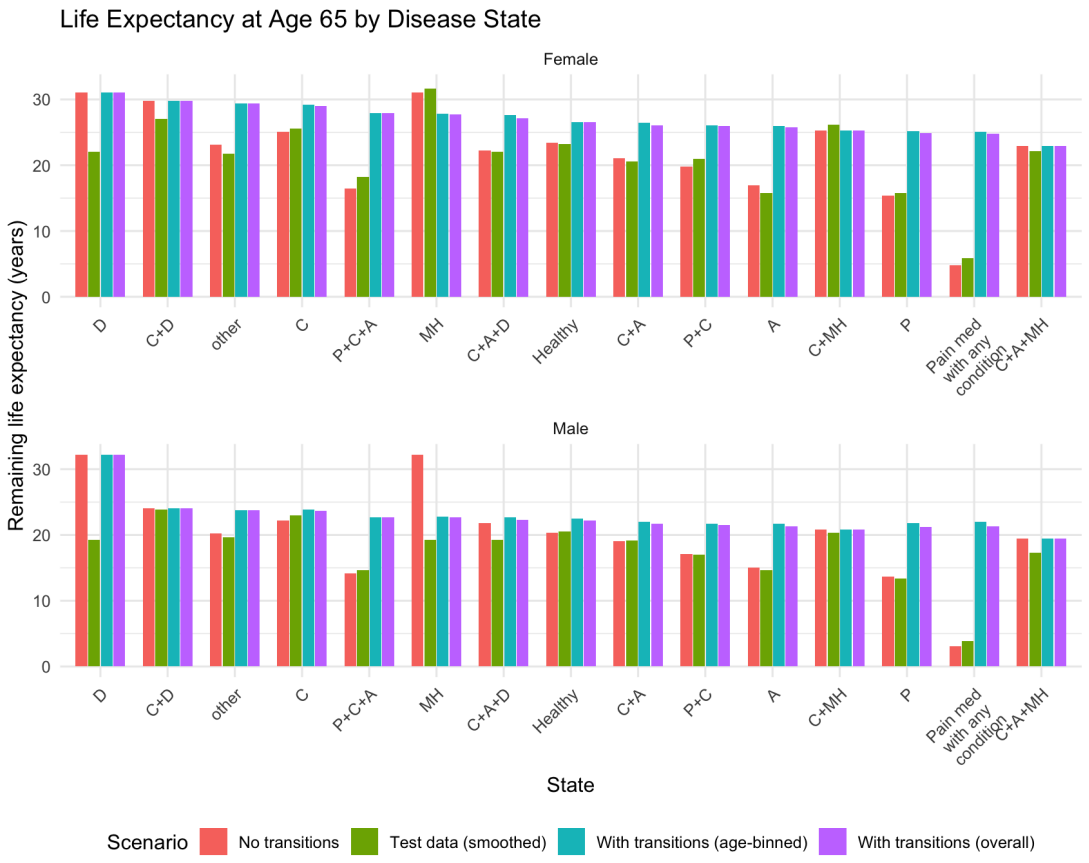


Figure 4.20: Life expectancy without transitions and with transitions considering both the age binned transitions and the complete population transitions.

ty/transition assumptions:

1. *No transitions* (identity transition matrix): individuals are assumed to remain in their starting state. This serves as a baseline that ignores progression or recovery.

2. *With transitions (overall)*: a single transition matrix estimated from the full cohort is applied at all ages.

3. *With transitions (age–binned)*: transition matrices vary based on age bins 50-70, 70-90, and above 90, allowing state movement to reflect age patterns observed in the data.

4. *Test data (smoothed)*: life expectancy computed directly from the smoothed observed q_x for the test period (empirical benchmark).

Across nearly all states, females have higher life expectancy than males, consistent with the trend seen in overall population mortality. The mortality for states with

transitions show less deviation, which is expected as the short-tail training data implies a heavy bias to near death mortality expectations. Introducing transitions typically *reduces* life expectancy relative to no transitions because some members move into worse states over time. However, the opposite trend is observed for patients with diabetes and mental health conditions only. This is due to the low mortality associated with these conditions. As a result, transitions out of these states (e.g., into more severe or co-morbid conditions) are associated with worse life expectancy. The age–binned specification tracks the empirical benchmark more closely than the single overall matrix, highlighting the strong age dependence of health state movements. However, the impact of these age-binned variations is expected to be more significant for shorter tail predictions, which are more relevant for pricing products such as term insurance.

4.10 Model limitations

1. Like all models, the proposed Markov chain model had numerous simplifying assumptions. This includes:
    - All individuals' age is at the reporting date, thus age change between the reporting date and age of death has not been accounted for. This assumption is made due to the limitations in the environment, requiring us to remove all date and month variables from the dataset.
    - Deaths occur at the end of the year. (Additionally, in the case of a Markov chain model, deaths occur prior to transitions)
2. The dataset used for this study only consists of six years of data, out of which we need to account for a lag of three years (Section 4.1.2). The lack of a longer tail of information introduces the following limitations:
    (a) Limited capability to capture medical conditions with long-term effects on mortality, as diagnoses or treatments occurring prior to the observation window are not recorded.
    (b) Reduced ability to identify delayed impacts of health interventions, since longer follow-up periods would be required to observe their full influence on outcomes.
    (c) Potential underestimation of transition probabilities between health states, as rare events may not appear frequently enough within the restricted time frame.
    (d) Performance testing was done only on mortality rate (q_x), while ideally to account for the improvement seen in mortality due to transition probabilities we should also test on expected lifetime, but this is not possible due to the lack of data. This limitation should be revisited when more data is available as the performance in a longer time frame will allow us to test the transition probabilities calculated under Section 4.9 and the long term

performance of the mortality table, which will likely cause Markov chain to significantly outperform other static models.

(e) Another major limitation of using supervised models on death-based variables is that the training data only captures near-death indicators. With a longer observation period, we could incorporate health status histories well before death, providing a broader view of the factors influencing mortality. However, our current data only includes at most three years prior to death, which is insufficient to capture the full range of potential mortality determinants.

4.11 Conclusion

This chapter goes over the attempted models, providing outputs, and analysing the captured deviations and performances against the benchmark ALT. The models discussed in this chapter and their key characteristics have been summarised under Table 4.1. A key takeaway is that due to the short-tail nature of training data, the mortality calculations primarily reflect near-death health statuses. This limits the model's ability to capture the true dynamic nature of individual health statuses over longer periods. There is a significant impact of these transitions. Chapter 5, will formalise this analysis in the form of metrics to allow us to better select the final model.

Chapter 5

Model Performance

This chapter examines model performance and evaluates the monetary implications of this study for individuals entering retirement. It first introduces the proposed metrics, providing both formal definitions and interpretations in Section 5.1. Section 5.2 then describes how raw probabilities are smoothed to align with the ALT for use in life expectancy calculations. Section 5.3 explains how life expectancy is calculated, as it is required for metric calculations and converts mortality findings into a more interpretable measure. Section 5.4 presents the performance of the models introduced in Chapter 4, which are further standardised for comparison in Section 5.5.

Note that this chapter includes calculations involving financial terminology, see the Financial Glossary for definitions.

5.1 Proposed Metrics

As mentioned under Section 4.1, there are two main aspects the model needs to be tested on:

1. Accuracy, measuring how well mortality is predicted for a given subset.
2. Deviation, assessing how well the modelled groups (subsets) capture differences in mortality.

5.1.1 Accuracy Metric

To evaluate model accuracy, model results using the 2011–2013 period are compared against observed outcomes from the 2016 data for each input combination. This comparison indicates how accurately the model predicts observed outcomes. Several measures can be used for this evaluation, including mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), bias (mean error), calibration, and R^2. These are defined in Section 5.1.1.1.

All of these metrics provide insight into model performance and were reported for each model, however RMSE (with a weight adjustment) was selected as the primary accuracy metric for this thesis. The rationale for this choice is discussed further in Section 5.1.1.3.

5.1.1.1 Metric Definitions

Let y_i denote the observed mortality (or event rate) and $\hat{y}_i$ the model prediction, for $i = 1, \ldots, n$.

MSE. The mean square error defined as

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

is a measure of how much the model's predictions deviate from the observed values. Squaring the deviations penalises larger errors more heavily than the mean absolute error.

RMSE. Root Mean Squared Error, defined as $\text{RMSE} = \sqrt{\text{MSE}}$, expresses this measure on the same scale as the data, making it more interpretable. It represents the average deviation of predicted mortality from the observed mortality.

MAE. Mean Absolute Error defined as

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

does not penalise large deviations as strongly as the MSE and is therefore more robust to outliers.

Bias (Mean Error).

$$\text{Bias} = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)$$

Unlike MSE and MAE, the bias metric retains the direction of deviation, allowing identification of whether predictions are systematically too high (positive bias) or too low (negative bias).

R^2.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Here $\bar{y} = \sum_{i=1}^{n} y_i$ is the mean of the observed values. The R^2 metric represents the proportion of variance in the observed target explained by the model's predictions. It ranges up to 1, with higher values indicating better fit. In this context, it measures how well the model captures the pattern of observed mortality across cells or groups.

5.1.1.2 Metric Interpretation

A concise summary of each metric's interpretation, along with its advantages and limitations, is presented in Table 5.1. Guided by this comparison, the choice of the primary evaluation metric is justified in Section 5.1.1.3.

Table 5.1: Accuracy and calibration metrics with interpretation, pros and cons. The selected accuracy metric is provided in bold.

Metric	**Better**	**Interpretation (assume metric $= x$)**	**Pros**	**Cons**
MSE	Lower	Average squared deviation per prediction from the actual value	Convex; heavily penalises large errors	Squared units hard to interpret; highly sensitive to outliers
RMSE	**Lower**	**Average deviation per prediction from the actual value**	**Same units as target; highlights big misses; widely understood**.	**Outlier-sensitive; hides error direction; not additively interpretable.**
MAE	Lower	Average absolute error per prediction from the actual.	Robust to outliers; median-optimal loss.	Less sensitive to rare large errors
Bias (Mean Error)	Close to 0	On average, predictions exceed the actual by x when $x > 0$, and fall short of the actual by $\|x\|$ when $x < 0$.	Direct read on systematic level error; easy to correct via intercept.	Can be ≈ 0 with poor accuracy because negative and positive errors cancel.
R^2	Higher	About x% of variation in observed mortality across cells is explained.	Scale-free summary of pattern capture; good for ranking models.	Can be negative for poor fits; depends on variance in y; not a pure accuracy metric.

Notes. (1) No consideration has been put into the differentiability of the metric and this is not relevant to this research.

5.1.1.3 Selected Metric

All metrics discussed under Section 5.1.1.1 provide different intuition about the performance of the model, and are therefore reported for each model. But for the purpose of this model, the priority lies in getting a good estimate of the mortality and

we want to ensure that higher deviations from expected are more penalised as they will cause amplified inadequacies when pricing. This allows us to ensure that all the errors seen within the data are small and thus the pricing for all clusters/ disease combinations are as fair as possible. Hence, for as the main metric of accuracy we use the root mean square error.

Additionally, not all mortality predictions have equal impact on pricing or industry applications, as exposure varies by age. Thus, the selected metric is a weighted RMSE which accounts for the impact of differing exposure levels across ages, denoted wRMSE.

Formalisation of RMSE. As predictions are made for each age and gender within each cluster, define the index sets $\mathcal{C}$ (clusters), $\mathcal{G}$ (genders), and $\mathcal{X}$ (ages). The (unweighted) RMSE is

$$\text{RMSE} = \sqrt{\frac{\sum_{c\in\mathcal{C}}\sum_{g\in\mathcal{G}}\sum_{x\in\mathcal{X}}\left(q^{\text{pred}}_{x,g,c} - q^{\text{obs}}_{x,g,c}\right)^2}{|\mathcal{C}\times\mathcal{G}\times\mathcal{X}|}}$$

where $|\mathcal{C}\times\mathcal{G}\times\mathcal{X}|$ is the size of the Cartesian product denoting the number of prediction points.

A small adjustment is applied to account for the non-uniform distribution of observations across age bins. Because pricing impact depends on exposure at each age, incorporating exposure-based weights further improves the metric's relevance.

Exposure-weighted variant Let us assume that the count of observations in a given age x in the training set is n_x thus the proportion for each age is $\frac{n_x}{\sum_x n_x}$

Then, the proposed metric is,

$$\text{wRMSE} = \sqrt{\frac{\sum_{c\in\mathcal{C}}\sum_{g\in\mathcal{G}}\sum_{x\in\mathcal{X}}\frac{n_x}{\sum_x n_x}\left(q^{\text{pred}}_{x,g,c} - q^{\text{obs}}_{x,g,c}\right)^2}{|\mathcal{C}\times\mathcal{G}\times\mathcal{X}|}}$$

Interpretation. The RMSE (and its weighted version wRMSE) summarises the model's average deviation for each prediction. For example, a wRMSE $= 0.3\%$ indicates that, on average, predicted q_x values deviate by about 0.3 percentage points from observed values. Because the weight $\frac{n_x}{\sum_x n_x}$ emphasises ages with more observations, wRMSE reflects error where pricing impact is greatest and de-emphasises very sparse ages where noise is high. Two models can be compared directly since the weights sum to 1.

This measure captures overall accuracy and penalises large deviations (squared errors), especially at ages with higher weight. However, it does not tell you direction of error (whether the model over or under predicts).

Age weights The weights used in the model are calculated using the entire 2016 dataset. These weights are based solely on age, without any split by gender or cluster. Consequently, the same set of weights is applied across all models shown in Figure 5.1.

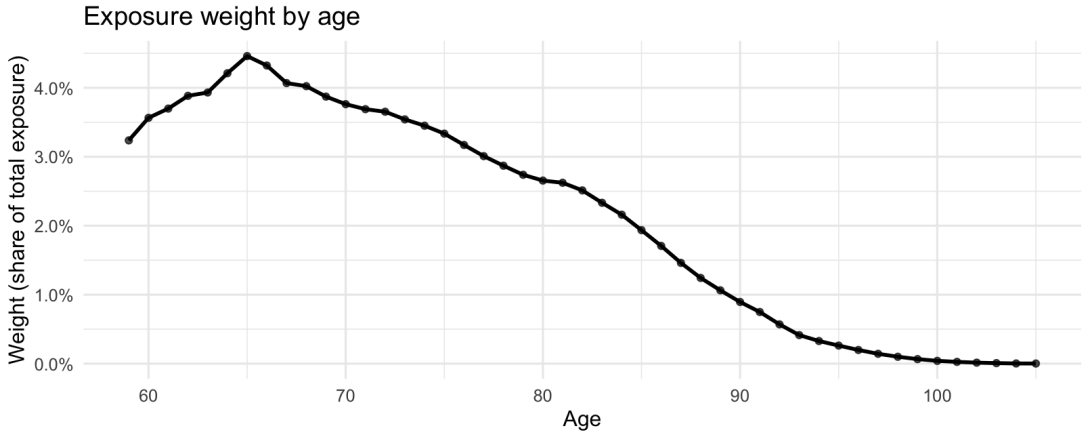


Figure 5.1: Proportion of observations by age in the 2016 dataset, used as model weights.

5.1.1.4 Log-Scale RMSE Metric

Because mortality rates typically follow an exponential growth pattern with age (as discussed in Section 3.7.1), evaluating prediction errors on the original scale can sometimes mask meaningful differences, especially at older ages where absolute mortality levels are much higher. To address this, we also calculate the root mean squared error on the logarithm of the mortality rate, $\log(q_x)$. I.e.,

$$RMSE_{log} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(log(y_i) - log(\hat{y}_i))^2}$$

The log transformation has two main benefits. First, it focuses on *relative* rather than absolute errors, evaluating the model by percentage deviation instead of raw difference. Consequently, proportional errors are treated equally regardless of whether they occur at younger (low-mortality) or older (high-mortality) ages. Second, it reduces the influence of very large errors at older ages, allowing performance to be assessed more evenly across the entire age spectrum. In this way, the log-scale RMSE complements the standard RMSE by providing additional insight into how well the model captures the overall shape and growth rate of the mortality curve.

It is worth noting that this relative error approach is also implicitly incorporated into the selected metric (wRMSE) discussed in Section 5.1.1.3. However, the exposure-weighted RMSE (wRMSE) is preferred overall because it accounts for the fact that some disease combinations have observations concentrated at older ages while others begin at much younger ages. By weighting errors according to exposure, the selected metric better reflects the real-world pricing impact across these

heterogeneous groups. Moreover, because the weights are based on the exposure of the entire population rather than a specific subgroup, they inherently capture the demographic structure and the relative importance of each age in the portfolio. This ensures that the evaluation aligns with practical pricing considerations, where errors at ages with high exposure have disproportionately greater financial consequences. Nevertheless, the log-scale RMSE is reported for all models for comparison purposes.

5.1.2 Deviation Metric (Life–Expectancy Spread)

We summarise how well the model captures *between-cluster* differences using the normalised squared difference between predicted life expectancy at age 65 for each cluster and the ALT predicted life expectancy at age 65. Let $e_{65,c,g}$ be the model's life expectancy at 65 for cluster c and gender $g \in \{\mathrm{F}, \mathrm{M}\}$, computed from the model's adjusted rates q_x^* (Section 5.2), and $e_{65,g}^{\mathrm{ALT}}$ be the ALT life expectancy at 65 for the same gender. As no cluster is considerably small, no weight has been added to this metric.

$$\boxed{\mathrm{Metric}_{\mathrm{Dev}} \;=\; \frac{\sum_{c,g} \left(e_{65,c,g} - e_{65,g}^{\mathrm{ALT}}\right)^2}{|\mathcal{C} \times \mathcal{G}|}}$$

where:

$\mathcal{C}$ = set of clusters, $\mathcal{G}$ = set of genders

$e_{65,c}^{\mathrm{pred}}$ = predicted life expectancy at 65 for cluster c

e_{65}^{ALT} = ALT life expectancy at 65 (baseline)

Interpretation: measured in *years*^{2}; 0 means the model reproduces ALT at 65 in every cluster. Larger values indicate greater separation from the ALT baseline. This is equivalent to using MSE on the life expectancy, but unlike the accuracy metric where we intended to reduce the value, we intend on maximising the deviation metric.

Notes.

- We use the *predicted* e_{65} (rather than the directly observed life expectancy) to avoid end–of–life bias arising from the short three–year observation window (2011–2013). Because many individuals in the dataset are still alive beyond the observation period, their actual lifetime is censored, which would lead to systematic underestimation of life expectancy if we relied on observed data alone. Using model–predicted e_{65} allows us to extrapolate the mortality curve beyond the study window and obtain an unbiased estimate of expected lifetime at age 65.

- We will also report square–rooted version ($\sqrt{\mathrm{Metric}_{\mathrm{Dev}}}$) in *years* for interpretability.

- Model selection is performed to minimise accuracy error (RMSE) while achieving sufficient separation.

5.2 Adjustments from Crude Annual Mortality

Our model outputs *crude annual death probabilities* q_x (for age $x \geq 60$), calculated using Eq. (3.1). However, raw probabilities can be noisy and may not align directly with official life tables. To make them comparable to the ALT, we process the raw data in two main steps using the same approach followed by the ALT [Australian Government Actuary, 2014]:

1. **Smooth the raw probabilities** to remove random fluctuations.
2. **Apply the ALT adjustment** so that the probabilities follow the same conventions as the official life tables.

The result is a set of adjusted probabilities $\mathbf{q}_x$ that can be used to build standard life tables and compute measures such as life expectancy.

Step 1: Smoothing. At older ages, data can be very noisy because fewer people remain alive, so random variation dominates. We smooth the raw probabilities q_x on the *logit* scale:

$$\text{logit}(q_x) = \log \frac{q_x}{1 - q_x}, \qquad \widehat{q}_x = \text{logit}^{-1}\{s(x)\},$$

where $s(x)$ is a smooth curve fitted through the noisy data points and $\widehat{q}_x$ denotes the smoothened values. This keeps the overall trend but removes sudden jumps caused by random variations.

Step 2: ALT Adjustment.

The ALT use a specific conversion formula [Australian Government Actuary, 2014] to transform mortality measures into a form suitable for life expectancy calculations. This process begins with the *central rate of mortality*, denoted m_x, which measures the intensity of deaths within a given age interval and is defined as:

$$m_x = \frac{D_x}{E_x}$$

where:

- D_x is the number of deaths between ages x and $x + 1$,
- E_x is the total person–years of exposure lived by the population during that interval.

The central rate is linked to the probability of death q_x (the probability of dying before reaching age $x + 1$) through the standard life table assumption of a uniform distribution of deaths:

$$q_x = \frac{m_x}{1 + \frac{1}{2}m_x}, \qquad m_x = \frac{q_x}{1 - \frac{1}{2}q_x}.$$

The ALT use this relationship to convert central rates into annual death probabilities using the following adjustment formula:

$$q_x^* = \frac{\widehat{m}_x \left(1 - \frac{1}{12}\frac{q_{x-1}}{p_{x-1}}\right)}{1 + \frac{5}{12}\widehat{m}_x}, \qquad p_x^* = 1 - q_x^*$$

where p_x^* denotes the probability of survival. This adjustment ensures that life tables constructed from central mortality rates follow the same assumptions about how deaths are distributed within each age interval.

In our modelling framework, the model directly estimates q_x (the annual probability of death) rather than starting from the central mortality rate m_x. Since the ALT adjustment is intended to convert m_x into q_x, **it is unnecessary in this context**. Our model already produces mortality probabilities on the correct scale, fully compatible with life table construction and life expectancy calculation without any additional transformation. This not only simplifies the process but also avoids potential numerical approximation errors introduced by conversion.

Step 3: Life Table Construction Here we calculate the life expectancy using the smoothed probabilities. This is only be performed for the final output model, and the resultant life tables are provided in Appendix A. We use the adjusted probabilities q_x^* to build a standard life table. Starting with a population of $l_{60} = 100{,}000$ (people at age 60), we calculate:

$$d_x = l_x q_x^* \qquad l_{x+1} = l_x - d_x \qquad L_x \approx l_{x+1} + \tfrac{1}{2}d_x \qquad T_x = \sum_{y \geq x} L_y \qquad e_x^\circ = \frac{T_x}{l_x}.$$

Here l_x is the number of survivors at age x, d_x is the number of the deaths, L_x is the number of years lived between x and $x+1$, T_x is the total number of years lived by the cohort from age x onwards, and e_x° the remaining life expectancy at age x. For the last open age interval 115+, we assume deaths occur at a constant rate.

5.3 Computation of Period Life Expectancy at Age 65

To compute the expected remaining lifetime at age 65 [Orford and Hennington, 2024; Dickson et al., 2013], denoted e_{65}, for each gender and state cluster, we first convert the annual mortality probabilities q_x obtained from the training data into a continuous-time force of mortality μ_x. For an individual aged x, the force of mortality is defined as:

$$\mu_x = -\ln(1 - q_x),$$

where q_x is the one-year death probability between ages x and $x+1$.

Given μ_x, the cumulative survival probability from age 60 to age $65 + j$ is:

$$p_{65}(j) = \exp\left(-\sum_{k=0}^{j-1} \mu_{65+k} \right),$$

for $j = 1, 2, \ldots, H$, where H is the maximum projection horizon (here $H = 50$ years). The period life expectancy at age 60 is then calculated as:

$$e_{65} = \frac{1}{2} + \sum_{j=1}^{H} p_{65}(j).$$

The $\frac{1}{2}$ term provides the standard continuity correction as the ages are in whole numbers.

5.3.1 Implementation

In practice, for each gender-state combination, we:

1. Filter the estimated q_x values for ages $x \geq 60$.
2. Convert q_x to $\mu_x = -\ln(1 - q_x)$, applying a numerical safeguard to ensure $q_x \in (0, 1)$.
3. Compute the cumulative hazard $\sum_{k=0}^{j-1} \mu_{65+k}$ and obtain survival probabilities $p_{65}(j)$.
4. Sum the survival probabilities over a 50-year horizon and add the $\frac{1}{2}$ adjustment.

This yields e_{65} for each cluster and gender directly from the period mortality rates estimated in the training step. The formula is consistent with standard actuarial practice for period life expectancy.

5.4 Model Performance and Comparison

This section goes over the performance of the models discussed under Chapter 4. We provide all the metrics discussed under Section 5.1. As the clusters and algorithms are different between models the ALT is used as a comparison. We further provide the model performance against each gender. The metrics are summarised under Section 5.5.

5.4.1 K-means Clustering

K-means clustering was performed on both the indicator dataset and the disease dataset, this will allow us to compare the performance of the model using the two datasets. In particular, it allows us to identify the impact of the simplification (Section 3.5) on the model performance (Section 5.4.3).

5.4.1.1 Indicator Dataset

Unfortunately, the relevant model values were not vetted in time for the thesis.

5.4.1.2 Disease Dataset

The accuracy metrics for the overall model using the k-means algorithm on the disease dataset are shown in Table 5.2. This also includes the performance of the ALT for each of the clusters. Additionally, the performance of the model split by gender (along with the ALT) is shown in Table 5.3.

Table 5.2: Model performance (overall): ALT vs k-means (disease-only).

Source	n	MSE	RMSE	MAE	Bias	Cali	R2
k-means	5.38e+02	4.051e-04	2.013e-02	9.273e-03	1.955e-03	9.634e-01	9.635e-01
ALT	5.38e+02	1.143e-03	3.380e-02	2.284e-02	-1.974e-03	1.020e+00	8.928e-01

Table 5.3: Model performance by gender: ALT vs k-means (disease-only).

Gender	Source	n	MSE	RMSE	MAE	Bias	Cali	R2
Female	k-means	2.74e+02	3.022e-04	1.738e-02	8.414e-03	3.537e-04	9.989e-01	9.711e-01
Female	ALT	2.74e+02	8.950e-04	2.992e-02	2.039e-02	7.941e-04	9.957e-01	9.143e-01
Male	k-means	2.64e+02	5.118e-04	2.262e-02	1.016e-02	3.617e-03	9.333e-01	9.582e-01
Male	ALT	2.64e+02	1.399e-03	3.741e-02	2.538e-02	-4.847e-03	1.047e+00	8.729e-01

As shown in Tables Table 5.2 and Table 5.3, all accuracy metrics indicate that the k-means algorithm outperforms the ALT model. Using the selected metric, the weighted root mean square error (wRMSE) for this model is 0.00783. This implies that, after accounting for the distribution of observations, the predicted mortality rates deviate from the expected values by an average of 0.783%. When the wRMSE is calculated separately by gender, female predictions deviate by approximately 0.805%, while male predictions deviate by 0.762%.

In contrast, the deviation in life expectancy predictions from the ALT model is more substantial. On average, the k-means predictions differ by 5.148 years from the ALT estimates (mean squared error = 26.501; root mean squared error = 5.148). A gender-specific analysis reveals that this deviation is greater for females (6.17 years) than for males (3.87 years).

The predicted life expectancy along with the actual life expectancy for each for each cluster split by gender is shown in Figure 5.2. The prediction and actual (calculated based on observed q_x) were close which is a reflection for the accuracy shown by the model predictions.

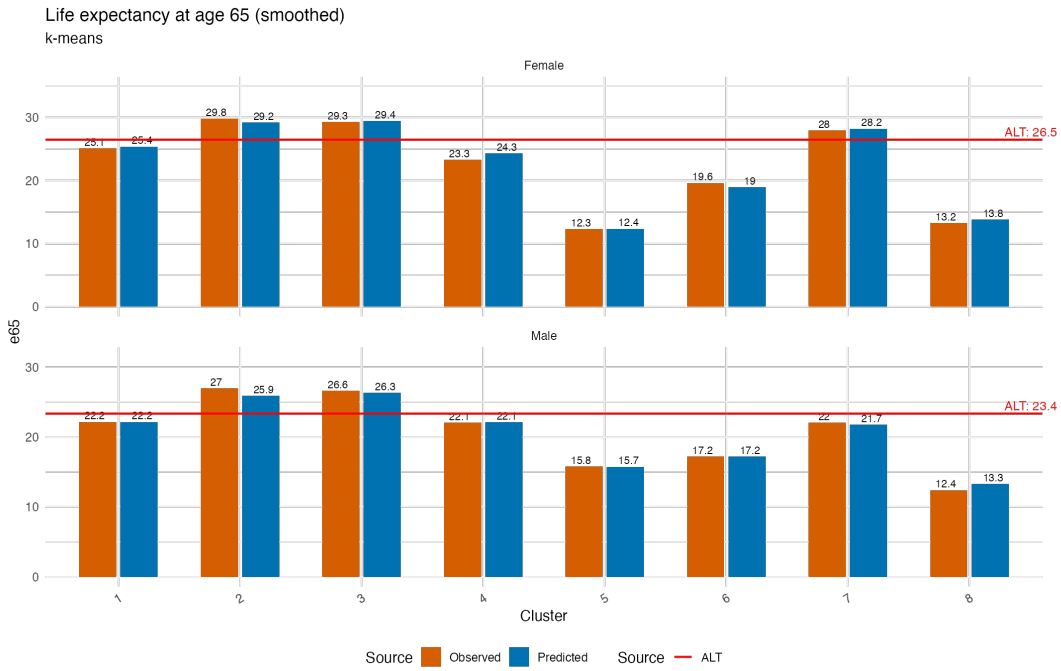


Figure 5.2: Life expectancy for each cluster in the k-means clustering model.

5.4.2 Survival Tree

The decision tree algorithm did not have comparable performance, and thus only the survival tree is included. The survival tree was trained on both the indicator and the disease dataset, the performance of both the performance metrics have been provided for this along with an analysis of the impact of the summarisation of the variables provided under Section 5.4.3.

5.4.2.1 Indicator Dataset

Tables 5.4 and 5.5 show that the survival tree trained on the indicator dataset materially outperforms the ALT benchmark across all metrics. Overall, the survival model attains an RMSE of 0.03817 versus 0.1373 for ALT, a lower MAE (0.01852 vs 0.06848), smaller absolute bias (0.006974 vs 0.03853 in magnitude), a calibration slope closer to one (0.9279 vs 0.8102), and a higher R^2 (0.9426 vs 0.2726).

Even when split by gender, the survival tree model outperforms ALT on all metrics, and its bias remains closer to zero for both groups.

Table 5.4: Model performance (overall): ALT vs Survival (indicator dataset).

Source	n	MSE	RMSE	MAE	Bias	Cali	R2
Survival	6.38e+02	1.457e-03	3.817e-02	1.852e-02	6.974e-03	9.279e-01	9.426e-01
ALT	6.41e+02	1.886e-02	1.373e-01	6.848e-02	-3.853e-02	8.102e-01	2.726e-01

Table 5.5: Model performance by gender: ALT vs Survival (indicator dataset).

Gender	Source	n	MSE	RMSE	MAE	Bias	Cali	R2
Female	Survival	3.28e+02	1.188e-03	3.447e-02	1.613e-02	6.154e-03	9.234e-01	9.365e-01
Female	ALT	3.29e+02	1.137e-02	1.066e-01	5.448e-02	-2.423e-02	7.994e-01	3.956e-01
Male	Survival	3.10e+02	1.741e-03	4.172e-02	2.104e-02	7.842e-03	9.304e-01	9.455e-01
Male	ALT	3.12e+02	2.677e-02	1.636e-01	8.325e-02	-5.362e-02	8.202e-01	2.002e-01

Figure 5.3 summarises life expectancy across terminal nodes for the survival tree trained on the indicator dataset. The model identifies clear between-group differences in longevity while maintaining strong absolute fit within groups, consistent with the accuracy and calibration results reported above. The deviation in life expectancy captured by the survival tree on the indicator dataset is significantly higher than that captured by k-means.

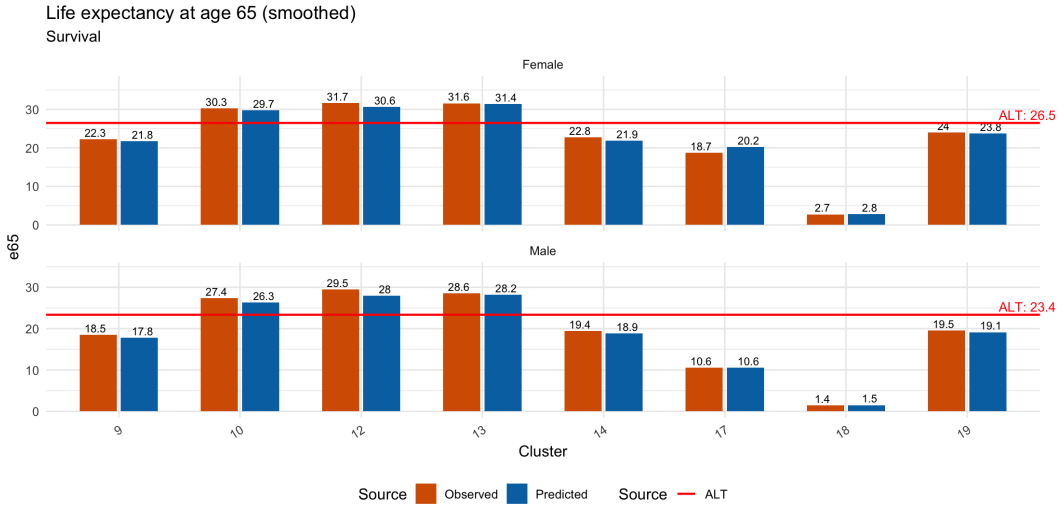


Figure 5.3: Life expectancy at age 65 across terminal nodes for the survival tree (indicator dataset).

5.4.2.2 Disease Dataset

As shown in Table 5.6 and Table 5.7, the survival tree outperforms performs compared to the ALT benchmark in most metrics. The overall RMSE of 0.0465 and weighted RMSE (wRMSE) of 0.0439 are considerably higher than those observed for the k-means model. When split by gender, female predictions show a wRMSE of 0.0388, while male predictions are even less accurate at 0.0478. These results suggest that the survival tree lacks the precision required for robust mortality modelling in this context. Although the R^2 of 0.914 indicates a moderate degree of variance explained, it is still below the performance of the clustering and Markov-based approaches.

Table 5.6: Model performance (overall): ALT vs Survival (disease-only).

Source	n	MSE	RMSE	MAE	Bias	Cali	R2
Survival	4.96e+02	2.166e-03	4.654e-02	2.145e-02	7.954e-03	8.889e-01	9.141e-01
ALT	5.03e+02	1.900e-02	1.378e-01	6.608e-02	-5.452e-02	8.506e-01	3.086e-01

Table 5.7: Model performance by gender: ALT vs Survival (disease-only).

Gender	Source	n	MSE	RMSE	MAE	Bias	Cali	R2
Female	Survival	2.46e+02	1.452e-03	3.810e-02	1.780e-02	5.261e-03	9.092e-01	9.253e-01
Female	ALT	2.51e+02	1.478e-02	1.216e-01	5.760e-02	-4.651e-02	8.414e-01	3.808e-01
Male	Survival	2.50e+02	2.869e-03	5.357e-02	2.504e-02	1.060e-02	8.769e-01	9.061e-01
Male	ALT	2.52e+02	2.320e-02	1.523e-01	7.452e-02	-6.250e-02	8.586e-01	2.528e-01

The life expectancy comparisons in Figure 5.4 show that, although the overall accuracy of the survival model is lower, it captures a higher degree of deviation in life expectancy across groups. That is, though the model is able to identify splits with higher mortality differences, the prediction within each split is not as accurate. This outcome is consistent with the expected trade-off between accuracy and deviation, where models that explain more variation often do so at the cost of predictive precision.

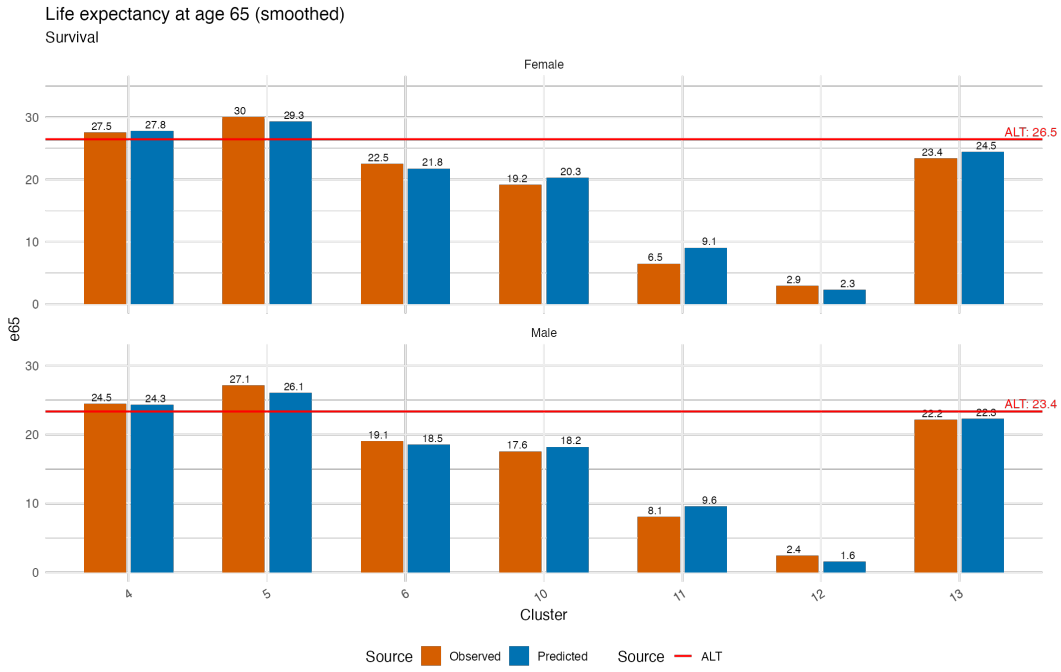


Figure 5.4: Life expectancy for each terminal node in the survival model.

5.4.3 Comparison between the Indicator Dataset and the Disease Dataset

Since the two splits are different between the two models, no direct comparison can be performed. To enable comparison, each model's metrics are normalised by the corresponding ALT metric computed on the same rows (i.e., the same split used by that model). In other words, we report *relative* metrics in Table 5.8.

Table 5.8: Relative performance to ALT (Model / ALT). Values < 1 for error metrics indicate improvement over ALT; values > 1 for Cali and R^2 indicate higher calibration slope and goodness-of-fit than ALT.

Dataset	Rel. MSE	Rel. RMSE	Rel. MAE	Bias Ratio	Cali Ratio	R^2 Ratio
Indicator	0.0773	0.2780	0.2704	-0.1810	1.1453	3.4578
Disease-only	0.1140	0.3377	0.3246	-0.1459	1.0450	2.9621

Interpretation and implications. Normalising each model to its own ALT shows that both specifications reduce error, with the indicator version doing best overall. That said, the extra gain from indicators over the disease-only tree is small and comes at the cost of interpretability. The gap between datasets is likely due to the loss of data when creating the disease dataset as not all indicator variables are used, as the survival tree on the disease dataset still outperforms the ALT.

Potential improvement. A potential further improvement of this research is to refine the disease-only specification rather than further intensifying indicator summarisation. With appropriate medical input, the disease set used in the model could be revisited and expanded beyond the current six conditions to capture higher variation and information, clinically meaningful severity groupings, and interactions (e.g., cardiovascular–diabetes clusters). Incorporating these additions; while preserving transparent, clinically interpretable features, should recover information lost to summarisation and improve both calibration and accuracy without sacrificing auditability.

5.4.4 Regression

As shown in Table 5.9 and Table 5.10, the regression model improves substantially on the ALT benchmark and delivers a strong balance of performance. Although the overall RMSE of 0.0522 and weighted RMSE of 0.0496 are higher than those of the k-means and Markov models, the regression achieves the highest deviation values across all models (overall 9.48; 10.1 for females and 8.71 for males). This indicates that the regression captures the greatest degree of heterogeneity in the data, making it particularly effective at distinguishing differences across subgroups. This is as expected as it accommodates for all combinations of diseases, it covers a much larger number of potential sub groups. The R^2 of 0.915 remains strong, confirming that the model explains a large proportion of the observed variance.

Table 5.9: Model performance (overall): ALT vs Regression (disease-only).

Source	n	MSE	RMSE	MAE	Bias	Cali	R2
Regression	9.61e+02	2.726e-03	5.221e-02	2.782e-02	8.092e-03	8.627e-01	9.152e-01
ALT	9.61e+02	2.366e-02	1.538e-01	8.582e-02	-6.838e-02	9.068e-01	2.318e-01

Table 5.10: Model performance by gender: ALT vs Regression (disease-only).

Gender	Source	n	MSE	RMSE	MAE	Bias	Cali	R2
Female	Regression	4.72e+02	1.579e-03	3.973e-02	2.424e-02	-1.669e-04	9.535e-01	9.267e-01
Female	ALT	4.72e+02	1.787e-02	1.337e-01	7.536e-02	-5.749e-02	9.163e-01	3.060e-01
Male	Regression	4.89e+02	3.834e-03	6.192e-02	3.128e-02	1.606e-02	8.193e-01	9.168e-01
Male	ALT	4.89e+02	2.925e-02	1.710e-01	9.592e-02	-7.888e-02	8.931e-01	1.794e-01

The life expectancy comparisons in Figure 5.5 illustrate this behaviour: while predictions are less precise than those of k-means, the wider spread of predicted life expectancies highlights the regression model's ability to reflect variation across the population. This is also expected due to the higher number of categories.

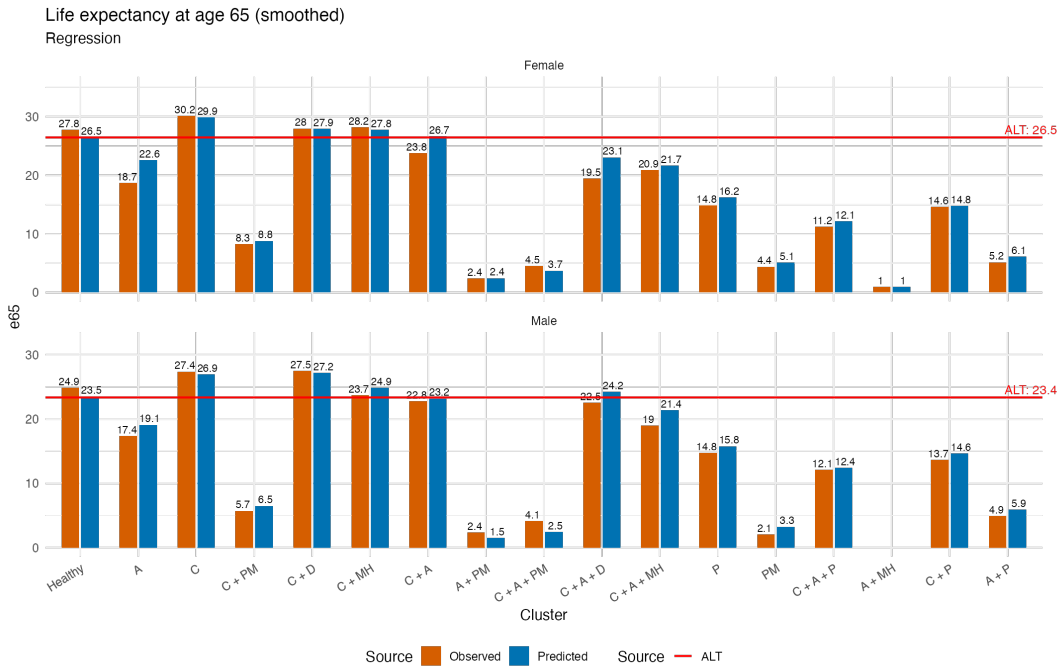


Figure 5.5: Life expectancy for all observed combinations in testing dataset.

5.4.5 Markov Chain

For the Markov chain, the state selection is done based on most common disease combination. The performance of the model is checked on model outputs without incorporating transitions, this allows for a fair comparison between model performance as the transitions allow for long term impacts that can not be verified using the testing data.

As shown in Table 5.11 and Table 5.12, the Markov chain model provides the most balanced trade-off between accuracy and deviation. It records strong accuracy (overall RMSE of 0.0296 and weighted RMSE of 0.0188) while also achieving high deviation values (overall 7.20; 8.57 for females and 5.50 for males), which are considerably higher than those obtained under k-means. This balance suggests that the Markov model is not only precise but also able to capture meaningful variation across health states. Its explanatory power is further supported by an R^2 of 0.951, only marginally below the k-means result.

Table 5.11: Model performance (overall): ALT vs Markov (disease-only).

Source	n	MSE	RMSE	MAE	Bias	Cali	R2
Markov	8.08e+02	8.756e-04	2.959e-02	1.526e-02	3.258e-03	9.400e-01	9.511e-01
ALT	8.11e+02	7.964e-03	8.924e-02	4.596e-02	-2.610e-02	1.086e+00	5.609e-01

Table 5.12: Model performance by gender: ALT vs Markov (disease-only).

Gender	Source	n	MSE	RMSE	MAE	Bias	Cali	R2
Female	Markov	4.05e+02	7.187e-04	2.681e-02	1.461e-02	1.634e-03	9.707e-01	9.549e-01
Female	ALT	4.05e+02	6.378e-03	7.986e-02	4.248e-02	-2.172e-02	1.068e+00	6.230e-01
Male	Markov	4.03e+02	1.033e-03	3.215e-02	1.591e-02	4.889e-03	9.159e-01	9.493e-01
Male	ALT	4.06e+02	9.546e-03	9.770e-02	4.943e-02	-3.047e-02	1.106e+00	5.064e-01

The life expectancy comparisons in Figure 5.6 confirm this balance, predictions remain close to observed values while still reflecting variation between groups, making the Markov model particularly suitable for actuarial projection and transition-based analyses.

5.4.6 RMSE on Log Mortality

Overall, every model outperforms its ALT benchmark. Within the disease dataset, k-means and the survival tree achieve the lowest errors, followed by the Markov and regression models. For the one model tested on both datasets (the survival tree), the indicator specification edges out the disease specification, indicating a modest accuracy gain from indicator encoding. Overall, the indicator-based survival tree is the strongest performer, but its advantage over the disease-based version is small.

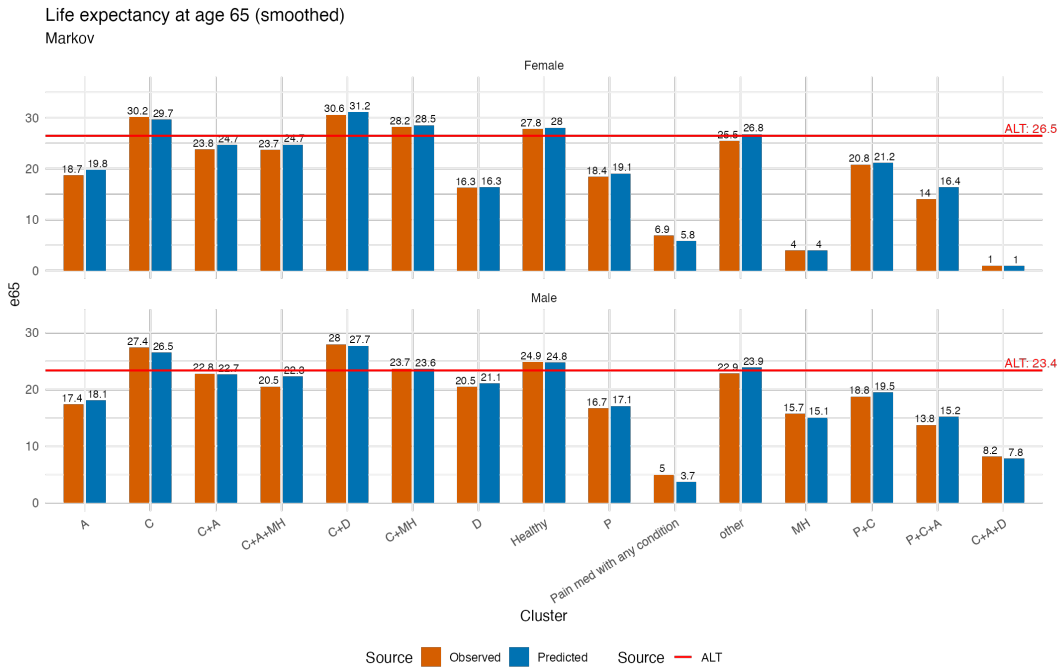


Figure 5.6: Life expectancy for each state in the Markov chain model.

Table 5.13: Log-RMSE vs ALT (overall): merging Indicator (ID_ST) and Disease datasets. Survival Tree appears twice (by dataset).

Model	Indicator dataset		Disease dataset	
	Model	ALT	Model	ALT
k-means			0.185	0.500
Survival	0.182	1.080	0.187	1.100
Regression			0.266	1.270
Markov			0.211	0.862

5.5 Summary and Final Model Selection

The Markov model emerges as the strongest overall performer because it balances accuracy with meaningful explanatory power (Table 5.15). While the k-means algorithm achieves the lowest error metrics (RMSE and wRMSE), the Markov approach offers the second-best error performance while also achieving relatively high deviation values. Since higher deviation indicates that the model captures more heterogeneity across states and groups, the Markov model demonstrates that it is both accurate and sensitive to variation in the data. This is especially evident in the female and male splits, where its deviation values are consistently strong compared to other models. At the same time, the Markov model maintains a high R^2(0.951), which is only marginally below k-means (0.963). Taken together, these results show that the Markov model provides the best trade-off: it preserves strong predictive accuracy

Table 5.14: Log-RMSE vs ALT by gender: merging Indicator (ID_ST) and Disease datasets. Survival Tree appears twice (by dataset).

Model	Gender	Indicator (ID_ST)		Disease (ST)	
		Model	ALT	Model	ALT
k-means	Female			0.200	0.537
k-means	Male			0.167	0.459
Survival	Female	0.189	1.040	0.179	1.150
Survival	Male	0.173	1.130	0.194	1.050
Regression	Female			0.290	1.290
Regression	Male			0.240	1.250
Markov	Female			0.220	0.892
Markov	Male			0.201	0.832

Table 5.15: Model Performance Summary. All provided metrics are for models trained on the disease dataset.

Metric	k-means	Survival	Regression	Markov
# clusters	8	7	17	12
RMSE.	**0.0201**	0.0465	0.0522	0.0296
wRMSE.	**0.0071**	0.0439	0.0496	0.0188
wRMSE- F	**0.0070**	0.0388	0.0322	0.0156
wRMSE- M	**0.0071**	0.0478	0.0602	0.0214
Dev.	5.06	8.42	**9.48**	7.20
Dev. (F)	5.98	9.02	**10.1**	8.57
Dev. (M)	3.92	7.77	**8.71**	5.50
R^2	**0.9635**	0.9141	0.9152	0.9511

while also capturing richer variation across the population. This makes it particularly suitable for actuarial and health modelling applications, where recognising differences across subgroups and transitions between health states is as important as minimising error.

5.6 Conclusion

This chapter assigns numerical values on the performance of the respective models to be able to identify which is the best performing model. We assess both the accuracy of clusters and the deviation between clusters. While there is a clear trade-off between these, we allowed for a higher importance to be provided on accuracy to ensure that predictions are reliable. Overall, the Markov chain model was selected as it was deemed to have the best overall values.

Chapter 6

Results

This chapter presents the final proposed model, evaluates its performance (Section 6.1), and quantifies the monetary implications for retirees (Section 6.3). Section 6.2 derives the formula for the mortality calculations using transitions.

Some sections use financial terminology; see the Financial Glossary for definitions. Additional computational details are provided in the corresponding subsections and appendices.

6.1 Proposed Model

The final model is a non-standard Markov chain with 15 states, each representing a common combination of diseases. Unlike typical mortality-based Markov chains that include death as a state, we use age and gender within each state to predict mortality. We make the key assumption that death occurs before any state transitions. When calculating the mortality for each health state, we account for the impact of age and gender in the same manner as the ALT, allowing for this model to be a strictly better model than the ALT.

The model outperforms alternative approaches by capturing the dynamic nature of disease progression, an essential feature for modelling retirees whose health evolves over time. This is especially important since the models are based on only three years of data, making it unlikely that other approaches can capture the long-term effects of retirees' changing health states.

Potential models and their benefits are discussed in Chapter 4. As shown in Chapter 5, when comparing model performance, the proposed model achieves high deviation while maintaining an acceptable level of accuracy. It also has the additional benefit of explicitly incorporating relevant health transitions.

6.1.1 Transition Probabilities

The proposed Markov chain model captures the changes in individual health status through a transition matrix. The transition matrix specifies the probability of an individual moving from one state to another at the end of each time step (in this case, a year).

As mentioned in Section 4.9.2.1, considering transitions for all age and gender combinations will result in a sparse vector due to lack of sufficient data. Hence, to model the impact of gender and age on transition probabilities, we used three age bins: 50-70, 70-90, and 90 and above. Statistical significance indicates that the impact of gender on prediction falls below a significance level of 5%, while the impact of age bins remains prominent.

The final model transition probabilities are provided in Figure 6.1, Figure 6.2, and Figure 6.3 for the age bins 50-70, 70-90, and 90 and above respectively. The algorithm for incorporating these transition probabilities into the model is described in Section 6.1.2. Life expectancies, with and without transitions, are shown in Figure 4.20. A clear indicator of the proposed model's improved accuracy is the reduced variation in average life expectancy between clusters.

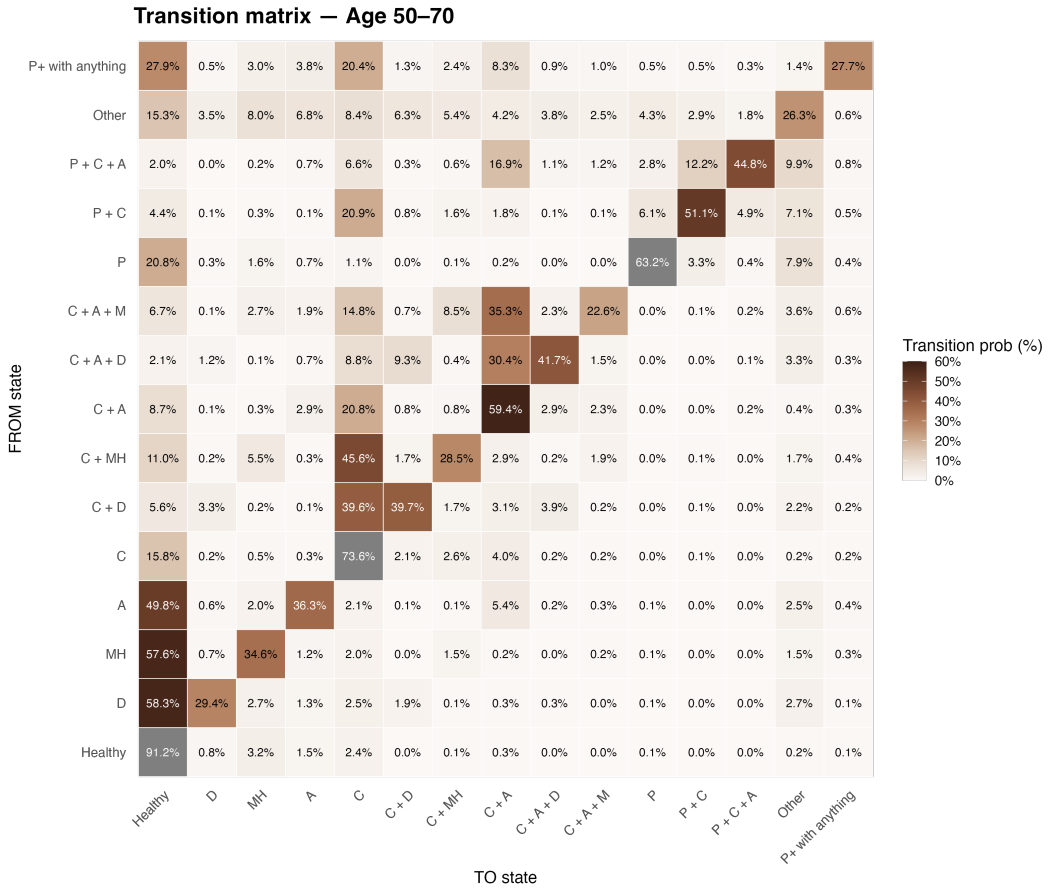


Figure 6.1: Transition probabilities for those aged 50-70 for the Markov chain process.

6.1.2 Incorporating Transition Probabilities into Mortality Calculations

We extend standard life table methods to a multi-state Markov model in which each state represents a health status. The model assumes that *deaths occur before transitions*

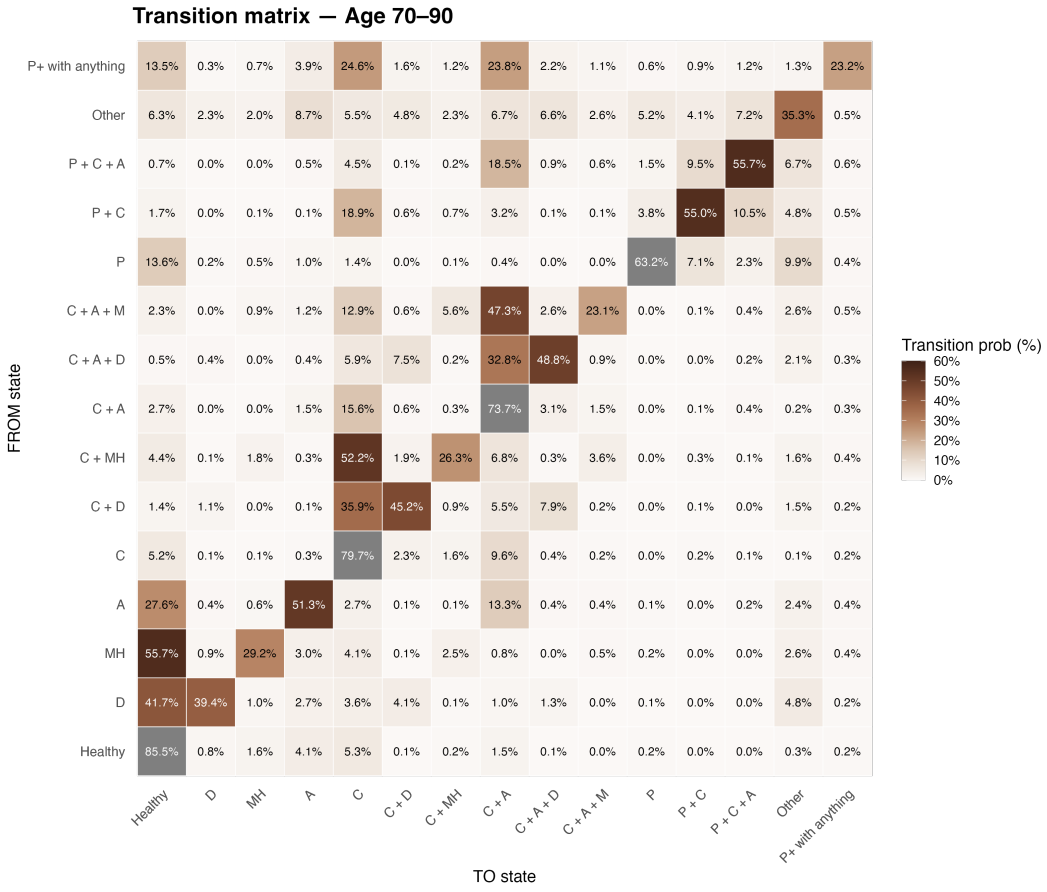


Figure 6.2: Transition probabilities for those aged 70-90 for the Markov chain process.

within each year. Mortality depends on the state at the start of the year, and only survivors transition to states for the following year.

We first show how the model updates the state distribution $\ell_{x,g}$. We then show how to use our model to compute the probability an individual aged x with gender g in state s dies within the next t years, ${}_tq^{(s)}_{x,g}$. As this is the format that classical single-state life tables give, we explain that this is sufficient to calculate standard actuarial formulas.

6.1.2.1 Notation and Assumptions

Let us denote,

- x as age where $x \in \{50, 51, ..., \omega\}$ and ω is the limiting age (maximum possible age), g as gender, and s as health state where $s \in \{1, ..., S\}$ (S=15).
- $q^{(s)}_{x,g}$: one-year mortality probability for age x, gender g, and state s.
- x_0 as the starting age and s_0 as the starting state. Thus life expectancy at age x_0

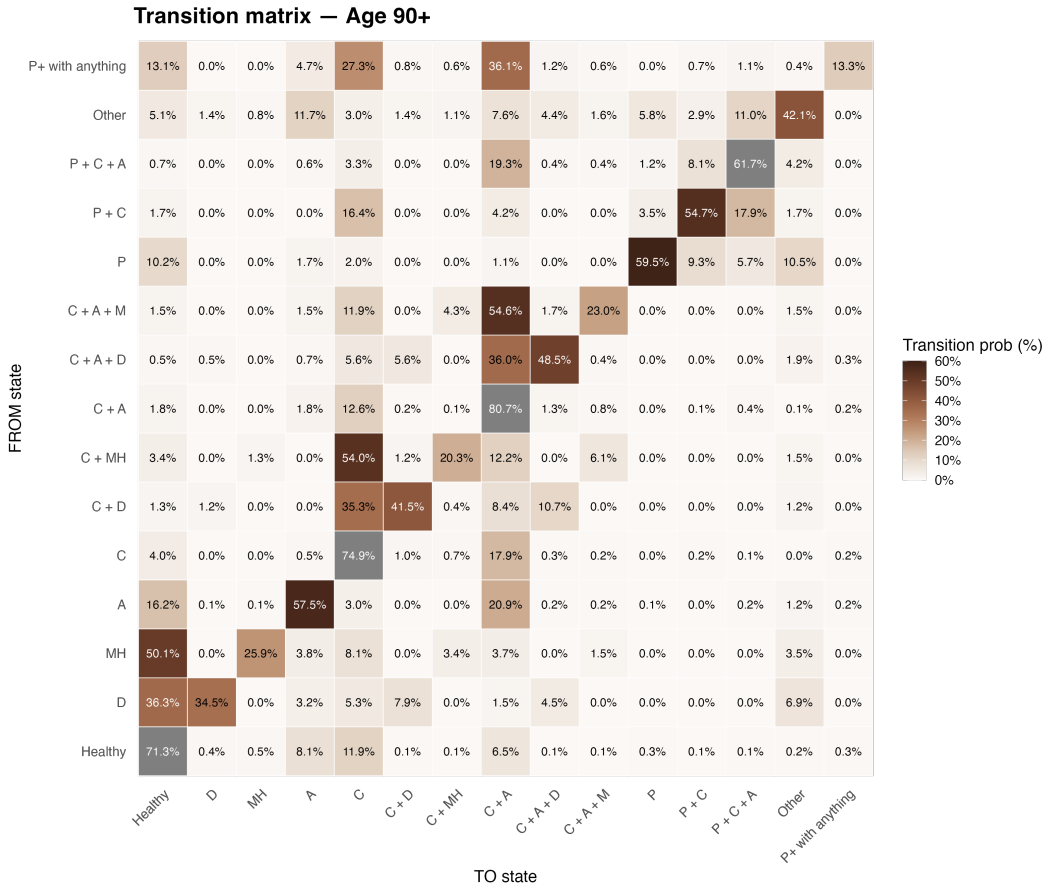


Figure 6.3: Transition probabilities for those aged 90 and above for the Markov chain process.

starting from state s_0 depends on $q_{x_0,g}^{(s_0)}$ and $q_{x,g}^{(s)}$ for $x \in \{x_0+1,\ldots,\omega-1\}$ and $s \in \{1,...,S\}$.

- a as age bin, one of 50-70, 70-90, and 90+, and $a(x)$ as the function that maps an age x to an age bin.

- $\mathbf{T}_a$ as the $S \times S$ transition matrix between states among *survivors*, where the entry in the s^{th} row and r^{th} column $\mathbf{T}_a[s,r]$ denote the probability of transitioning from state s at age x to state r at age $x+1$ for all $x \in a$. Thus each row sums to one. Note that transition probabilities are independent of gender and only depend on age bin a.

- $\boldsymbol{\ell}_{x,g} = (\ell_{x,g}^{(1)},\ldots,\ell_{x,g}^{(S)})$: the state distribution at age x and gender g (technically conditioned on x_0 and s_0), which is a $S \times 1$ vector of the proportion alive at the *start* of age x in each state. At the starting age x_0:

$$\ell_{x_0,g}^{(s_0)} = 1, \qquad \ell_{x_0,g}^{(s)} = 0 \quad \forall s \neq s_0.$$

Despite it being defined as a column vector, we will use it in row vector form with transition matrices.

- ${}_tq^{(s)}_{x,g}$ as the probability for an individual aged x (and with gender g) in state s dying within the next t years.

- ${}_tp^{(s)}_{x,g}$ as the probability for an individual aged x (and with gender g) in state s surviving t years, so ${}_tp^{(s)}_{x,g} = P(\text{alive at time } t \mid \text{start in state } s \text{ at age } x \text{ and gender } g)$ and ${}_tq^{(s)}_{x,g} = 1 - {}_tp^{(s)}_{x,g}$.

6.1.2.2 Updating State Distributions

We assume each year proceeds as: (i) possible death within the year using state-specific one-year mortality $q^{(i)}_{x,g}$; (ii) survivors then transition at year-end according to the transition matrix $\mathbf{T}_{a(x)}$ appropriate for age x.

$\mathbf{T}_{a(x)}$ already represents (ii). To represent (i), let $\mathbf{q}_{x,g} = \left(q^{(1)}_{x,g}, \ldots, q^{(S)}_{x,g}\right)^\top$ and define the survival operator

$$S_{x,g} \;=\; D(\mathbf{1} - \mathbf{q}_{x,g})\,,$$

where $D(\cdot)$ forms a diagonal matrix from a given vector. Then we can combine them to get

$$A_{x,g} \;=\; S_{x,g}\,\mathbf{T}_{a(x)},$$

which maps the *state distribution of survivors at the start of age x* to the state distribution of survivors at the start of age $x+1$:

$$\boldsymbol{\ell}^\top_{x+1,g} = \boldsymbol{\ell}^\top_{x,g} A_{x,g}$$

6.1.2.3 Calculating ${}_tp^{(s)}_{x,g}$ and ${}_tq^{(s)}_{x,g}$

To get ${}_1p^{(s)}_{x,g}$, we can start with the state distribution as one-hot vector e_s to get the resulting state probabilities for age $x+1$ from state x, $e_s^T A_{x,g}$. Then to get the probability of surviving to age $x+1$ we can add up these state probabilities, so

$${}_1p^{(s)}_{x,g} = e_s^T A_{x,g}\mathbf{1} \qquad {}_1q^{(s)}_{x,g} \;=\; 1 - {}_1p^{(s)}_{x,g} = q^{(s)}_{x,g}.$$

Now note that $e_s^T A_{x,g}\mathbf{1} = (A_{x,g}\mathbf{1})[s]$, so we can calculate the $S \times 1$ vector of ${}_1p^{(s)}_{x,g}$ for each s as

$$\boxed{{}_1\mathbf{p}_{x,g} \;=\; A_{x,g}\mathbf{1}}$$

and similarly for ${}_1q^{(s)}_{x,g}$

$$\boxed{{}_1\mathbf{q}_{x,g} \;=\; \mathbf{1} - {}_1\mathbf{p}_{x,g}\,.}$$

Define the t-year survivor propagation matrix

$$M_t \;=\; \prod_{j=0}^{t-1} A_{x+j,g} \;=\; A_{x,g} A_{x+1,g} \cdots A_{x+t-1,g}.$$

Then the vector of t-year survival probabilities (one entry per start state) is

$$\boxed{{}_t\mathbf{p}_{x,g} \;=\; M_t \mathbf{1}\,,}$$

and

$$\boxed{{}_t\mathbf{q}_{x,g} \;=\; \mathbf{1} - {}_t\mathbf{p}_{x,g}\,.}$$

Recursive computation. Set ${}_0\mathbf{p}_{x,g} = \mathbf{1}$ and for $t \geq 1$,

$$\boxed{{}_t\mathbf{p}_{x,g} \;=\; A_{x,g}\, {}_{t-1}\mathbf{p}_{x+1,g}\,,} \qquad \boxed{{}_t\mathbf{q}_{x,g} \;=\; \mathbf{1} - {}_t\mathbf{p}_{x,g}\,.}$$

This makes explicit the yearly structure: survive within the year via $S_{x,g}$, transition among survivors via $\mathbf{T}_{a(x)}$, and iterate.

6.1.2.4 Mortality-Based Calculations with ${}_tq_{x,g}$

Now that transition probabilities have been incorporated into the definition of ${}_tq_{x,g}$, subsequent mortality-based calculations retain the same structure as in the classical single-state life table. The cumulative probability of death within t years, the survival probability over t years, and related life expectancy measures can all be expressed directly in terms of ${}_tq_{x,g}$. Consequently, standard actuarial formulas for life expectancy, annuities, and insurance values remain valid with no change to their form, however they now reflect the impact of both mortality and state transitions thanks to ${}_tq_{x,g}$.

6.2 Transition-Incorporated Life Expectancy Predictions Alongside ALT

As shown in Figure 6.4, the predicted life expectancy from the proposed model is often higher than the ALT baseline at younger entry ages. This occurs because the inclusion of health state transitions allows individuals to move into healthier states over time, which increases expected longevity compared with the static ALT assumptions that do not account for dynamic health improvements.

However, at older ages the model life expectancy estimates reduce and converge towards the ALT. Transitions at advanced ages capture the onset of co-morbidity, functional decline, and increased mortality risk associated with worsening health conditions. As a result, predicted life expectancy falls significantly for many health states beyond the mid-70s. This divergence reflects a key advantage of the transition-based approach: it captures both positive and negative trajectories of health over the

life course, rather than assuming a fixed risk profile.

Figure 6.4 illustrates these dynamics across entry ages and health states, with separate lines shown for males and females. Each panel corresponds to a specific baseline state (such as "Healthy," "C+MH," or "P"), and within each panel the model-based predictions (solid lines) are plotted against the ALT benchmarks (dotted lines). The model predicts higher life expectancy for all ages, indicating longevity gains due to the possibility of transitioning to better health states. In contrast, the higher decline in the older ages making it approach the ALT shows where the impact of comorbidity transitions begin to outweigh earlier improvements. The gender-specific lines further highlight that these effects differ in magnitude between males and females, with females generally showing higher life expectancy across most states.

The noticeable slope difference in predicted life expectancy, around age 70 or 90 depending on the disease, is primarily linked to the transition probabilities changing to account for worsening health status then. The use of smaller age bins would aid in a more accurate representation of the life expectancy relevant to the respective age. When multiple ages are grouped into a single bin, transition probabilities are averaged across a heterogeneous group of individuals, masking the variability in health pathways within that range. With narrower bins we would be able to identify the decline better. We believe that with smaller transition age bins we would be able to see a trend with the model life expectancy being higher than the ALT in the younger ages and lower than the ALT in older ages. Due to time constraints smaller bins were not exported.

It is also important to note that the ALT model, by construction, prices annuities without accounting for transitions between health states. While this leads to some degree of cross-subsidisation between individuals in different health conditions, with healthier individuals effectively subsidising those in poorer health, the more significant cross-subsidy arises between younger and older entrants (especially given pricing occurs with loadings when using the ALT).

Figure 6.5 shows that at younger entry ages, the proposed dynamic model prices slightly below the ALT, while at older ages it rises above the ALT. This pattern likely emerges because the ALT represents an average over all possible health paths, effectively smoothing improvement and deterioration into one curve. By contrast, the transition-based model explicitly captures both directions of change. Among younger individuals, although many improve in health over time, the model also accounts for those who remain unwell or worsen, reducing their expected lifetime and thus the fair annuity value. For older individuals, the ALT assumes steady decline, but the transition model recognises that a subset maintain or even recover stability, extending their expected lifetime and increasing the annuity price. In essence, the ALT captures only the "average" trajectory, whereas the transition model retains the heterogeneity of individual health paths, resulting in lower prices where deterioration risk dominates and higher prices where resilience becomes more influential. This richer, state-based structure explains the reversal observed in the figure and demonstrates how the dynamic model mitigates the need for broad loadings by aligning prices more closely with the true diversity of ageing pathways.

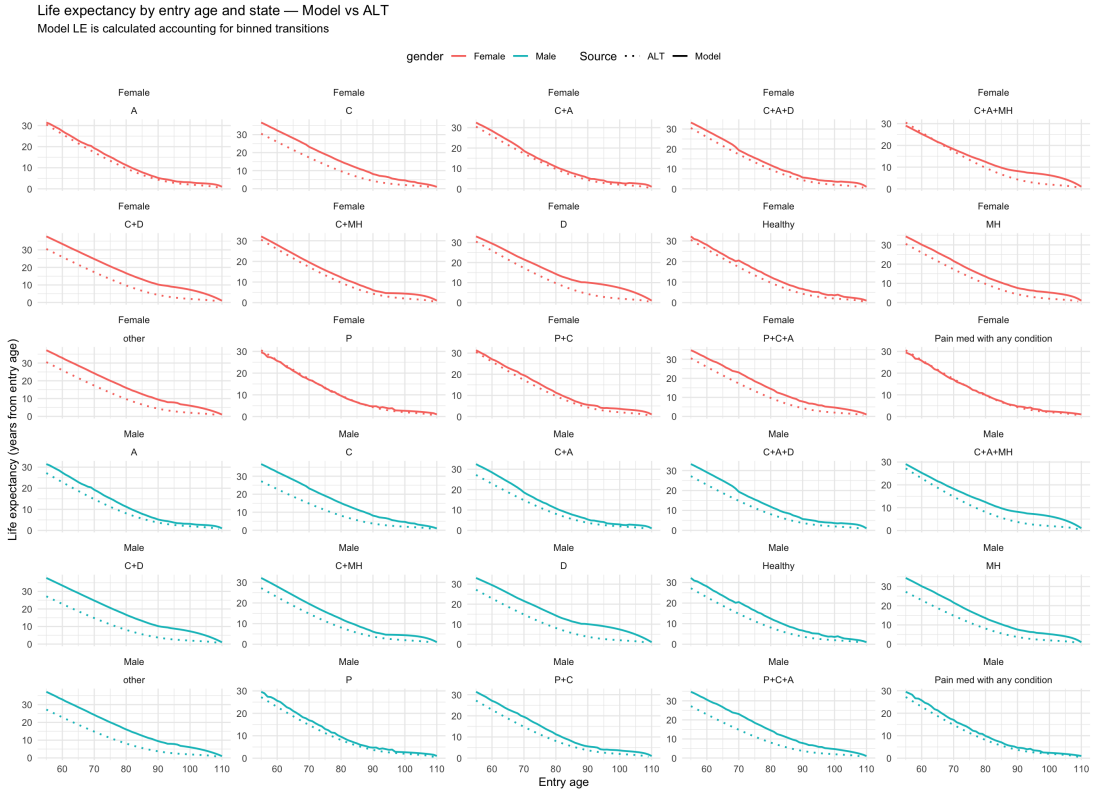


Figure 6.4: Life expectancy by age for the model against ALT.

6.3 Monetary Implications for Retirement

Upon retirement, individuals typically access their superannuation savings in one of two ways [Australian Taxation Office, 2025]:

1. converting the balance into an *income stream*, or

2. withdrawing the balance as a *lump sum*.

A combination of both methods is also common in practice.

This section examines the impact of the proposed mortality model on each option. In both cases, the choice of discount rate plays a central role in valuing future payments. For the following calculations we assume a discount rate of $d = 3\%$.

6.3.1 Income Stream

The most common approach to convert the superannuation balance to an income stream is to invest the balance in an annuity product, which is a savings product that will pay you a certain amount till death for a single initial payment. In this approach, the proposed mortality model will influence the pricing of the annuity.

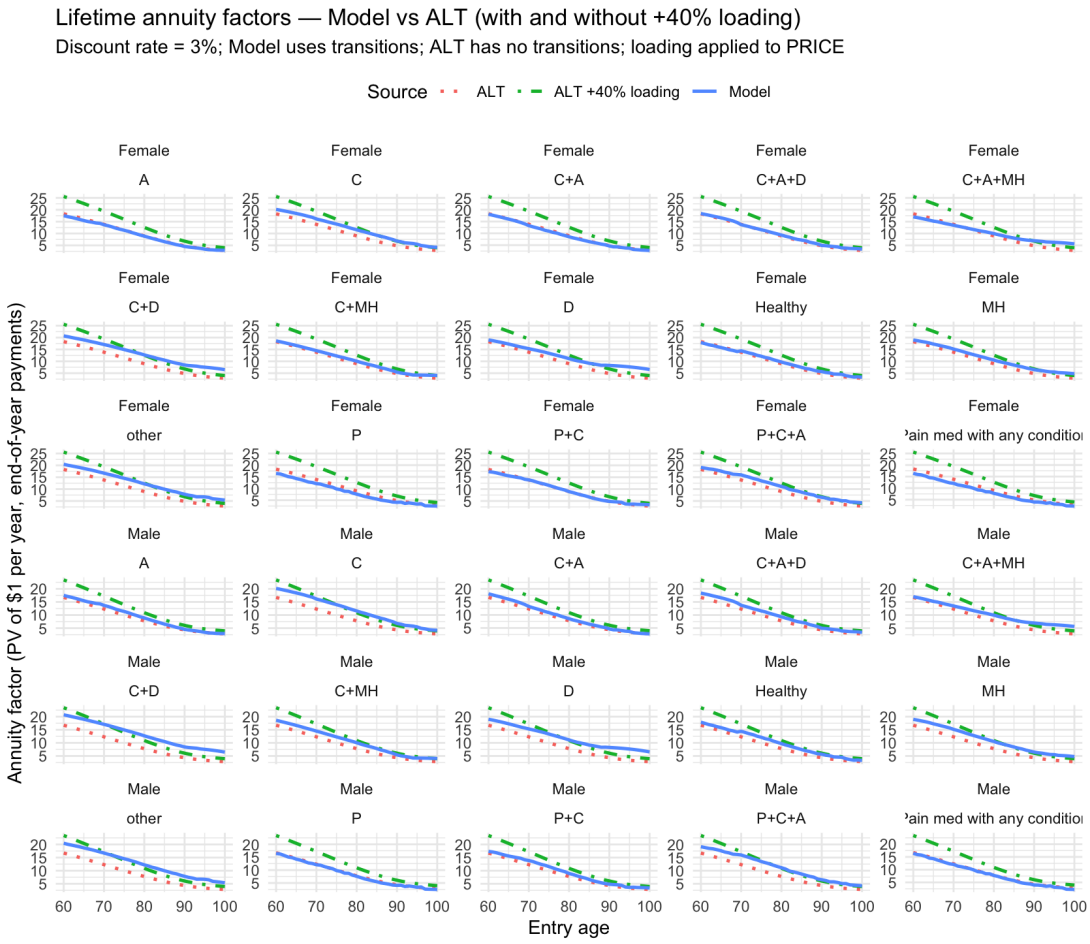


Figure 6.5: The price of a $1 income stream till death assuming discount rate is 3%. *Note:* the loading is an adjustment the insurance and superannuation companies make to account for the deviation seen from the population average. With more granular estimates the need for loading reduces.

For the calculation of the income steam, we have used the expected savings at the point of retirement which is around $ 500,000 of savings at age 60 [Australian Retirement Trust, 2025]. This is performed in contrast to calculating the annual annuity income for males from a $100,000 investment at age 65 as calculated in Huang et al. [2023] as it gives a more clear understanding of annual income and value obtained from the superannuation balance (see Section 6.3.1.1).

6.3.1.1 Annuity Calculation

Let q_x denote the mortality and $p_x = 1 - q_x$ denote the survival rate. Extending this, the probability of surviving k years from age x is given by

$$ {}_kp_x = p_x \times p_{x+1} \times \cdots \times p_{x+k-1}, $$

which captures the likelihood of being alive at each successive age.

An annuity pays a fixed amount at regular intervals for as long as the annuitant is alive. To value such an annuity, each future payment must be discounted to the present using a discount factor v,

$$v = \frac{1}{1+i},$$

where i is the effective annual interest rate. A payment made k years in the future is therefore worth v^k today.

The present value of a whole life annuity paying 1 per year to a life aged x is then given by

$$a_x = \sum_{k=0}^{\infty} v^k \, {}_kp_x$$

where v^k accounts for the time value of money, and ${}_kp_x$ ensures payments are only made if the individual is alive at time k In practice, the summation is truncated at a sufficiently high age (e.g., age 120) since the survival probabilities become negligible.

Finally, if the annuity pays R per year, the price of the annuity is simply

$$\text{Price} = R \times a_x$$

The price therefore depends critically on the mortality assumptions via q_x and on the choice of interest rate i. Higher mortality rates and higher discount rates lead to fewer expected payments and lower present values respectively, both of which reduce the annuity price.

6.3.1.2 Missing Entries due to Export Constrains

There are age, gender combinations within clusters that do not meet the requirements for exporting (Section 3.1). Due to the presence of these, interpolation was used to estimate these missing bounds using the two adjacent mortality values for the specific age.

6.3.1.3 Results

Due to the dynamic nature of the proposed model, younger retirees are expected to have higher life expectancy than those under the ALT. This is shown in Figure 6.6 where the model predicts higher life expectancy for all individuals aged 65. As expected, the life expectancy of of females outperform that of males in all disease combinations.

The resulting annual income from a $500,000 annuity investment is shown in Figure 6.7, varying significantly between clusters. Higher annuity payments correspond to shorter expected lifespans. At entry age 65, the model indicates higher annuity income for males than females, consistent with shorter male life expectancy, which reduces the annuity factor and increases annual payouts. Within each gender, bars

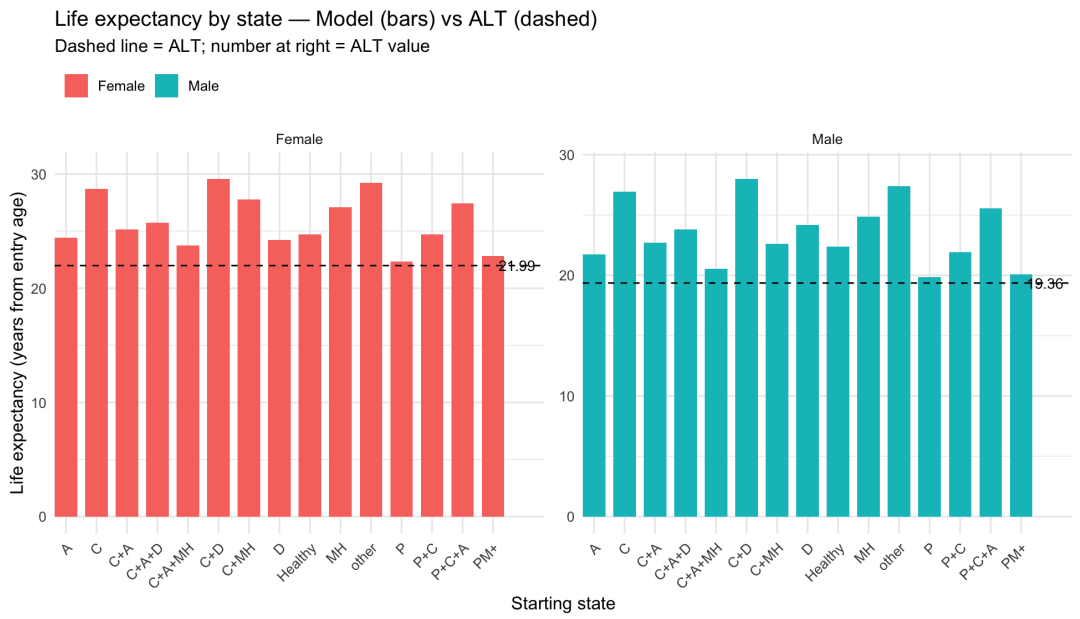


Figure 6.6: Life expectancy for a 65 year old upon retirement split by gender, compared to the ALT (dashed).

above the dashed ALT line represent higher forecast mortality (shorter life, higher income), while those below indicate lower mortality (longer life, lower income).

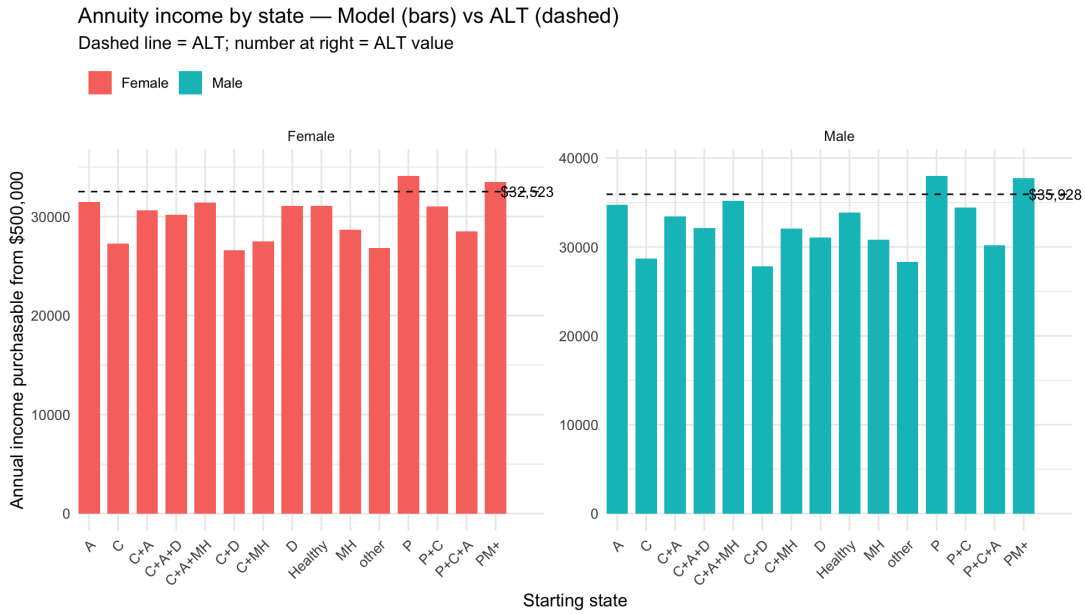


Figure 6.7: Annuity pricing for a 65 year old at retirement, compared to the ALT (dashed).

Although the proposed model produces lower income estimates for most clusters when compared to the ALT-based prediction, the higher predictive performance of

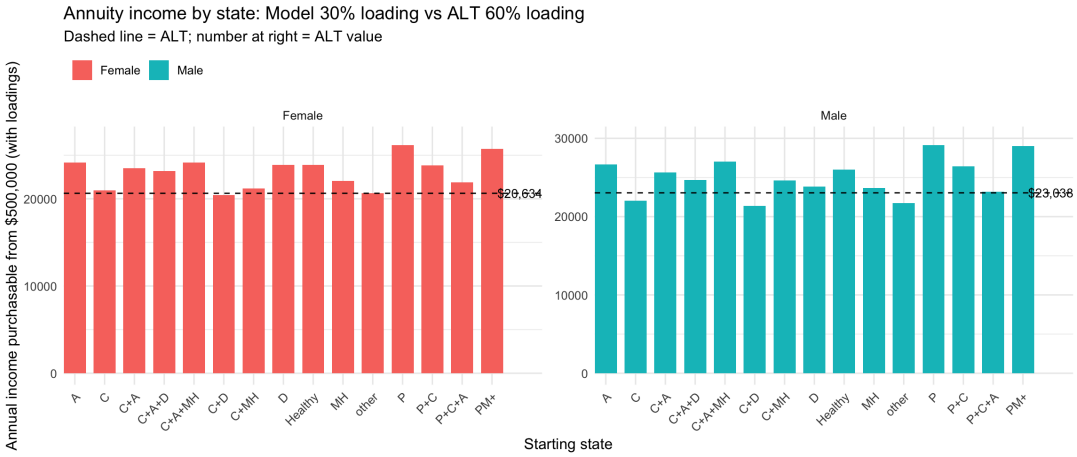


Figure 6.8: Annuity pricing for a 65 year old at retirement with 30% margin, compared to the ALT with 60% margin (dashed).

our model may actually increase income. Insurers typically set prices above these point estimates to allow for model uncertainty, parameter error, adverse selection, capital costs, regulatory and solvency requirements, and administration expenses. These loadings ensure that premiums remain adequate when experience deviates from the central estimate. If the model improves mortality prediction, the required uncertainty margin narrows, which reduces model and parameter risk allowances (though does not affect other loadings). To illustrate this, we assume a 60% margin for the ALT and a 30% margin for the proposed model, and show the resulting annual payments is higher in Figure 6.8.

6.3.2 Withdrawal from a Lump Sum

There is a growing demand for lump sum payments in retirement [Australian Bureau of Statistics, 2024], accompanied by an increasing need for more accurate life expectancy estimates. As more people rely on defined contribution schemes rather than guaranteed pensions, they are increasingly responsible for managing their own savings. This shift means that instead of receiving a regular income for life, many retirees receive a single lump sum and must decide how to make it last. Making those decisions is far from simple: if someone underestimates how long they will live, they risk running out of money. If they overestimate, they might restrict their spending too much and miss opportunities to enjoy their retirement. A realistic understanding of life expectancy is therefore essential. It helps retirees plan withdrawals that balance immediate needs with long-term security, reducing the chances of either exhausting their savings too soon or leaving a large unused balance behind.

We now compare withdrawing from a lump sum using the ALT predicted life expectancy to our predicted life expectancy. To do this, we first need to determine a lump sum withdrawal strategy and estimate the pension value and super balance.

There are numerous ways to withdraw from a lump sum, ranging from simple

fixed-value withdrawals to more sophisticated approaches. Some models draw a constant proportion of the remaining balance each year, while others base withdrawals on the maximum expected lifespan or on the estimated duration of a non-disabled lifetime [MacDonald et al., 2013]. Each method attempts to balance spending needs with the risk of outliving one's savings, but they vary in complexity and the assumptions they require. For the purposes of the following analysis, we assume constant withdrawals over the expected lifetime. This simplifies the modelling process while still capturing the core trade-off retirees face between maintaining a sustainable income and preserving their capital.

Once their super balance is exhausted, an individual will revert to the age pension, conditional on asset and income tests. This would be roughly $1,079.70 a fortnight at most [Services Australia, 2025]. Thus, to account for an average situation we have considered an income of $539.85 per fortnight. For a comfortable retirement, a rough balance of $600,000 is required [UniSuper]. Thus, we use an assumption of $600,000 as the balance to identify the impact on a comfortable lifestyle.

We show our comparison in Figure 6.9. Our improved life expectancy (LE) predictions allow the lump sum to be distributed more efficiently over time. Because the same $600,000 must support income for a longer period, the sustainable annual withdrawal is lower; however, income lasts longer, and the transition to the Age Pension occurs at older ages. In essence, more accurate LE forecasts mean accepting a slightly lower annual withdrawal in exchange for a steadier standard of living and reduced risk of depleting savings prematurely.

For states with shorter LE, the model instead favours higher withdrawals earlier in retirement, leading to an earlier shift to pension reliance. This approach enhances retirees' quality of life during a larger portion of retirement and promotes a more sustainable balance between superannuation and pension resources (Section 6.3.3).

In summary, Figure 6.9 shows three key impacts of incorporating health-informed life expectancy into retirement income projections:

1. **Smaller Drop in Lifestyle Standards**: Compared with the ALT-based projection, the health-informed model generally results in a smoother decline in income rather than a sudden fall when private savings run out. This is visible in the figure as a less pronounced drop in income lines, indicating that retirees can maintain a more consistent standard of living later in life.

2. **Shorter Period on Age Pension**: The figure also shows that the period spent relying solely on the Age Pension is typically shorter under the health-informed model. Because withdrawals are spread more appropriately across the expected lifetime, private savings tend to last longer, reducing the time retirees depend entirely on government support.

3. **Improved Retirement Planning**: These differences highlight the value of using individualised life expectancy estimates. By aligning withdrawal strategies with more realistic longevity expectations, retirees are less likely to experience

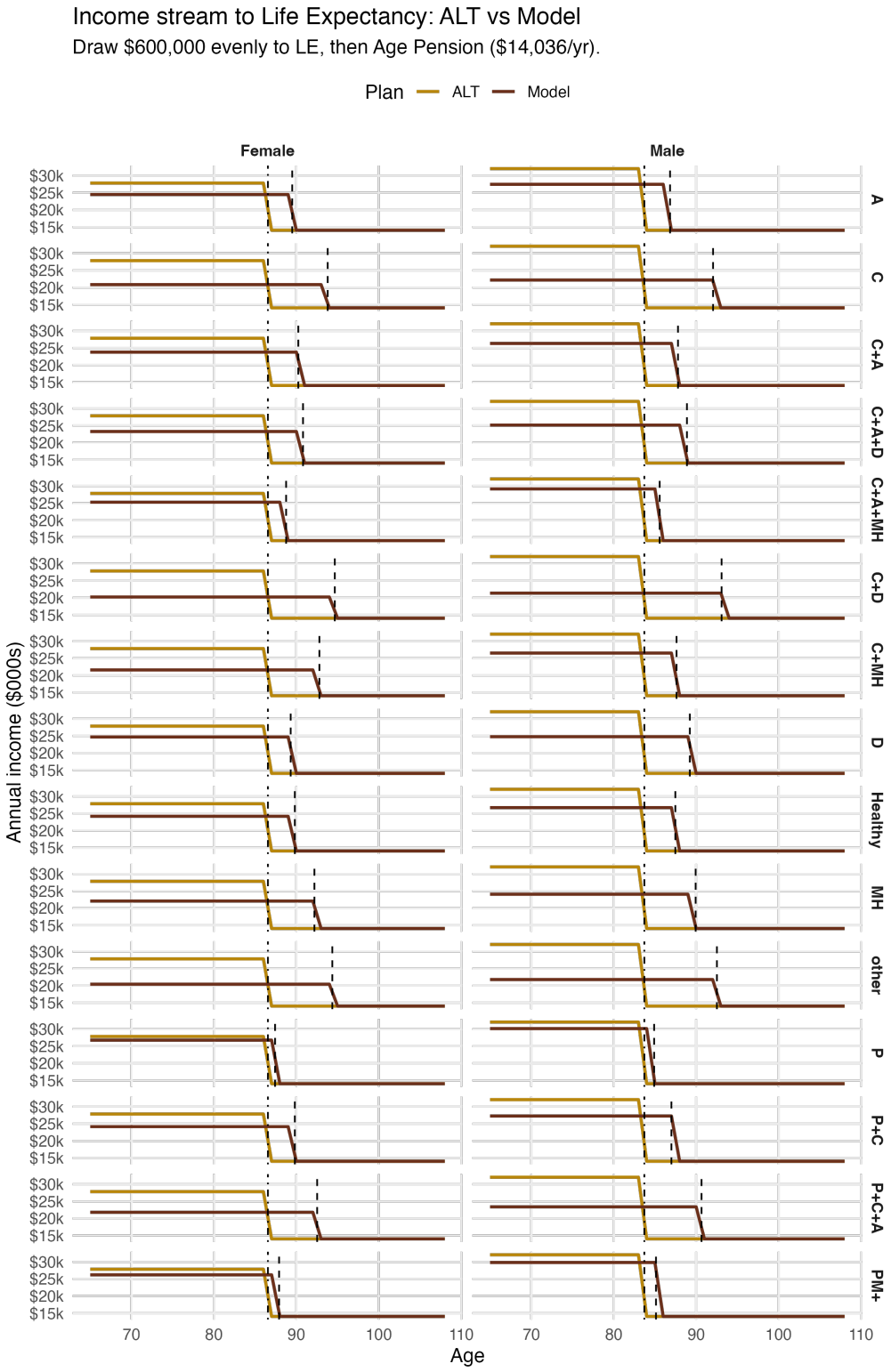


Figure 6.9: Estimated income stream from a $600,000 superannuation balance, assuming an annual Age Pension of $14,036. Income is drawn evenly to life expectancy.

steep income shocks or extended periods on the pension, resulting in more stable and sustainable retirement outcomes.

6.3.3 Implications for Dependence on the Age Pension

The frequency with which individuals live beyond their projected life expectancy is illustrated in Figure 6.10 and Table 6.1). Planning level withdrawals only up to the average life expectancy means that those who live longer will deplete their superannuation sooner and may need to rely on the Age Pension for the remainder of their lives, subject to income and asset eligibility tests. Incorporating life expectancy estimates from the proposed model helps reduce this risk by tailoring projections to individual health profiles, leading to more sustainable retirement income strategies.

Figure 6.10 illustrates how the proposed model improves the accuracy of life expectancy projections by reducing the proportion of individuals who live significantly beyond their expected lifespan. The shaded regions (both blue and red) represents the share of the population outliving standard life expectancy assumptions. Under the proposed model, the portion outliving will be limited to the red portion, indicating fewer individuals depleting their superannuation earlier than expected and subsequently relying on the Age Pension. This shift demonstrates the model's effectiveness in aligning retirement planning with actual longevity patterns, reducing pressure on public support systems and improving individual financial outcomes.

Across all states and genders, the share of who outlive LE falls from **9.9%** under ALT to **6.7%**, a reduction of about **32%** in the number of people outliving the planning horizon (from 1,086,756 to 737,293; Table 6.1). Hence, by using the proposed model, fewer retirees will deplete savings prematurely, thereby reducing years of dependence on the Age Pension.

Importantly, no single planning horizon based on life expectancy can ever drive the "outlive" count to *zero*: there is always a positive survival tail beyond LE. Setting the horizon to an extreme maximum age (e.g., 110–115) would make the count negligible, but it is not optimal, doing so inflates the annuity factor and forces very low withdrawals for everyone. For example, the LE for a 65-year-old female is approximately 21 years (till around age 86), yet a very small proportion live beyond 110–115 years. Planning everyone to such an extreme horizon would roughly halve (or worse) the sustainable annual income, implying a material drop in living standards for the vast majority who will never reach those ages. A more practical approach is to plan withdrawals based on life expectancy rather than an extreme maximum age. The proposed model's improved LE estimates enable allocating a given lump sum more efficiently over the retirement horizon, significantly reducing reliance on the Age Pension.

The proposed model reduces Age Pension reliance for approximately 349,463 individuals. At an average annual pension payment of $14,036, this equates to around $5 billion in savings each year, which could be redirected towards individuals with greater financial need. Importantly, this reduction in reliance does not leave those affected in a worse position. Instead, it reflects more accurate planning and better

alignment between retirement income strategies and individual longevity, allowing retirees to draw on their superannuation for longer before transitioning to public support.

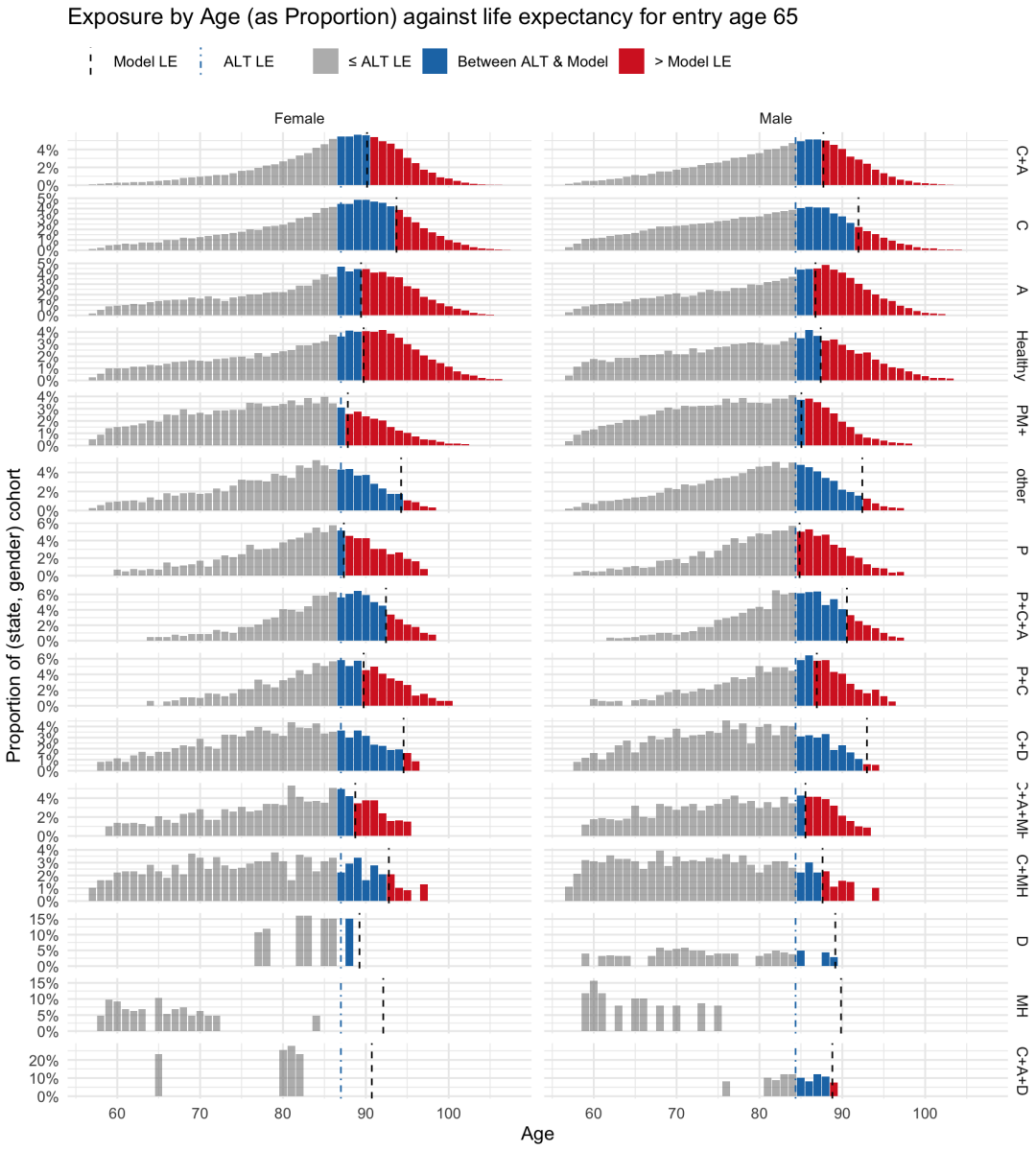


Figure 6.10: Distribution of deaths beyond expected life expectancy. The proposed model reduces the share of individuals who outlive projections, lowering the risk of early asset depletion and extended pension reliance.

Table 6.1: Outliving thresholds by state and gender (ALT vs Higher of Model/ALT)

State	Gender	Count > ALT	Count > Higher	% > ALT	% > Higher
A	Female	21770	21770	11.3%	11.3%
A	Male	38670	38670	20.9%	20.9%
C	Female	101245	77795	3.9%	3.0%
C	Male	336110	101547	11.1%	3.4%
C+A	Female	126684	126684	10.6%	10.6%
C+A	Male	291805	254126	28.8%	25.1%
C+A+D	Female	445	445	21.0%	21.0%
C+A+D	Male	0	0	0.0%	0.0%
C+A+MH	Female	2098	2098	5.0%	5.0%
C+A+MH	Male	6873	3430	12.8%	6.4%
C+D	Female	5391	1572	2.8%	0.8%
C+D	Male	13923	1545	7.2%	0.8%
C+MH	Female	1423	1423	1.5%	1.5%
C+MH	Male	7631	911	4.0%	0.5%
D	Female	274	274	1.2%	1.2%
D	Male	713	713	23.4%	23.4%
Healthy	Female	16072	16072	2.9%	2.9%
Healthy	Male	43483	26219	5.9%	3.6%
MH	Female	0	0	0.0%	0.0%
MH	Male	0	0	0.0%	0.0%
P	Female	2706	2706	7.0%	7.0%
P	Male	4655	4655	13.9%	13.9%
P+C	Female	3202	3202	8.9%	8.9%
P+C	Male	7723	7723	20.2%	20.2%
P+C+A	Female	5007	5007	15.2%	15.2%
P+C+A	Male	9237	9237	32.9%	32.9%
PM+	Female	2954	2954	8.8%	8.8%
PM+	Male	5987	5987	14.3%	14.3%
other	Female	10954	10954	6.0%	6.0%
other	Male	19721	9574	12.4%	6.0%
Total	All	1086756	737293	9.9%	6.7%

6.4 Summary

The proposed dynamic model improves calibration and produces more coherent life-expectancy profiles across health clusters. It changes planning horizons for each cohort in a way that aligns with observed pathways of multi-morbidity. Because the model retains the standard actuarial structure, it integrates seamlessly into annuity valuation and spending rule calculations without adding complexity. In drawdown examples, improved survival inputs trade slightly lower annual withdrawals for longer-lasting private income and a later step-down to the Age Pension.

Key takeaways

- Allowing for transitions narrows unwarranted gaps in life expectancy between clusters, which is a practical indicator of improved model accuracy.
- Using transition-adjusted survival in retirement calculations supports steadier spending paths and reduces the risk that members exhaust savings earlier than planned.
- The ALT benchmark creates *cross subsidies* by age as well as by health: younger entrants are priced too tightly, since ALT does not reflect improvement and deterioration in health seen based on age. Using transition-adjusted survival reduces these subsidies. With finer age bands for transitions, the model captures turning points more precisely, so the reduction in cross subsidies (and the fairness gains in pricing and planning) is even larger.
- The proposed model enables better financial planning while reducing dependence on retirement income, resulting in approximately $5 million in annual Age Pension savings, and improving the quality of life for those individuals.

6.5 Conclusion

This chapter detailed and evaluated the final 15–state Markov chain model (Section 6.1). Transition matrices per age bin ($50-70$, $70-90$, $90+$) capture the dynamic evolution of health in retirement. This model provides life–expectancy (LE) profiles across clusters which show improved performance compared to the ALT.

The monetary analysis (Section 6.3) showed that accurate mortality prediction matter economically. For annuity pricing, the ALT was shown to be underpricing the products for a 65-year retiree, which the industry accounts for by adding substantial margins when pricing. Because the proposed model exhibits lower variability and higher predictive precision, it supports the use of narrower margins, which in turn leads to more accurately priced products and better outcomes for individuals. For lump-sum draw-downs, better LE forecasts spread a fixed balance over a longer horizon, lowering the annual withdrawal but pushing the step-down to the Age Pension to older ages (Figure 6.9). At the system level, planning using the proposed model cuts the fraction who outlive the planning horizon from **9.9%** to **6.7%** a $\sim$32% reduction (1,086,756 to 737,293 (Table 6.1)). This implies that approximately $4.9 billion

p.a. in Age Pension payments could be avoided, though this figure is illustrative and subject to means-testing and behavioural factors.

In summary, modelling mortality on a dynamic health–state basis improves both predictive accuracy and financial relevance. It supports more stable private income streams for retirees, delays premature pension reliance, and offers a practical guideline: use the *higher of* Model and ALT LE estimates for default planning.

Chapter 7

Conclusion

This chapter begins by summarising the main contributions of the thesis in Section 7.1. It then outlines the key features of the proposed model in Section 7.2, followed by the validation procedures and performance evaluation used to determine the final specification in Section 7.3. The practical and financial implications of the model are discussed in Section 7.4 and Section 7.5, respectively. Finally, Section 7.6 presents the study's limitations and directions for future research, and Section 7.7 provides concluding remarks.

7.1 Summary of Contributions

Mortality calculations in Australia are currently based primarily on age and gender. Huang et al. [2023] extended this approach by incorporating demographic variables into mortality modelling for retirees. This thesis advances mortality modelling by further incorporating health-based information. It proposes a model that integrates key medical information and accounts for changes in health status over time.

In this thesis, a dynamic mortality model for Australian retirees has been developed and validated. The model incorporates health data to estimate mortality while allowing for changes in health state. Due to data limitations, the model includes only six medical conditions: Parkinson's disease, use of anti-thrombotic medication, cardiac-related disease, mental health conditions, pain medication use, and diabetes. It defines health states based on the 15 most common conditions in the data, and estimates mortality by age and gender within each state. Survivors update their health status annually using empirically derived transition probabilities. This dynamic modelling approach generates survival inputs that reflect the presence and progression of multi-morbidity over time. Further, this approach reduces artificial dispersion in life expectancy across clusters, while preserving the standard actuarial machinery for life tables, annuities, and present values.

7.1.1 Key Achievements

1. Development of a state-based Markov model that jointly models disease progression and mortality, allowing for the estimation of dynamic survival probabilities that evolve with changes in health status.

2. Integration of nationally representative health and medication data into mortality prediction, a first in the Australian context, enabling the model to capture complex relationships between clinical variables and longevity outcomes.

3. Demonstration of superior predictive accuracy and calibration compared to traditional benchmarks such as the Australian Life Tables (ALT), improving both individual-level predictions and cohort-level forecasts.

4. Translation of predictive improvements into tangible financial impacts, demonstrating how more accurate survival estimates can inform annuity pricing, optimise withdrawal strategies, and reduce reliance on the Age Pension.

7.2 Novelty and Significance of the Dynamic Probabilistic Model

The proposed model differs from a traditional Markov chain model for mortality prediction, as death is not treated as a state within the model. Instead, the model assumes deaths occur prior to any state transitions within a given year, which allows transitions to "death" be dependant on age, gender, and current health state. This better incorporates the complex nature of the mortality predictions. Since we only have annual data records, all health state transitions are modelled to occur once per year.

Given the three-year observational data window, this modelling approach ,which additionally captures state transitions, is more reasonable than trying to directly capture raw death rates. In the latter method, observed deaths are likely to be skewed by near-death state observations, leading to biased mortality estimates. By using transition tables, the proposed model mitigates this issue as it smoothens late-stage clustering effects, providing a more robust and reliable basis for estimating mortality dynamics.

The proposed modelling framework is novel in several key aspects:

1. Dynamic and state-based modelling

2. Integration of national-scale health data

3. Interpretability and practical application

as explained in the following sections.

7.2.1 Dynamic and State-Based Modelling

Conventional mortality models assume that mortality risk remains constant within demographic groups. In contrast, the proposed model treats mortality as a dynamic process that evolves with changes in health state. By explicitly modelling transitions, the framework captures the temporal dimension of mortality risk. This approach

produces survival curves that are sensitive to individual health trajectories and reflect real-world patterns of disease progression.

7.2.2 Integration of National-Scale Health Data

This study is among the first in Australia to integrate national-scale administrative health data into a mortality model. By leveraging millions of individual-level observations, the model captures complex interactions between disease prevalence and mortality outcomes. This scale not only enhances predictive performance but also ensures that the model reflects population-level patterns relevant to policy and product design.

7.2.3 Interpretability and Practical Application

Despite its technical sophistication, the model remains interpretable and practically useful. It can be distilled into a concise set of clinically meaningful variables, enabling its adoption by insurers, superannuation funds, and policymakers. The framework supports the design of simple health questionnaires that, when combined with demographic data, allow for more accurate and equitable mortality risk assessments in real-world applications.

7.3 Validation and Benchmarks

The proposed approach was tested against several other models, using the ALT as the benchmark. The evaluation process assessed both discrimination and calibration. One of the key challenges was the dependence of the mortality function on the specific data split. To ensure a fair comparison, model performance was benchmarked against the ALT using the same data partition. All models outperformed the benchmark, emphasising the importance of integrating medical status into mortality predictions. Deviations in life expectancy were compared across clusters as an aggregated measure of mortality within each group. This allowed us to assess whether the clustering effectively captured meaningful variations in mortality trends, ensuring that the models reflect underlying differences in health outcomes.

7.4 Implications for Practice and Policy

For funds and insurers, improved survival inputs support more stable spending paths and more reliable pricing. When used for level draw-downs strategies, the same balance is distributed over a horizon that reflects likely health trajectories. This reduces the risk of premature asset depletion and delays the point at which private income declines, causing dependence on Age Pension level. It also reduces pressure on public expenditure while ensuring member outcomes remain transparent and defensible.

7.5 Monetary Implications

The societal implications of this work are far-reaching. At the individual level, the model enables retirees to make better-informed decisions about draw-down strategies and annuity purchases. By aligning withdrawal plans with realistic survival expectations, individuals can maintain a more stable standard of living, reduce the risk of premature asset depletion, and avoid sharp income shocks later in life.

At the system level, the model's impact is even more significant. It reduces the proportion of retirees who outlive their planning horizon from 9.9% under ALT-based models to 6.7%, a reduction of roughly 3.2%. This translates into approximately 349,463 fewer individuals prematurely relying on the Age Pension, based on the 2014-2016 dataset. With an average annual pension payment of $14,036, this equates to an estimated reduction in public expenditure of around $5 billion per year. These savings could be redirected towards supporting individuals with greater financial need, improving the sustainability of Australia's retirement income system and easing long-term fiscal pressures on government.

For the insurance and superannuation industries, the model also enables more accurate pricing of longevity-linked products. Because it reduces uncertainty around survival outcomes, providers can price annuities with narrower margins, improving affordability for consumers while maintaining financial soundness. This results in a more equitable distribution of longevity risk and ensures that products are better aligned with individual risk profiles.

Beyond its technical contributions, this research carries broader societal significance. It demonstrates that mortality is not merely a demographic outcome but the culmination of dynamic health trajectories. By recognising this complexity and embedding it within a probabilistic framework, the model provides a foundation for a more nuanced understanding of ageing and longevity.

The approach also has potential applications beyond retirement modelling. It can inform public health strategies, guide resource allocation, and support the design of targeted interventions for at-risk populations. Moreover, by linking mortality modelling with real-world healthcare data, this work paves the way for a new era of evidence-based policymaking that is both data-driven and person-centred.

7.6 Limitations and Future Work

7.6.1 Dataset

The following limitations exist within the MBS and PBS dataset and represent potential areas for improvement as new data become available.

- The current dataset only includes retirees aged 55 and above as of 2011, limiting generalisability to younger cohorts. As the population ages and develops chronic conditions earlier, incorporating pre-retirement health trajectories would extend this research and improve mortality prediction for all individu-

als. Notably, the inclusion of health information remains most crucial for older populations.

- The MBS dataset has missing entries for columns `dialysis_date`, `chemo_proc_date`, `rad_oncology_date`, `neuro_surg_date`, and `spinal_surg_date` which are all influential columns in the mortality of retirees. Incorporating these variables in future data releases would further improve mortality estimation.

- As listed under Section 3.4, many variables contained only missing or zero values and were therefore omitted from this study. Ideally, these data issues will be resolved in future releases, allowing the model to be revisited and expanded to include these variables.

- The PLIDA dataset does not contain individuals not captured within the Medicare system (this also includes non-residents and emigrants along with others who do not meet the eligibility criteria). This exclusion may introduce survivorship bias, particularly in long-term survival estimates.

- Due to constraints on the size and structure of the dataset, date variables were simplified to record only the year rather than the full date. As a result, the constructed `age` variable represents an approximation rather than an individual's true chronological age. This simplification introduces a potential source of deviation between modelled and actual ages, which may slightly affect mortality estimates, particularly near age boundaries where year-based rounding can shift individuals between adjacent age groups. Future work incorporating complete date information would enable more precise age calculations and finer-grained survival estimates.

- The model assumes static demographic features (e.g., sex, year of birth) and does not incorporate other variables such as income or homeownership status examined in prior studies. Integrating these demographic factors would produce a more comprehensive model that combines both demographic and health information.

- Exposure time was limited to three historical years and one test year due to computational and data constraints. Extending the training window would allow for more robust transition estimates, especially for slower-progressing conditions.

7.6.2 Model

- The model assumes that transitions depend only on the most recent state (Markov-property). However, this is a simplifying assumption, as health is influenced by medical history; incorporating longer-term data impacts would improve model accuracy.

- As noted in Siebert et al. [2009], a key limitation of disease-based Markov models, including this one, is the lack of consideration of disease severity. The model only considers the presence of the condition and the implications of that on the mortality, thus implying the mortality is averaged across all those who have the relevant condition. A potential improvement is to incorporate measures of disease severity to estimate mortality more precisely.

- The model focuses on high-level disease categories and does not account for specific drugs or interactions between medications. Due to limited medical expertise and time constraints, this analysis was not undertaken, though it remains a promising direction for future research.

- Transition probabilities were estimated in relatively broad age bins to stabilise estimates given the limited sample size. Ideally, finer-grained bins would be used to capture more nuanced changes in health transitions and improve the fairness and precision of mortality estimation. However, due to data sparsity, narrowing the bands substantially would have resulted in unstable or unreliable estimates. Future work with larger datasets or longer observation windows could support smaller transition intervals without compromising statistical robustness.

- While the transition probabilities are estimated from empirical data, they are not currently smoothed over age or calendar year. This leads to small irregularities in estimated rates. Incorporating smoothing techniques (e.g., penalised splines or logit-LOESS) would improve stability and realism, especially for policy applications.

- Data limitations constrain the model to six diseases, but ideally it should be extended to incorporate additional conditions as data become available. Incorporating the most common causes of death identified in Appendix C would be an ideal extension.

- The current model uses 2011–2016 data and should be rerun on more recent data to ensure that mortality and disease estimates reflect changes in medical exposure and healthcare advancements over the past decade.

- The mortality model assumes independence between health and external shocks (e.g., pandemics or economic downturns). Adding a stochastic or scenario-based layer could capture macro-level risks and their interactions with health states.

- The model does not currently handle competing risks explicitly (e.g., individuals with multiple severe conditions are assigned a single aggregated transition or death risk). A more refined competing-risks framework would enable condition-specific death pathways and allow consideration of disease severity in combination.

7.7 Concluding Remarks

This thesis demonstrates that a carefully designed dynamic mortality model can materially improve retirement planning and pension allocation. The proposed model adopts a non-traditional, state-based approach that achieves an effective trade-off between practicality, accuracy, and calibration. It integrates cleanly with standard life-table methods and remains simple to implement, transparent, and auditable. At the *individual* level, health-informed survival inputs extend planning horizons, smooth income paths, reduce premature reliance on the Age Pension, and support fairer pricing of longevity-linked products through narrower and more credible margins. At the *national* level, more accurate survival forecasts reduce premature reliance on the Age Pension, improve the targeting of public support, and strengthen fiscal sustainability, while helping policymakers and funds design products and guidelines that are equitable, defensible, and aligned with observed health trajectories. Overall, this work demonstrates that incorporating available health information within a dynamic probabilistic framework can deliver better retiree outcomes and measurable system-wide gains in a transparent and computationally efficient manner.

Owing to the model's relatively simple structure, the accompanying questionnaire would only need to record whether an individual currently has any of six conditions: **Parkinson's**, **Diabetes**, **Cardiac**, **Anti-thrombotic**, **Pain medication**, or **Mental health**. The questionnaire could also ask whether the individual has experienced any of these conditions within the past one or two years, helping to account for short-term lapses in health status. Extending the timeframe further, however, would likely reduce accuracy, as such lapses are not explicitly modelled and the annual structure of the data makes transition probabilities less reliable for conditions persisting over longer periods. As a result, mortality estimates based on conditions that no longer reflect a person's current health state would lose much of their predictive value.

Appendix A

Life table for final model

The transition matrices can be integrated onto a table calculating ${}_tq_x$ for each state. Though all the relevant tables have been generated to keep the thesis concise, the relevant tables have been provided for conditions cardiac (C) and pain medication (PM). The remaining tables follow the same structure and are available on request.

A.1 How to Read the ${}_tq_x$ Tables

The tables report cumulative death probabilities over the next t years for an individual aged x at the start of the interval, conditioning on the stated *start cluster* and gender. They are intended to be read exactly like a life table, but the risks reflect transition dynamics across health states ("transitions then death") from the fitted multi-state model.

What Each Element Means

- **Rows (x):** Exact age at valuation. For example, the row labelled $x = 70$ pertains to lives aged 70 at time 0 in the specified cluster.
- **Columns (_${}_tq_x$):** Cumulative probability of death *within* t years from age x, i.e.

$$ {}_tq_x \;=\; \Pr(T \le t \mid \text{alive at age } x,\ \text{start cluster, gender}), $$

  where T is the future lifetime measured in years. Thus ${}_1q_x$ is 1-year mortality, ${}_5q_x$ is 5-year mortality, and so on. By construction, ${}_tq_x$ is non-decreasing in t.
- **"Transitions then death":** The probability is taken over all paths of the underlying Markov model in which the life may transition between non-death states (e.g., condition clusters) before entering the absorbing death state within t years.
- **Start cluster & gender:** Each table conditions on the label shown in the caption (e.g., `C` for cardiac; `Pain med with any condition`) and on gender. Do not mix rows across tables when analyzing a single cohort.

- **Blanks at the right edge:** If later-horizon entries are blank, the horizon exceeds the estimation support for that $(x, \text{cluster}, \text{gender})$ combination; no values are imputed.

Quick Conversions and Derived Quantities

- **t-year survival:** ${}_tp_x = 1 - {}_tq_x$.
- **Year-by-year (incremental) death probability between $t-1$ and t:**

$$q_{x+t-1:x+t} = {}_tq_x - {}_{t-1}q_x.$$

- **Conditional survival from s to t ($0 < s < t$):**

$${}_{t-s}p_{x+s} = \frac{1 - {}_tq_x}{1 - {}_sq_x}.$$

- **Conditional t-year mortality given survival to s:**

$${}_{t-s|s}q_x = \frac{{}_tq_x - {}_sq_x}{1 - {}_sq_x}.$$

How to Read a Cell in the Provided Tables

Pick the appropriate table (cluster+gender), the row for the attained age x, and the column ${}_tq_x$. For instance, in Table A.1 the entry on row $x = 70$ and column ${}_{10}q_{70}$ is the probability that a 70-year-old female in cluster `C` dies within the next 10 years, allowing for transitions among non-death states beforehand. If that cell were 0.1019, it would mean a 10.19% risk of death by age 80 under the model assumptions; the corresponding 10-year survival would be $1 - 0.1019 = 0.8981$.

Table A.1: ${}_tq_x$ (Probability of death within t years, provided the gender = Female, Initial state = C). Rounded to 2DP, full precision available upon request.

x	$1q_x$	$2q_x$	$3q_x$	$4q_x$	$5q_x$	$6q_x$	$7q_x$	$8q_x$	$9q_x$	$10q_x$	$11q_x$	$12q_x$	$13q_x$	$14q_x$	$15q_x$	$16q_x$	$17q_x$	$18q_x$	$19q_x$	$20q_x$	$21q_x$	$22q_x$	$23q_x$	$24q_x$	$25q_x$	$26q_x$	$27q_x$	$28q_x$	$29q_x$	$30q_x$	$31q_x$	$32q_x$	$33q_x$	$34q_x$	$35q_x$	$36q_x$	$37q_x$	$38q_x$	$39q_x$	$40q_x$	$41q_x$	$42q_x$	$43q_x$	$44q_x$	$45q_x$	$46q_x$	$47q_x$	$48q_x$	$49q_x$	$50q_x$
60	0.01	0.01	0.02	0.02	0.03	0.03	0.04	0.04	0.05	0.05	0.06	0.06	0.07	0.07	0.08	0.09	0.09	0.10	0.11	0.12	0.13	0.14	0.16	0.17	0.19	0.20	0.22	0.24	0.27	0.30	0.33	0.37	0.41	0.45	0.51	0.57	0.63	0.67	0.72	0.75	0.79	0.81	0.84	0.86	0.88	0.89	0.91	0.92	0.93	0.94
61	0.01	0.01	0.02	0.02	0.03	0.03	0.04	0.04	0.05	0.05	0.06	0.06	0.07	0.08	0.08	0.09	0.10	0.11	0.12	0.13	0.14	0.15	0.17	0.18	0.20	0.22	0.24	0.26	0.30	0.33	0.37	0.41	0.45	0.51	0.57	0.62	0.67	0.72	0.75	0.78	0.81	0.84	0.86	0.88	0.89	0.91	0.92	0.93	0.94
62	0.01	0.01	0.02	0.03	0.03	0.04	0.04	0.04	0.05	0.06	0.06	0.07	0.07	0.08	0.09	0.10	0.11	0.12	0.12	0.14	0.15	0.16	0.18	0.20	0.22	0.24	0.26	0.29	0.33	0.37	0.41	0.45	0.51	0.57	0.62	0.67	0.72	0.75	0.78	0.81	0.84	0.86	0.88	0.89	0.91	0.92	0.93	0.94	
63	0.01	0.01	0.02	0.03	0.03	0.04	0.04	0.05	0.05	0.06	0.06	0.07	0.08	0.09	0.09	0.10	0.11	0.12	0.13	0.15	0.16	0.18	0.19	0.21	0.23	0.26	0.29	0.33	0.36	0.40	0.45	0.50	0.57	0.62	0.67	0.71	0.75	0.78	0.81	0.84	0.86	0.87	0.89	0.90	0.92	0.93	0.94		
64	0.01	0.02	0.02	0.03	0.03	0.04	0.04	0.05	0.06	0.06	0.07	0.08	0.08	0.09	0.10	0.11	0.12	0.13	0.15	0.16	0.18	0.19	0.21	0.23	0.26	0.29	0.33	0.36	0.40	0.45	0.50	0.57	0.62	0.67	0.71	0.75	0.78	0.81	0.84	0.86	0.87	0.89	0.90	0.92	0.93	0.94			
65	0.01	0.02	0.02	0.03	0.03	0.04	0.05	0.05	0.06	0.07	0.07	0.08	0.09	0.10	0.11	0.12	0.13	0.14	0.16	0.17	0.19	0.21	0.23	0.25	0.29	0.32	0.36	0.40	0.45	0.50	0.56	0.62	0.67	0.71	0.75	0.78	0.81	0.83	0.86	0.87	0.89	0.90	0.92	0.93	0.94				
66	0.01	0.02	0.02	0.03	0.04	0.04	0.05	0.06	0.06	0.07	0.08	0.09	0.10	0.10	0.11	0.13	0.14	0.16	0.17	0.19	0.21	0.23	0.25	0.29	0.32	0.36	0.40	0.44	0.50	0.56	0.62	0.67	0.71	0.75	0.78	0.81	0.83	0.86	0.87	0.89	0.90	0.92	0.93	0.94					
67	0.01	0.02	0.03	0.03	0.04	0.05	0.05	0.06	0.07	0.07	0.08	0.09	0.10	0.11	0.12	0.14	0.15	0.17	0.18	0.20	0.23	0.25	0.28	0.32	0.36	0.40	0.44	0.50	0.56	0.62	0.67	0.71	0.75	0.78	0.81	0.83	0.85	0.87	0.89	0.90	0.92	0.93	0.94						
68	0.01	0.02	0.03	0.04	0.04	0.05	0.06	0.06	0.07	0.08	0.09	0.10	0.11	0.12	0.13	0.15	0.17	0.18	0.20	0.22	0.25	0.28	0.32	0.36	0.40	0.44	0.50	0.56	0.62	0.67	0.71	0.75	0.78	0.81	0.83	0.85	0.87	0.89	0.90	0.92	0.93	0.94							
69	0.01	0.02	0.03	0.04	0.05	0.05	0.06	0.07	0.08	0.09	0.10	0.11	0.12	0.13	0.15	0.17	0.18	0.20	0.22	0.25	0.28	0.32	0.36	0.40	0.44	0.50	0.56	0.62	0.67	0.71	0.75	0.78	0.81	0.83	0.85	0.87	0.89	0.90	0.91	0.93	0.93								
70	0.01	0.02	0.04	0.05	0.06	0.07	0.07	0.08	0.09	0.10	0.11	0.13	0.14	0.15	0.17	0.19	0.21	0.23	0.26	0.29	0.33	0.36	0.40	0.45	0.50	0.57	0.62	0.67	0.71	0.75	0.78	0.81	0.83	0.85	0.87	0.89	0.90	0.92	0.93	0.94									
71	0.01	0.02	0.04	0.05	0.06	0.07	0.08	0.09	0.10	0.11	0.12	0.14	0.15	0.17	0.19	0.20	0.23	0.25	0.29	0.32	0.36	0.40	0.45	0.50	0.56	0.62	0.67	0.71	0.75	0.78	0.81	0.83	0.85	0.87	0.89	0.90	0.92	0.93	0.93										
72	0.02	0.03	0.05	0.06	0.07	0.08	0.09	0.10	0.11	0.12	0.13	0.15	0.17	0.18	0.20	0.23	0.25	0.28	0.32	0.36	0.40	0.44	0.50	0.56	0.62	0.67	0.71	0.75	0.78	0.81	0.83	0.85	0.87	0.89	0.90	0.91	0.93	0.93											
73	0.02	0.03	0.05	0.06	0.07	0.08	0.09	0.10	0.11	0.12	0.14	0.16	0.18	0.20	0.22	0.25	0.28	0.32	0.36	0.40	0.44	0.50	0.56	0.62	0.67	0.71	0.75	0.78	0.81	0.83	0.85	0.87	0.89	0.90	0.91	0.93	0.93												
74	0.02	0.03	0.05	0.07	0.08	0.09	0.10	0.11	0.13	0.14	0.16	0.18	0.20	0.22	0.24	0.28	0.32	0.35	0.39	0.44	0.50	0.56	0.62	0.67	0.71	0.75	0.78	0.81	0.83	0.85	0.87	0.89	0.90	0.91	0.92	0.93													
75	0.02	0.04	0.06	0.07	0.08	0.10	0.11	0.12	0.14	0.16	0.17	0.19	0.22	0.24	0.28	0.32	0.35	0.39	0.44	0.50	0.56	0.62	0.66	0.71	0.75	0.78	0.81	0.83	0.85	0.87	0.89	0.90	0.91	0.92	0.93														
76	0.02	0.04	0.06	0.07	0.09	0.10	0.12	0.13	0.15	0.17	0.19	0.21	0.24	0.27	0.31	0.35	0.39	0.43	0.49	0.56	0.61	0.66	0.71	0.74	0.78	0.80	0.83	0.85	0.87	0.89	0.90	0.91	0.92	0.93															
77	0.02	0.04	0.06	0.08	0.09	0.11	0.13	0.14	0.16	0.18	0.21	0.23	0.27	0.30	0.34	0.38	0.43	0.49	0.55	0.61	0.66	0.70	0.74	0.77	0.80	0.83	0.85	0.87	0.89	0.90	0.91	0.92	0.93																
78	0.02	0.04	0.06	0.08	0.10	0.12	0.14	0.15	0.17	0.20	0.22	0.26	0.30	0.34	0.38	0.43	0.48	0.55	0.61	0.66	0.70	0.74	0.77	0.80	0.83	0.85	0.87	0.88	0.90	0.91	0.92	0.93																	
79	0.03	0.05	0.08	0.10	0.11	0.13	0.15	0.17	0.20	0.22	0.26	0.30	0.34	0.38	0.42	0.48	0.55	0.61	0.66	0.70	0.74	0.77	0.80	0.83	0.85	0.87	0.88	0.90	0.91	0.92	0.93																		
80	0.03	0.05	0.08	0.10	0.13	0.14	0.17	0.19	0.22	0.25	0.29	0.33	0.37	0.42	0.48	0.54	0.60	0.65	0.70	0.74	0.77	0.80	0.83	0.85	0.87	0.88	0.90	0.91	0.92	0.93																			
81	0.03	0.06	0.09	0.11	0.13	0.16	0.18	0.21	0.24	0.28	0.32	0.37	0.41	0.47	0.54	0.60	0.65	0.69	0.73	0.77	0.80	0.82	0.85	0.87	0.88	0.90	0.91	0.92	0.93																				
82	0.04	0.07	0.10	0.13	0.15	0.18	0.21	0.24	0.28	0.32	0.36	0.41	0.47	0.54	0.60	0.65	0.69	0.73	0.77	0.80	0.82	0.85	0.86	0.88	0.90	0.91	0.92	0.93																					
83	0.04	0.07	0.11	0.14	0.17	0.19	0.23	0.27	0.31	0.36	0.40	0.47	0.53	0.59	0.64	0.69	0.73	0.76	0.79	0.82	0.84	0.86	0.88	0.90	0.91	0.92	0.93																						
84	0.05	0.08	0.12	0.16	0.19	0.22	0.27	0.31	0.35	0.40	0.46	0.53	0.59	0.64	0.69	0.73	0.76	0.79	0.82	0.84	0.86	0.88	0.89	0.91	0.92	0.93																							
85	0.05	0.09	0.14	0.17	0.21	0.25	0.29	0.34	0.39	0.45	0.52	0.58	0.64	0.68	0.72	0.76	0.79	0.82	0.84	0.86	0.88	0.89	0.91	0.92	0.93																								
86	0.05	0.10	0.14	0.19	0.24	0.28	0.33	0.38	0.44	0.51	0.57	0.63	0.68	0.72	0.75	0.78	0.81	0.84	0.86	0.87	0.89	0.90	0.92	0.93																									
87	0.06	0.11	0.17	0.22	0.26	0.31	0.37	0.43	0.50	0.57	0.62	0.67	0.71	0.75	0.78	0.81	0.83	0.85	0.87	0.89	0.90	0.92	0.93																										
88	0.07	0.13	0.20	0.25	0.30	0.36	0.42	0.49	0.56	0.62	0.67	0.71	0.75	0.78	0.81	0.83	0.85	0.87	0.89	0.90	0.91	0.92																											
89	0.08	0.17	0.24	0.30	0.35	0.42	0.49	0.56	0.61	0.66	0.71	0.74	0.78	0.80	0.83	0.85	0.87	0.88	0.90	0.91	0.92																												
90	0.09	0.20	0.30	0.36	0.42	0.50	0.56	0.62	0.67	0.71	0.75	0.78	0.81	0.83	0.85	0.87	0.89	0.90	0.91	0.92																													
91	0.10	0.21	0.32	0.39	0.47	0.54	0.60	0.65	0.70	0.73	0.77	0.80	0.82	0.84	0.86	0.88	0.90	0.91	0.92																														
92	0.11	0.23	0.35	0.44	0.51	0.58	0.63	0.68	0.72	0.75	0.79	0.81	0.84	0.86	0.87	0.89	0.90	0.92																															
93	0.12	0.26	0.39	0.48	0.55	0.61	0.66	0.70	0.74	0.77	0.80	0.83	0.85	0.87	0.88	0.90	0.91																															
94	0.15	0.31	0.43	0.51	0.58	0.63	0.68	0.72	0.76	0.79	0.81	0.84	0.86	0.88	0.89	0.90																															
95	0.15	0.33	0.45	0.53	0.59	0.64	0.69	0.73	0.76	0.79	0.82	0.84	0.86	0.88	0.89																															
96	0.18	0.36	0.48	0.55	0.61	0.66	0.71	0.74	0.78	0.80	0.83	0.85	0.87	0.89																															
97	0.19	0.36	0.48	0.56	0.62	0.67	0.71	0.75	0.78	0.81	0.83	0.85	0.87																															
98	0.19	0.37	0.49	0.56	0.62	0.67	0.71	0.75	0.78	0.81	0.83	0.85																															
99	0.19	0.37	0.49	0.57	0.62	0.67	0.71	0.75	0.78	0.81	0.83																																
100	0.20	0.38	0.50	0.57	0.63	0.68	0.72	0.75	0.79	0.81																																	
101	0.21	0.39	0.51	0.58	0.63	0.68	0.72	0.76	0.79																																		
102	0.21	0.39	0.51	0.58	0.63	0.68	0.72	0.76																																			
103	0.21	0.39	0.51	0.58	0.63	0.68	0.72																																				
104	0.22	0.40	0.51	0.58	0.64	0.68																																					
105	0.22	0.40	0.51	0.58	0.64																																						
106	0.22	0.40	0.51	0.58																																							
107	0.22	0.40	0.51																																							
108	0.22	0.40																																								
109	0.22																																								
110																																									

Table A.2: ${}_tq_x$ (P[death within t years]), transitions then death; start cluster: C, gender: Male). Rounded to 2DP, full precision available upon request.

x	$1q_x$	$2q_x$	$3q_x$	$4q_x$	$5q_x$	$6q_x$	$7q_x$	$8q_x$	$9q_x$	$10q_x$	$11q_x$	$12q_x$	$13q_x$	$14q_x$	$15q_x$	$16q_x$	$17q_x$	$18q_x$	$19q_x$	$20q_x$	$21q_x$	$22q_x$	$23q_x$	$24q_x$	$25q_x$	$26q_x$	$27q_x$	$28q_x$	$29q_x$	$30q_x$	$31q_x$	$32q_x$	$33q_x$	$34q_x$	$35q_x$	$36q_x$	$37q_x$	$38q_x$	$39q_x$	$40q_x$	$41q_x$	$42q_x$	$43q_x$	$44q_x$	$45q_x$	$46q_x$	$47q_x$	$48q_x$	$49q_x$	$50q_x$
60	0.01	0.02	0.02	0.03	0.04	0.05	0.05	0.06	0.07	0.07	0.08	0.09	0.09	0.10	0.10	0.11	0.12	0.13	0.15	0.16	0.17	0.19	0.20	0.22	0.24	0.26	0.29	0.31	0.34	0.37	0.40	0.45	0.49	0.53	0.57	0.61	0.65	0.68	0.71	0.73	0.76	0.78	0.80	0.82	0.83	0.85	0.86	0.87	0.88	0.89	0.90
61	0.01	0.02	0.03	0.04	0.04	0.05	0.06	0.06	0.07	0.08	0.09	0.09	0.10	0.11	0.12	0.13	0.13	0.14	0.16	0.17	0.18	0.20	0.22	0.24	0.26	0.28	0.31	0.33	0.37	0.40	0.44	0.49	0.53	0.57	0.61	0.65	0.68	0.71	0.73	0.76	0.78	0.80	0.82	0.83	0.85	0.86	0.87	0.88	0.89	0.90	
62	0.01	0.02	0.03	0.04	0.04	0.05	0.06	0.07	0.07	0.08	0.09	0.10	0.11	0.12	0.13	0.14	0.15	0.17	0.18	0.20	0.22	0.24	0.26	0.28	0.31	0.33	0.37	0.40	0.44	0.49	0.53	0.57	0.61	0.64	0.68	0.71	0.73	0.76	0.78	0.80	0.81	0.83	0.85	0.86	0.87	0.88	0.89	0.90		
63	0.01	0.02	0.03	0.04	0.05	0.05	0.06	0.07	0.08	0.09	0.09	0.10	0.11	0.12	0.14	0.15	0.16	0.18	0.19	0.21	0.23	0.26	0.28	0.30	0.33	0.36	0.40	0.44	0.48	0.53	0.57	0.61	0.64	0.67	0.70	0.73	0.75	0.78	0.80	0.81	0.83	0.84	0.86	0.87	0.88	0.89	0.90			
64	0.01	0.02	0.03	0.04	0.05	0.05	0.06	0.07	0.08	0.09	0.10	0.11	0.12	0.13	0.14	0.16	0.17	0.19	0.21	0.23	0.25	0.27	0.30	0.32	0.36	0.39	0.44	0.48	0.53	0.57	0.61	0.64	0.67	0.70	0.73	0.75	0.77	0.79	0.81	0.83	0.84	0.86	0.87	0.88	0.89	0.90				
65	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.08	0.09	0.10	0.11	0.13	0.14	0.15	0.17	0.19	0.20	0.22	0.25	0.27	0.29	0.32	0.36	0.39	0.43	0.48	0.52	0.56	0.60	0.64	0.67	0.70	0.73	0.75	0.77	0.79	0.81	0.83	0.84	0.86	0.87	0.88	0.89	0.90					
66	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10	0.11	0.12	0.14	0.15	0.16	0.18	0.20	0.22	0.24	0.27	0.29	0.32	0.35	0.39	0.43	0.48	0.52	0.56	0.60	0.64	0.67	0.70	0.73	0.75	0.77	0.79	0.81	0.83	0.84	0.86	0.87	0.88	0.89	0.90						
67	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10	0.12	0.13	0.14	0.16	0.18	0.19	0.22	0.24	0.26	0.29	0.31	0.35	0.38	0.43	0.47	0.52	0.56	0.60	0.64	0.67	0.70	0.72	0.75	0.77	0.79	0.81	0.83	0.84	0.86	0.87	0.88	0.89	0.90							
68	0.02	0.02	0.04	0.05	0.06	0.07	0.08	0.09	0.10	0.11	0.13	0.14	0.15	0.17	0.19	0.21	0.24	0.26	0.28	0.31	0.35	0.38	0.42	0.47	0.52	0.56	0.60	0.63	0.67	0.70	0.72	0.75	0.77	0.79	0.81	0.83	0.84	0.86	0.87	0.88	0.89	0.90								
69	0.02	0.03	0.05	0.06	0.07	0.08	0.09	0.10	0.11	0.13	0.14	0.16	0.17	0.19	0.21	0.24	0.26	0.28	0.31	0.35	0.38	0.42	0.47	0.52	0.56	0.60	0.63	0.67	0.70	0.72	0.75	0.77	0.79	0.81	0.82	0.84	0.85	0.87	0.88	0.89	0.90									
70	0.02	0.03	0.06	0.07	0.08	0.10	0.11	0.12	0.14	0.15	0.16	0.18	0.20	0.22	0.24	0.27	0.29	0.32	0.36	0.39	0.43	0.48	0.52	0.57	0.60	0.64	0.67	0.70	0.73	0.75	0.77	0.79	0.81	0.83	0.84	0.86	0.87	0.88	0.89	0.90										
71	0.02	0.04	0.06	0.08	0.09	0.10	0.12	0.13	0.14	0.16	0.18	0.20	0.22	0.24	0.26	0.29	0.32	0.35	0.39	0.43	0.48	0.52	0.56	0.60	0.64	0.67	0.70	0.73	0.75	0.77	0.79	0.81	0.83	0.84	0.86	0.87	0.88	0.89	0.90											
72	0.02	0.04	0.07	0.08	0.10	0.11	0.12	0.14	0.15	0.17	0.19	0.21	0.24	0.26	0.28	0.31	0.35	0.38	0.43	0.47	0.52	0.56	0.60	0.64	0.67	0.70	0.72	0.75	0.77	0.79	0.81	0.83	0.84	0.86	0.87	0.88	0.89	0.90												
73	0.02	0.04	0.07	0.09	0.10	0.12	0.13	0.15	0.17	0.18	0.21	0.23	0.25	0.28	0.31	0.34	0.38	0.42	0.47	0.52	0.56	0.60	0.63	0.67	0.70	0.72	0.75	0.77	0.79	0.81	0.83	0.84	0.86	0.87	0.88	0.89	0.90													
74	0.02	0.04	0.08	0.10	0.11	0.13	0.14	0.16	0.18	0.20	0.23	0.25	0.28	0.30	0.34	0.38	0.42	0.47	0.51	0.56	0.60	0.63	0.66	0.69	0.72	0.75	0.77	0.79	0.81	0.82	0.84	0.85	0.87	0.88	0.89	0.90														
75	0.03	0.05	0.08	0.10	0.12	0.14	0.16	0.17	0.20	0.22	0.25	0.27	0.30	0.34	0.37	0.42	0.46	0.51	0.55	0.59	0.63	0.66	0.69	0.72	0.74	0.77	0.79	0.81	0.82	0.84	0.85	0.87	0.88	0.89	0.90															
76	0.03	0.05	0.09	0.11	0.13	0.15	0.17	0.19	0.21	0.24	0.26	0.29	0.33	0.37	0.41	0.46	0.51	0.55	0.59	0.63	0.66	0.69	0.72	0.74	0.76	0.79	0.80	0.82	0.84	0.85	0.86	0.88	0.89	0.90																
77	0.03	0.06	0.09	0.11	0.13	0.16	0.18	0.21	0.23	0.26	0.29	0.32	0.36	0.40	0.45	0.50	0.54	0.58	0.62	0.65	0.68	0.71	0.74	0.76	0.78	0.80	0.82	0.84	0.85	0.86	0.88	0.89	0.90																	
78	0.03	0.06	0.10	0.12	0.15	0.17	0.20	0.22	0.25	0.28	0.32	0.35	0.40	0.45	0.49	0.54	0.58	0.62	0.65	0.68	0.71	0.74	0.76	0.78	0.80	0.82	0.83	0.85	0.86	0.87	0.89	0.90																		
79	0.04	0.07	0.11	0.14	0.16	0.19	0.22	0.24	0.27	0.31	0.35	0.39	0.44	0.49	0.53	0.58	0.61	0.65	0.68	0.71	0.73	0.76	0.78	0.80	0.82	0.83	0.85	0.86	0.87	0.88	0.89																			
80	0.04	0.08	0.12	0.15	0.18	0.21	0.23	0.26	0.30	0.34	0.39	0.44	0.49	0.53	0.57	0.61	0.65	0.68	0.71	0.73	0.76	0.78	0.80	0.82	0.83	0.85	0.86	0.87	0.88	0.89																				
81	0.05	0.08	0.13	0.17	0.19	0.22	0.25	0.29	0.33	0.38	0.43	0.48	0.52	0.57	0.61	0.64	0.67	0.70	0.73	0.75	0.77	0.79	0.81	0.83	0.84	0.86	0.87	0.88	0.89																					
82	0.05	0.10	0.15	0.18	0.21	0.24	0.28	0.32	0.37	0.42	0.47	0.52	0.56	0.60	0.64	0.67	0.70	0.72	0.75	0.77	0.79	0.81	0.83	0.84	0.86	0.87	0.88	0.89																						
83	0.05	0.10	0.15	0.19	0.22	0.27	0.30	0.36	0.41	0.46	0.51	0.55	0.59	0.63	0.66	0.69	0.72	0.74	0.77	0.79	0.81	0.82	0.84	0.85	0.87	0.88	0.89																							
84	0.06	0.11	0.16	0.21	0.25	0.29	0.34	0.40	0.45	0.50	0.54	0.58	0.62	0.65	0.69	0.71	0.74	0.76	0.78	0.80	0.82	0.84	0.85	0.86	0.88	0.89																								
85	0.07	0.12	0.18	0.23	0.27	0.33	0.38	0.44	0.48	0.53	0.57	0.61	0.65	0.68	0.71	0.73	0.76	0.78	0.80	0.82	0.83	0.85	0.86	0.87	0.88																									
86	0.07	0.13	0.20	0.25	0.31	0.37	0.42	0.47	0.52	0.56	0.60	0.64	0.67	0.70	0.73	0.75	0.77	0.79	0.81	0.83	0.84	0.86	0.87	0.88																										
87	0.08	0.15	0.22	0.28	0.35	0.40	0.46	0.51	0.55	0.59	0.63	0.66	0.69	0.72	0.74	0.77	0.79	0.81	0.82	0.84	0.85	0.87	0.88																											
88	0.08	0.15	0.24	0.31	0.37	0.43	0.48	0.53	0.57	0.61	0.65	0.68	0.71	0.73	0.76	0.78	0.80	0.82	0.83	0.85	0.86	0.87																												
89	0.10	0.22	0.31	0.38	0.43	0.49	0.53	0.58	0.61	0.65	0.68	0.71	0.73	0.76	0.78	0.80	0.82	0.83	0.85	0.86	0.87																												
90	0.11	0.24	0.35	0.42	0.47	0.52	0.57	0.61	0.64	0.68	0.70	0.73	0.75	0.78	0.80	0.81	0.83	0.85	0.86	0.87																												
91	0.14	0.28	0.38	0.45	0.50	0.55	0.59	0.63	0.66	0.69	0.72	0.74	0.77	0.79	0.81	0.82	0.84	0.85	0.87																												
92	0.16	0.30	0.41	0.47	0.52	0.56	0.60	0.64	0.67	0.70	0.73	0.75	0.78	0.80	0.81	0.83	0.84	0.86																												
93	0.19	0.33	0.43	0.49	0.54	0.59	0.62	0.66	0.69	0.72	0.74	0.76	0.79	0.80	0.82	0.84	0.85																												
94	0.19	0.33	0.44	0.50	0.55	0.59	0.63	0.66	0.69	0.72	0.74	0.77	0.79	0.81	0.82	0.84																												
95	0.20	0.34	0.44	0.50	0.55	0.59	0.63	0.66	0.69	0.72	0.75	0.77	0.79	0.81	0.82																												
96	0.20	0.35	0.46	0.51	0.56	0.60	0.64	0.67	0.70	0.73	0.75	0.77	0.79	0.81																												
97	0.23	0.38	0.48	0.54	0.58	0.62	0.65	0.68	0.71	0.74	0.76	0.78	0.80																												
98	0.24	0.39	0.48	0.54	0.58	0.62	0.65	0.69	0.71	0.74	0.76	0.78																												
99	0.24	0.39	0.49	0.54	0.58	0.62	0.65	0.69	0.71	0.74	0.76																													
100	0.24	0.39	0.49	0.54	0.58	0.62	0.66	0.69	0.71	0.74																														
101	0.26	0.40	0.49	0.54	0.59	0.62	0.66	0.69	0.72																															
102	0.26	0.41	0.49	0.54	0.59	0.62	0.66	0.69																																
103	0.26	0.41	0.49	0.54	0.59	0.62	0.66																																	
104	0.26	0.41	0.49	0.54	0.59	0.62																																		
105	0.26	0.41	0.49	0.54	0.59																																			
106	0.26	0.41	0.49	0.54																																				
107	0.26	0.41	0.49																																					
108	0.26	0.41																																						
109	0.26																																							
110																																							

Table A.3: ${}_tq_x$ (Probability of death within t years, provided the gender = Male, Initial state = C). Rounded to 2DP, full precision available upon request.

x	$1q_x$	$2q_x$	$3q_x$	$4q_x$	$5q_x$	$6q_x$	$7q_x$	$8q_x$	$9q_x$	$10q_x$	$11q_x$	$12q_x$	$13q_x$	$14q_x$	$15q_x$	$16q_x$	$17q_x$	$18q_x$	$19q_x$	$20q_x$	$21q_x$	$22q_x$	$23q_x$	$24q_x$	$25q_x$	$26q_x$	$27q_x$	$28q_x$	$29q_x$	$30q_x$	$31q_x$	$32q_x$	$33q_x$	$34q_x$	$35q_x$	$36q_x$	$37q_x$	$38q_x$	$39q_x$	$40q_x$	$41q_x$	$42q_x$	$43q_x$	$44q_x$	$45q_x$	$46q_x$	$47q_x$	$48q_x$	$49q_x$	$50q_x$
60	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10	0.11	0.12	0.13	0.14	0.15	0.16	0.17	0.18	0.19	0.21	0.23	0.25	0.27	0.29	0.32	0.35	0.38	0.42	0.46	0.50	0.53	0.57	0.60	0.64	0.68	0.71	0.75	0.77	0.80	0.82	0.84	0.86	0.87	0.88	0.90	0.91	0.92	0.92	0.93
61	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10	0.11	0.12	0.12	0.13	0.14	0.15	0.16	0.17	0.19	0.20	0.22	0.24	0.26	0.29	0.31	0.34	0.38	0.42	0.46	0.49	0.53	0.57	0.60	0.64	0.68	0.71	0.74	0.77	0.80	0.82	0.84	0.85	0.87	0.88	0.89	0.91	0.92	0.92	0.93
62	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10	0.11	0.12	0.13	0.14	0.14	0.15	0.17	0.18	0.19	0.21	0.24	0.26	0.28	0.31	0.34	0.37	0.41	0.45	0.49	0.53	0.56	0.60	0.64	0.67	0.71	0.74	0.77	0.79	0.82	0.84	0.85	0.87	0.88	0.89	0.90	0.91	0.92	0.92	0.93
63	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10	0.11	0.12	0.13	0.14	0.15	0.16	0.17	0.19	0.21	0.23	0.25	0.28	0.30	0.33	0.37	0.41	0.45	0.49	0.52	0.56	0.59	0.63	0.67	0.71	0.74	0.77	0.79	0.81	0.83	0.85	0.87	0.88	0.89	0.90	0.91	0.92	0.93		
64	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10	0.11	0.12	0.13	0.14	0.15	0.16	0.18	0.20	0.22	0.24	0.27	0.30	0.32	0.36	0.40	0.44	0.48	0.52	0.56	0.59	0.63	0.67	0.70	0.74	0.77	0.79	0.81	0.83	0.85	0.87	0.88	0.89	0.90	0.91	0.92	0.93			
65	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10	0.11	0.12	0.13	0.14	0.16	0.17	0.19	0.21	0.24	0.26	0.29	0.32	0.35	0.40	0.44	0.48	0.51	0.55	0.59	0.63	0.67	0.70	0.73	0.76	0.79	0.81	0.83	0.85	0.86	0.88	0.89	0.90	0.91	0.92	0.93				
66	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10	0.11	0.12	0.13	0.15	0.16	0.18	0.21	0.23	0.26	0.28	0.31	0.35	0.39	0.43	0.47	0.51	0.55	0.58	0.62	0.66	0.70	0.73	0.76	0.79	0.81	0.83	0.85	0.86	0.88	0.89	0.90	0.91	0.92	0.93					
67	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10	0.11	0.13	0.14	0.16	0.18	0.20	0.22	0.25	0.28	0.31	0.34	0.38	0.43	0.47	0.51	0.54	0.58	0.62	0.66	0.70	0.73	0.76	0.78	0.81	0.83	0.85	0.86	0.88	0.89	0.90	0.91	0.92	0.93						
68	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.10	0.11	0.12	0.13	0.15	0.17	0.19	0.21	0.24	0.27	0.30	0.34	0.38	0.42	0.46	0.50	0.54	0.58	0.62	0.66	0.69	0.73	0.76	0.78	0.81	0.83	0.84	0.86	0.87	0.89	0.90	0.91	0.92	0.93							
69	0.01	0.02	0.03	0.04	0.05	0.07	0.08	0.09	0.10	0.11	0.12	0.14	0.16	0.18	0.21	0.23	0.26	0.29	0.33	0.37	0.42	0.46	0.50	0.53	0.57	0.61	0.65	0.69	0.72	0.75	0.78	0.80	0.82	0.84	0.86	0.87	0.89	0.90	0.91	0.92	0.93								
70	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10	0.12	0.14	0.16	0.18	0.21	0.23	0.26	0.29	0.33	0.37	0.42	0.46	0.50	0.54	0.58	0.63	0.67	0.71	0.75	0.78	0.80	0.83	0.85	0.86	0.88	0.89	0.90	0.91	0.92	0.93									
71	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.10	0.11	0.13	0.15	0.17	0.20	0.22	0.25	0.29	0.33	0.37	0.41	0.45	0.49	0.53	0.58	0.62	0.67	0.71	0.74	0.77	0.80	0.82	0.84	0.86	0.88	0.89	0.90	0.91	0.92	0.93										
72	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.09	0.10	0.12	0.14	0.16	0.19	0.22	0.25	0.28	0.32	0.36	0.40	0.44	0.48	0.53	0.57	0.62	0.66	0.70	0.74	0.77	0.80	0.82	0.84	0.86	0.88	0.89	0.90	0.91	0.92	0.93											
73	0.01	0.02	0.03	0.04	0.05	0.07	0.08	0.10	0.12	0.14	0.16	0.19	0.21	0.24	0.28	0.32	0.36	0.40	0.44	0.48	0.53	0.57	0.62	0.66	0.70	0.74	0.77	0.80	0.82	0.84	0.86	0.88	0.89	0.90	0.91	0.92	0.93												
74	0.01	0.02	0.03	0.05	0.06	0.07	0.09	0.11	0.13	0.15	0.18	0.20	0.23	0.27	0.31	0.35	0.40	0.44	0.48	0.52	0.57	0.61	0.66	0.70	0.74	0.77	0.80	0.82	0.84	0.86	0.87	0.89	0.90	0.91	0.92	0.93													
75	0.01	0.02	0.04	0.05	0.06	0.08	0.10	0.12	0.14	0.17	0.20	0.23	0.26	0.30	0.35	0.39	0.43	0.47	0.52	0.56	0.61	0.66	0.70	0.74	0.77	0.79	0.82	0.84	0.86	0.87	0.89	0.90	0.91	0.92	0.93														
76	0.01	0.02	0.04	0.05	0.07	0.09	0.11	0.14	0.16	0.19	0.22	0.26	0.30	0.34	0.39	0.43	0.47	0.51	0.56	0.61	0.65	0.70	0.73	0.76	0.79	0.82	0.84	0.86	0.87	0.89	0.90	0.91	0.92	0.93															
77	0.01	0.03	0.04	0.06	0.08	0.10	0.13	0.15	0.18	0.21	0.25	0.29	0.34	0.38	0.42	0.46	0.51	0.56	0.60	0.65	0.69	0.73	0.76	0.79	0.81	0.84	0.85	0.87	0.88	0.90	0.91	0.92	0.93																
78	0.01	0.03	0.05	0.07	0.09	0.12	0.14	0.17	0.20	0.24	0.28	0.33	0.37	0.42	0.46	0.50	0.55	0.60	0.65	0.69	0.73	0.76	0.79	0.81	0.83	0.85	0.87	0.88	0.90	0.91	0.92	0.93																	
79	0.02	0.03	0.06	0.08	0.11	0.13	0.16	0.19	0.23	0.28	0.32	0.37	0.41	0.45	0.50	0.55	0.60	0.64	0.69	0.72	0.76	0.79	0.81	0.83	0.85	0.87	0.88	0.89	0.91	0.92	0.92																		
80	0.02	0.04	0.07	0.09	0.12	0.15	0.18	0.22	0.26	0.31	0.36	0.40	0.45	0.49	0.54	0.59	0.64	0.68	0.72	0.75	0.78	0.81	0.83	0.85	0.87	0.88	0.89	0.90	0.91	0.92																			
81	0.02	0.05	0.08	0.11	0.14	0.17	0.21	0.25	0.30	0.35	0.39	0.44	0.48	0.53	0.58	0.63	0.68	0.72	0.75	0.78	0.80	0.83	0.85	0.86	0.88	0.89	0.90	0.91	0.92																				
82	0.03	0.05	0.09	0.12	0.15	0.19	0.24	0.29	0.33	0.38	0.43	0.47	0.52	0.57	0.62	0.67	0.71	0.74	0.77	0.80	0.82	0.84	0.86	0.88	0.89	0.90	0.91	0.92																					
83	0.03	0.06	0.10	0.13	0.17	0.22	0.27	0.32	0.37	0.41	0.46	0.51	0.57	0.61	0.66	0.70	0.74	0.77	0.79	0.82	0.84	0.86	0.87	0.89	0.90	0.91	0.92																						
84	0.03	0.07	0.11	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.69	0.73	0.76	0.79	0.81	0.83	0.85	0.87	0.88	0.89	0.91	0.92																							
85	0.03	0.08	0.12	0.18	0.23	0.28	0.33	0.38	0.43	0.49	0.54	0.59	0.64	0.69	0.72	0.75	0.78	0.81	0.83	0.85	0.86	0.88	0.89	0.90	0.91																								
86	0.04	0.09	0.15	0.20	0.26	0.31	0.36	0.41	0.47	0.53	0.58	0.63	0.68	0.71	0.75	0.78	0.80	0.82	0.84	0.86	0.87	0.89	0.90	0.91																									
87	0.05	0.11	0.17	0.23	0.28	0.34	0.39	0.45	0.51	0.56	0.62	0.66	0.70	0.74	0.77	0.79	0.82	0.84	0.85	0.87	0.88	0.89	0.91																										
88	0.06	0.13	0.19	0.25	0.31	0.36	0.42	0.49	0.54	0.60	0.65	0.69	0.72	0.75	0.78	0.81	0.83	0.84	0.86	0.88	0.89	0.90																											
89	0.07	0.14	0.21	0.27	0.33	0.39	0.46	0.52	0.57	0.62	0.67	0.71	0.74	0.77	0.79	0.81	0.83	0.85	0.87	0.88	0.89																												
90	0.07	0.14	0.21	0.28	0.35	0.42	0.49	0.55	0.61	0.66	0.70	0.74	0.77	0.80	0.82	0.84	0.86	0.87	0.89	0.90																													
91	0.07	0.15	0.23	0.30	0.38	0.45	0.52	0.58	0.64	0.68	0.72	0.75	0.78	0.81	0.83	0.85	0.86	0.88	0.89																													
92	0.08	0.16	0.25	0.33	0.41	0.49	0.55	0.61	0.66	0.70	0.73	0.76	0.79	0.81	0.83	0.85	0.87	0.88																													
93	0.09	0.18	0.27	0.36	0.44	0.52	0.58	0.63	0.67	0.71	0.74	0.77	0.80	0.82	0.84	0.86	0.87																													
94	0.10	0.20	0.30	0.39	0.47	0.54	0.60	0.64	0.69	0.72	0.75	0.78	0.80	0.83	0.84	0.86																													
95	0.11	0.22	0.33	0.42	0.49	0.55	0.61	0.65	0.69	0.73	0.76	0.79	0.81	0.83	0.85																													
96	0.12	0.25	0.35	0.43	0.50	0.56	0.62	0.66	0.70	0.73	0.76	0.79	0.81	0.83																													
97	0.14	0.26	0.36	0.44	0.51	0.57	0.62	0.66	0.70	0.73	0.76	0.79	0.81																													
98	0.14	0.26	0.36	0.44	0.51	0.57	0.62	0.66	0.70	0.73	0.76	0.79																													
99	0.14	0.26	0.36	0.44	0.51	0.57	0.62	0.66	0.70	0.73	0.76																														
100	0.14	0.26	0.36	0.44	0.51	0.57	0.62	0.66	0.70	0.73																															
101	0.14	0.26	0.36	0.44	0.51	0.57	0.62	0.66	0.70																																
102	0.14	0.26	0.36	0.44	0.51	0.57	0.62	0.66																																	
103	0.14	0.26	0.36	0.44	0.51	0.57	0.62																																		
104	0.14	0.26	0.36	0.44	0.51	0.57																																			
105	0.14	0.26	0.36	0.44	0.51																																				
106	0.14	0.26	0.36	0.44																																					
107	0.14	0.26	0.36																																						
108	0.14	0.26																																							
109	0.14																																							
110																																								

Table A.4: ${}_tq_x$ (Probability of death within t years, provided the gender = Male, Initial state = Pain med with any other condition). Rounded to 2DP, full precision available upon request.

x	${}_1q_x$	${}_2q_x$	${}_3q_x$	${}_4q_x$	${}_5q_x$	${}_6q_x$	${}_7q_x$	${}_8q_x$	${}_9q_x$	${}_{10}q_x$	${}_{11}q_x$	${}_{12}q_x$	${}_{13}q_x$	${}_{14}q_x$	${}_{15}q_x$	${}_{16}q_x$	${}_{17}q_x$	${}_{18}q_x$	${}_{19}q_x$	${}_{20}q_x$	${}_{21}q_x$	${}_{22}q_x$	${}_{23}q_x$	${}_{24}q_x$	${}_{25}q_x$	${}_{26}q_x$	${}_{27}q_x$	${}_{28}q_x$	${}_{29}q_x$	${}_{30}q_x$	${}_{31}q_x$	${}_{32}q_x$	${}_{33}q_x$	${}_{34}q_x$	${}_{35}q_x$	${}_{36}q_x$	${}_{37}q_x$	${}_{38}q_x$	${}_{39}q_x$	${}_{40}q_x$	${}_{41}q_x$	${}_{42}q_x$	${}_{43}q_x$	${}_{44}q_x$	${}_{45}q_x$	${}_{46}q_x$	${}_{47}q_x$	${}_{48}q_x$	${}_{49}q_x$	${}_{50}q_x$
60	0.01	0.01	0.03	0.03	0.04	0.05	0.05	0.06	0.07	0.08	0.09	0.11	0.12	0.13	0.14	0.15	0.17	0.18	0.20	0.21	0.23	0.25	0.27	0.30	0.32	0.36	0.39	0.43	0.48	0.52	0.56	0.60	0.64	0.67	0.70	0.73	0.76	0.78	0.80	0.82	0.84	0.85	0.87	0.88	0.89	0.90	0.91	0.92	0.93	0.93
61	0.01	0.01	0.03	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10	0.11	0.12	0.14	0.15	0.16	0.18	0.19	0.20	0.22	0.24	0.27	0.29	0.32	0.35	0.39	0.43	0.47	0.51	0.56	0.60	0.63	0.67	0.70	0.73	0.75	0.78	0.80	0.82	0.83	0.85	0.86	0.88	0.89	0.90	0.91	0.92	0.92	0.93	
62	0.01	0.01	0.03	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.11	0.12	0.13	0.14	0.16	0.17	0.18	0.20	0.22	0.24	0.26	0.29	0.32	0.35	0.38	0.42	0.47	0.51	0.55	0.59	0.63	0.67	0.70	0.73	0.75	0.78	0.80	0.82	0.83	0.85	0.86	0.88	0.89	0.90	0.91	0.92	0.92	0.93		
63	0.01	0.02	0.03	0.03	0.04	0.05	0.06	0.08	0.09	0.10	0.11	0.12	0.14	0.15	0.16	0.18	0.19	0.21	0.23	0.26	0.28	0.31	0.34	0.38	0.42	0.47	0.51	0.55	0.59	0.63	0.66	0.70	0.73	0.75	0.78	0.80	0.82	0.83	0.85	0.86	0.88	0.89	0.90	0.91	0.92	0.92	0.93			
64	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.11	0.12	0.13	0.14	0.16	0.17	0.19	0.20	0.23	0.25	0.28	0.31	0.34	0.38	0.42	0.46	0.50	0.55	0.59	0.63	0.66	0.69	0.72	0.75	0.77	0.80	0.81	0.83	0.85	0.86	0.87	0.89	0.90	0.91	0.92	0.92	0.93				
65	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.09	0.10	0.11	0.12	0.14	0.15	0.17	0.18	0.20	0.22	0.25	0.27	0.30	0.33	0.37	0.41	0.46	0.50	0.54	0.59	0.62	0.66	0.69	0.72	0.75	0.77	0.79	0.81	0.83	0.85	0.86	0.87	0.89	0.90	0.91	0.91	0.92	0.93					
66	0.01	0.02	0.03	0.04	0.06	0.07	0.08	0.09	0.11	0.12	0.13	0.15	0.16	0.18	0.19	0.21	0.24	0.27	0.30	0.33	0.37	0.41	0.45	0.50	0.54	0.58	0.62	0.66	0.69	0.72	0.75	0.77	0.79	0.81	0.83	0.85	0.86	0.87	0.88	0.90	0.91	0.91	0.92	0.93						
67	0.01	0.02	0.04	0.05	0.06	0.07	0.09	0.10	0.11	0.13	0.14	0.16	0.17	0.19	0.21	0.24	0.26	0.29	0.32	0.36	0.40	0.45	0.49	0.54	0.58	0.62	0.65	0.69	0.72	0.74	0.77	0.79	0.81	0.83	0.84	0.86	0.87	0.88	0.89	0.90	0.91	0.92	0.93							
68	0.01	0.02	0.04	0.05	0.07	0.08	0.09	0.11	0.12	0.13	0.15	0.17	0.18	0.20	0.23	0.26	0.29	0.32	0.36	0.40	0.45	0.49	0.53	0.58	0.62	0.65	0.69	0.72	0.74	0.77	0.79	0.81	0.83	0.84	0.86	0.87	0.88	0.89	0.90	0.91	0.92	0.93								
69	0.01	0.02	0.04	0.04	0.06	0.07	0.08	0.09	0.11	0.12	0.14	0.16	0.17	0.19	0.22	0.25	0.28	0.31	0.35	0.39	0.44	0.48	0.53	0.57	0.61	0.65	0.68	0.71	0.74	0.76	0.79	0.81	0.83	0.84	0.86	0.87	0.88	0.89	0.90	0.91	0.92	0.93								
70	0.01	0.02	0.04	0.05	0.07	0.08	0.10	0.11	0.13	0.14	0.16	0.18	0.21	0.24	0.26	0.30	0.33	0.37	0.42	0.46	0.51	0.55	0.59	0.63	0.67	0.70	0.73	0.76	0.78	0.80	0.82	0.84	0.86	0.87	0.88	0.89	0.90	0.91	0.92	0.93										
71	0.01	0.03	0.04	0.06	0.07	0.08	0.10	0.12	0.13	0.15	0.17	0.20	0.23	0.26	0.29	0.33	0.37	0.41	0.45	0.50	0.55	0.59	0.63	0.67	0.70	0.73	0.76	0.78	0.80	0.82	0.84	0.85	0.87	0.88	0.89	0.90	0.91	0.92	0.93											
72	0.01	0.03	0.04	0.06	0.07	0.09	0.11	0.12	0.14	0.16	0.19	0.22	0.25	0.28	0.32	0.36	0.41	0.45	0.49	0.54	0.59	0.63	0.66	0.70	0.73	0.75	0.78	0.80	0.82	0.84	0.85	0.87	0.88	0.89	0.90	0.91	0.92	0.93												
73	0.01	0.03	0.05	0.06	0.08	0.10	0.11	0.13	0.15	0.18	0.21	0.24	0.27	0.31	0.35	0.40	0.44	0.49	0.54	0.58	0.62	0.66	0.69	0.73	0.75	0.78	0.80	0.82	0.83	0.85	0.86	0.88	0.89	0.90	0.91	0.92	0.92													
74	0.01	0.03	0.05	0.07	0.08	0.10	0.12	0.14	0.17	0.20	0.23	0.26	0.30	0.34	0.39	0.44	0.48	0.53	0.58	0.62	0.66	0.69	0.72	0.75	0.77	0.80	0.82	0.83	0.85	0.86	0.88	0.89	0.90	0.91	0.92	0.92														
75	0.01	0.03	0.05	0.07	0.09	0.11	0.13	0.16	0.19	0.22	0.25	0.29	0.34	0.38	0.43	0.48	0.53	0.57	0.61	0.65	0.69	0.72	0.75	0.77	0.79	0.81	0.83	0.85	0.86	0.88	0.89	0.90	0.91	0.92	0.92															
76	0.02	0.03	0.06	0.08	0.10	0.12	0.15	0.18	0.21	0.24	0.28	0.33	0.38	0.42	0.47	0.52	0.57	0.61	0.65	0.68	0.72	0.74	0.77	0.79	0.81	0.83	0.85	0.86	0.88	0.89	0.90	0.91	0.92	0.92																
77	0.02	0.04	0.06	0.08	0.10	0.13	0.17	0.20	0.23	0.27	0.32	0.37	0.41	0.46	0.51	0.56	0.60	0.64	0.68	0.71	0.74	0.76	0.79	0.81	0.83	0.84	0.86	0.87	0.88	0.89	0.90	0.91	0.92	0.92																
78	0.02	0.04	0.06	0.09	0.12	0.15	0.18	0.22	0.26	0.31	0.36	0.40	0.45	0.50	0.55	0.60	0.64	0.67	0.71	0.74	0.76	0.78	0.81	0.82	0.84	0.86	0.87	0.88	0.89	0.90	0.91	0.92	0.92																	
79	0.02	0.04	0.07	0.10	0.14	0.17	0.20	0.25	0.29	0.34	0.39	0.44	0.50	0.54	0.59	0.63	0.67	0.70	0.73	0.76	0.78	0.80	0.82	0.84	0.85	0.87	0.88	0.89	0.90	0.91	0.92																		
80	0.02	0.05	0.09	0.12	0.15	0.19	0.23	0.28	0.33	0.38	0.43	0.49	0.54	0.58	0.63	0.66	0.70	0.73	0.75	0.78	0.80	0.82	0.84	0.85	0.86	0.88	0.89	0.90	0.91	0.92																			
81	0.03	0.06	0.10	0.14	0.17	0.22	0.27	0.32	0.37	0.42	0.48	0.53	0.57	0.62	0.66	0.69	0.72	0.75	0.77	0.80	0.81	0.83	0.85	0.86	0.88	0.89	0.90	0.91	0.92																				
82	0.03	0.07	0.11	0.15	0.20	0.25	0.30	0.35	0.41	0.46	0.52	0.56	0.61	0.65	0.68	0.71	0.74	0.77	0.79	0.81	0.83	0.84	0.86	0.87	0.88	0.90	0.90	0.91																					
83	0.04	0.08	0.12	0.17	0.22	0.28	0.33	0.39	0.45	0.50	0.55	0.60	0.64	0.67	0.71	0.73	0.76	0.78	0.80	0.82	0.84	0.86	0.87	0.88	0.89	0.90	0.91																						
84	0.04	0.08	0.14	0.19	0.25	0.31	0.37	0.43	0.48	0.53	0.58	0.62	0.66	0.70	0.73	0.75	0.78	0.80	0.82	0.83	0.85	0.86	0.88	0.89	0.90	0.91																							
85	0.05	0.10	0.16	0.23	0.28	0.34	0.41	0.46	0.52	0.57	0.61	0.65	0.69	0.72	0.74	0.77	0.79	0.81	0.83	0.84	0.86	0.87	0.88	0.90	0.91																								
86	0.06	0.12	0.19	0.25	0.32	0.38	0.44	0.50	0.55	0.59	0.64	0.67	0.70	0.73	0.76	0.78	0.80	0.82	0.84	0.85	0.87	0.88	0.89	0.90																									
87	0.06	0.14	0.21	0.28	0.35	0.41	0.47	0.52	0.57	0.62	0.65	0.69	0.72	0.75	0.77	0.79	0.81	0.83	0.85	0.86	0.87	0.89	0.90																									
88	0.08	0.15	0.23	0.30	0.37	0.44	0.49	0.55	0.59	0.63	0.67	0.70	0.73	0.76	0.78	0.80	0.82	0.84	0.85	0.87	0.88	0.89																									
89	0.08	0.16	0.25	0.32	0.39	0.46	0.51	0.56	0.60	0.64	0.68	0.71	0.74	0.76	0.79	0.81	0.82	0.84	0.86	0.87	0.88																									
90	0.09	0.18	0.28	0.36	0.43	0.50	0.55	0.60	0.64	0.68	0.71	0.74	0.77	0.79	0.81	0.83	0.84	0.86	0.87	0.89																									
91	0.10	0.20	0.30	0.38	0.45	0.51	0.56	0.61	0.65	0.69	0.72	0.75	0.77	0.79	0.81	0.83	0.85	0.86	0.87																									
92	0.11	0.21	0.31	0.39	0.46	0.52	0.57	0.61	0.65	0.69	0.72	0.75	0.77	0.79	0.81	0.83	0.85	0.86																									
93	0.11	0.22	0.32	0.40	0.46	0.52	0.57	0.62	0.66	0.69	0.72	0.75	0.77	0.80	0.82	0.83	0.85																									
94	0.12	0.23	0.33	0.40	0.47	0.53	0.58	0.62	0.66	0.69	0.72	0.75	0.77	0.80	0.82	0.83																									
95	0.12	0.23	0.33	0.40	0.47	0.53	0.58	0.62	0.66	0.69	0.72	0.75	0.78	0.80	0.82																									
96	0.12	0.23	0.33	0.41	0.47	0.53	0.58	0.62	0.66	0.69	0.72	0.75	0.78	0.80																									
97	0.12	0.24	0.33	0.41	0.47	0.53	0.58	0.62	0.66	0.69	0.72	0.75	0.78																									
98	0.13	0.24	0.33	0.41	0.47	0.53	0.58	0.62	0.66	0.69	0.72	0.75																										
99	0.13	0.24	0.33	0.41	0.47	0.53	0.58	0.62	0.66	0.69	0.72																											
100	0.13	0.24	0.33	0.41	0.47	0.53	0.58	0.62	0.66	0.69																												
101	0.13	0.24	0.33	0.41	0.47	0.53	0.58	0.62	0.66																													
102	0.13	0.24	0.33	0.41	0.47	0.53	0.58	0.62																														
103	0.13	0.24	0.33	0.41	0.47	0.53	0.58																															
104	0.13	0.24	0.33	0.41	0.47	0.53																																
105	0.13	0.24	0.33	0.41	0.47																																	
106	0.13	0.24	0.33	0.41																																		
107	0.13	0.24	0.33																																			
108	0.13	0.24																																				
109	0.13																																					
110																																					

Appendix B

Additional Model Outputs

Additional plots and output regarding the models introduced in Chapter 4 are provided here.

B.1 Regression

Mortality plot for all the disease combination is provided under Figure B.1.

B.2 Decision Trees vs Survival Trees

A decision tree works by repeatedly dividing the input variables, creating regions in which the target values are relatively homogeneous and thus well predicted. The search for a good target variable was difficult due to the inability to determine q_x without the relevant clustering information. Thus, initially the most obvious variable at hand was used, `Age at death`. Using variable `years till death` and using an indicator variable for whether or not an individual died within a given year as the target were also attempted. Unlike the clustering model and the Markov chain model, using a supervised model will likely be able to identify variables based on their impact on the target variable and not solely based on the data structure of the input.

B.2.0.1 Age at Death

A decision tree model was initially trained using `age at death` as the target variable. This formulation implicitly assumes that the age to which an individual may live depends on the combination of diseases they experience. However, including age itself as an input variable, introduces a form of leakage: the model would rely almost entirely on age to predict `age at death`. Conversely, excluding age results in the possibility of a predicted death age before the current age. The accuracy when run without age is roughly 2% due to the lack of information provided. Thus, the performance was not comparable to other models and have been omitted from this report.

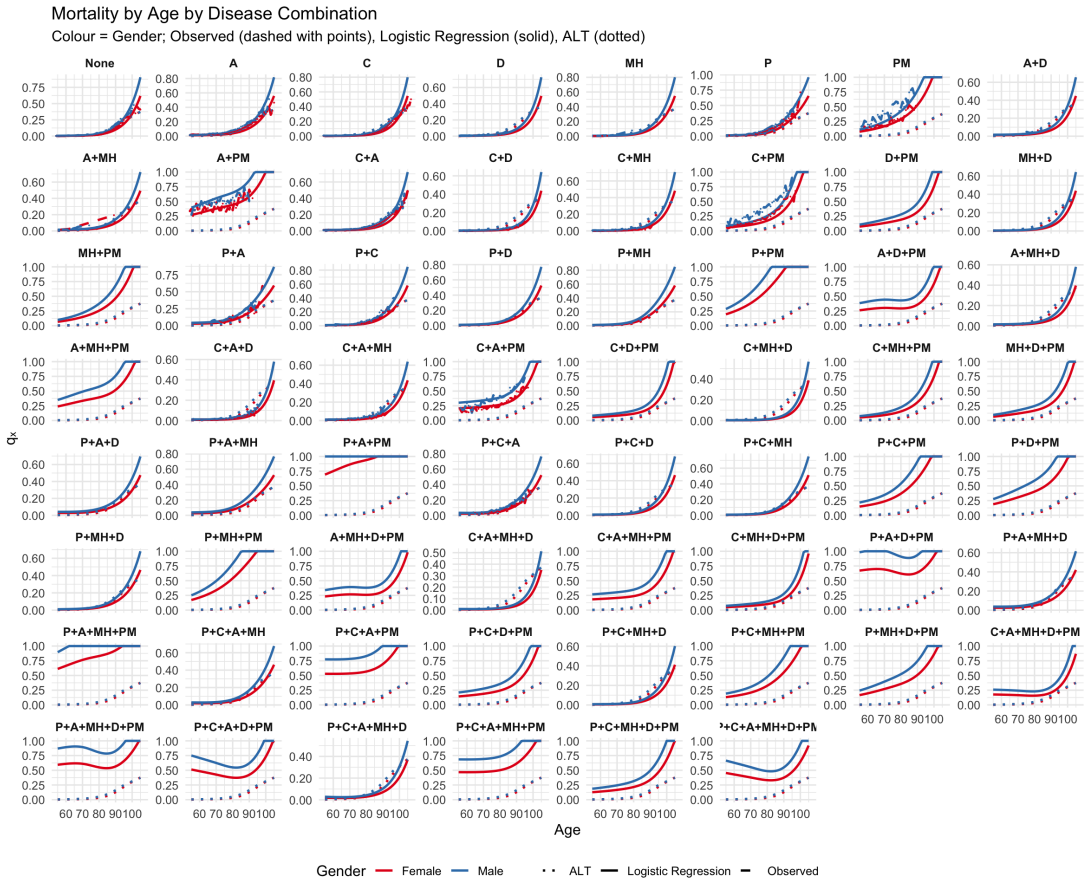


Figure B.1: Mortality for all disease combinations using regression.

B.2.0.2 Years Till Death

Because this model is used only to estimate q_x, we assess performance on the 2016 test set by comparing predicted and observed q_x. Although the framework could, in principle, produce life expectancy directly, the current dataset is not long tailed enough to estimate and validate that outcome. In estimating q_x, the model's performance matches that of a decision tree trained on the death indicator for the relevant year, as described in Section B.2.0.3. This section is retained to flag a future extension: with a longer time series, we could calibrate and evaluate life-expectancy predictions as well.

B.2.0.3 Indicator Death Variable

Methodology. An alternative approach redefined the decision tree's splitting criterion to maximise the difference in mortality rates between partitions. Specifically, the split maximised the mean of an indicator variable for death, equivalent to the

group's mortality rate:

$$\text{Mortality} = \frac{\text{number of deaths}}{\text{total people}} = \frac{\sum_{i=1}^{n} \mathbb{1}_{\{i \in \text{dead}\}}}{n} = \mathbf{E}\left[\mathbb{1}_{\{i \in \text{dead}\}}\right].$$

Hence, partitions $(P1, P2)$ were chosen to maximise $|Mortality(P1) - Mortality(P2)|$.

This differs from age-specific mortality since splits are driven by explanatory variables, but it ensures partitions with the largest overall mortality contrast.

For both the disease and the raw datasets, the model predicted the probability of death within a year for each terminal node. These probabilities can serve as mortality estimates or as clustering outputs. However, neither age nor gender emerged as key predictors, and achieving reasonable accuracy required trees so deep that interpretability was lost.

B.3 Markov Chain Model

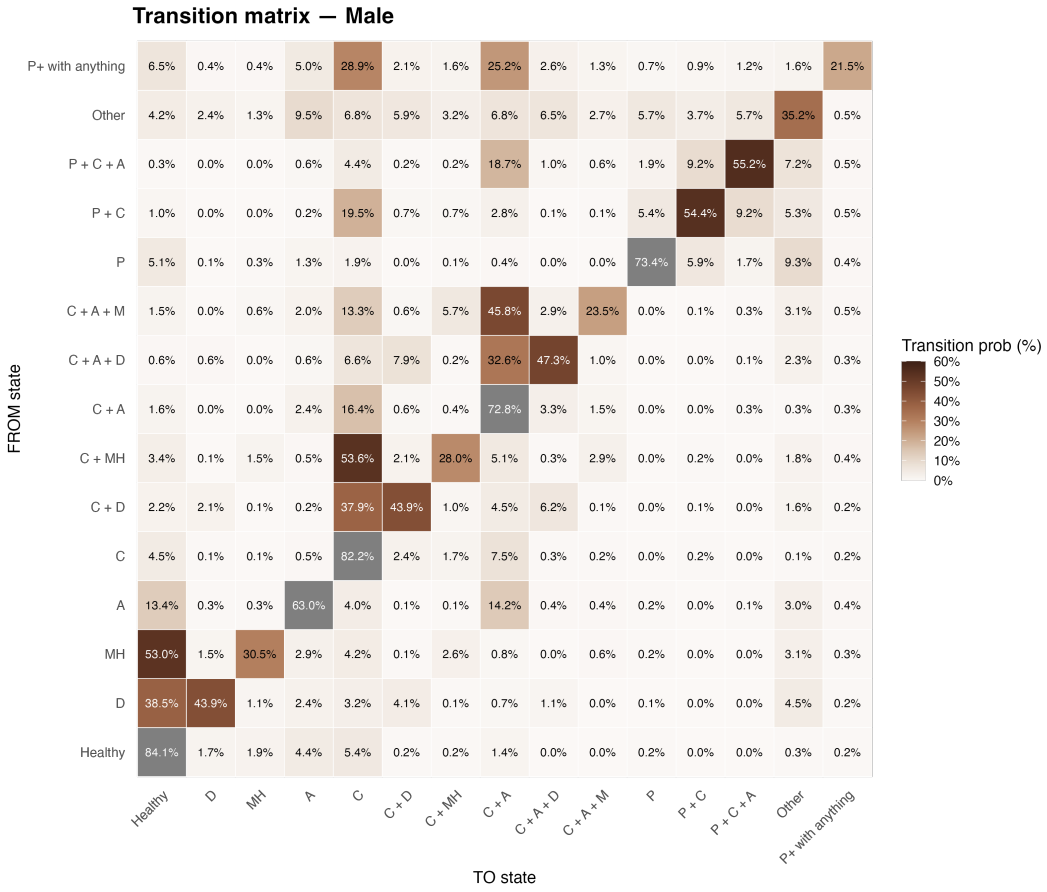


Figure B.2: Transition probabilities for males, from each of the health conditions listed in the Markov chain model to another within a year.

Figure B.3 and Figure B.2 represent the transition probabilities split by gender. As seen by these two plots, the transition probabilities vary slightly by gender and thus allowing us to not account for gender when considering transitions between states.

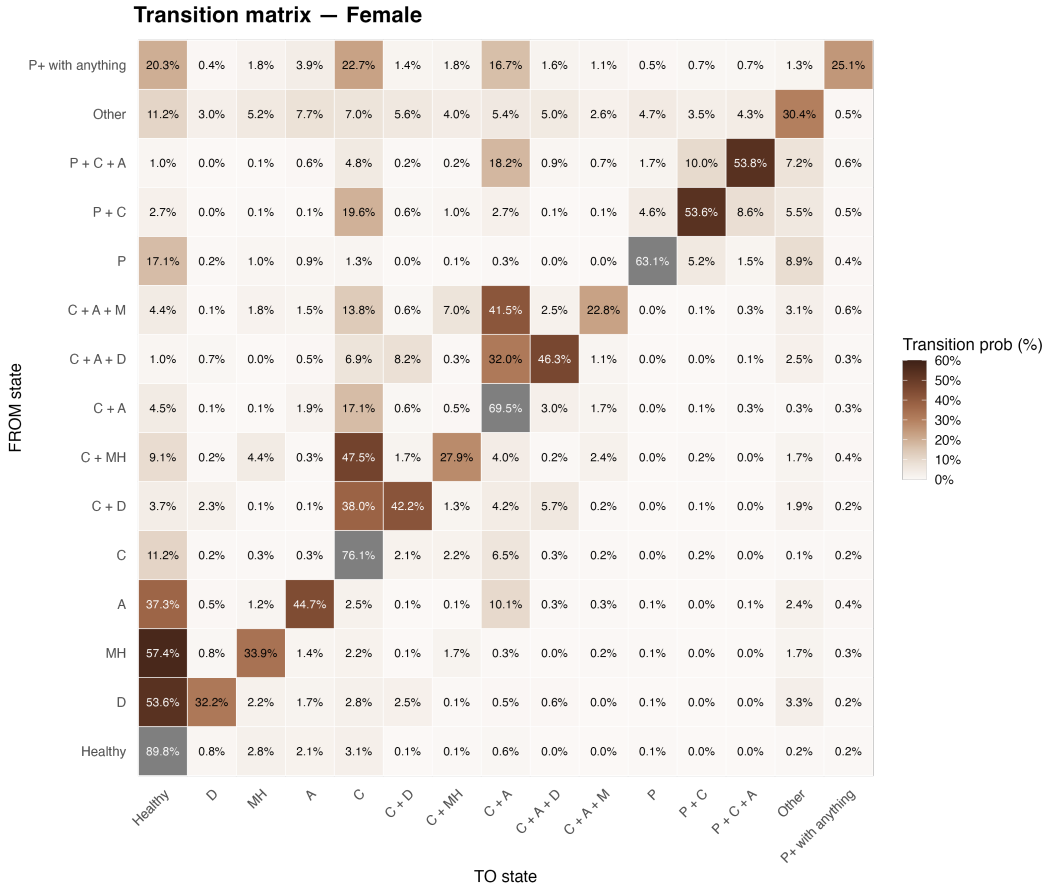


Figure B.3: Transition probabilities for females, from each of the health conditions listed in the Markov chain model to another within a year.

B.4 Transition Dynamics by Age

Figures B.4–B.6 show the estimated transition probabilities between health states across different age bands, presented as heatmaps. Each row represents an individual's current health state, and each column represents the state they are expected to be in one year later, including death. The colour intensity reflects the probability of each transition, with darker shades indicating higher likelihood. Together, these plots highlight how disease progression and mortality risk evolve with age: at younger ages, individuals are more likely to remain in the same state, while at older ages transitions into higher-risk states and death become more frequent.

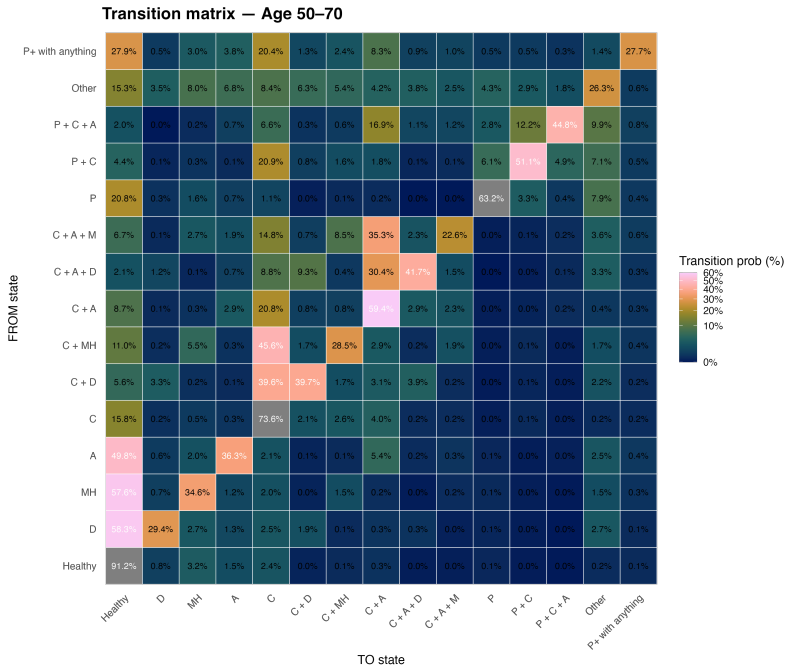


Figure B.4: Transition heatmap for ages 50–70.

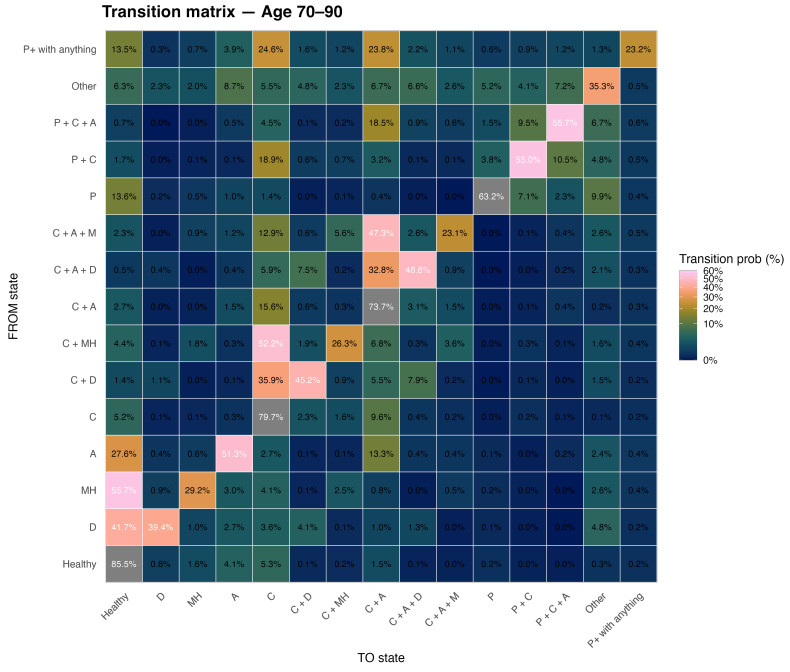


Figure B.5: Transition heatmap for ages 70–90.

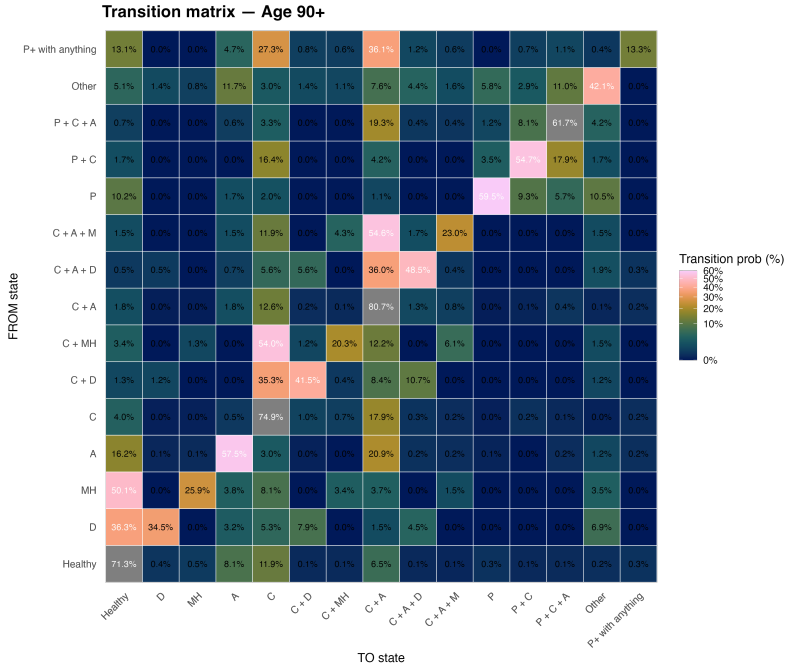


Figure B.6: Transition heatmap for ages 90 and above.

B.4.1 Life Expectancy

The Tables B.1 and B.2 report remaining life expectancy (in years) at age 65 by starting health state cluster. "*No transitions*" assumes the health state is fixed after age 65. "*With transitions (overall)*" allows annual transitions across states estimated on the full sample, while "*With transitions (age-binned)*" uses transition rates estimated within age bands to better reflect age-specific dynamics. "*Test data (smoothed)*" shows out-of-sample life expectancy using a smoothed mortality surface. State labels denote the starting cluster; combinations (e.g. `C+A`) indicate co-occurring conditions.

B.5 Additional Models on Raw MBS and PBS datasets

B.5.1 K-means Clustering

The models performed on the raw MBS and PBS variables were completed but unfortunately not vetted within the time of the thesis. Thus, were not able to be incorporated within this thesis.

B.5.2 Survival Tree

The survival tree induced from the fitted model is too large to be presented graphically within the main body of the thesis, thus has been provided in in the form of a serialised tree (see Listing B.1). The breadth and depth of the tree primarily reflect (i)

Table B.1: Life Expectancy at Age 65 — Female.

state	No transitions	With transitions (overall)	With transitions (age-binned)	Test data (smoothed)
A	16.95	25.81	25.97	15.81
C	25.12	28.98	29.17	25.57
C+A	21.04	26.06	26.50	20.62
C+A+D	22.29	27.14	27.67	22.05
C+A+MH	22.92	22.92	22.92	22.12
C+D	29.80	29.80	29.80	26.99
C+MH	25.24	25.24	25.24	26.15
D	31.03	31.03	31.03	22.05
Healthy	23.44	26.59	26.60	23.21
MH	31.03	27.70	27.84	31.66
P	15.35	24.93	25.18	15.75
P+C	19.83	25.97	26.06	20.93
P+C+A	16.43	27.97	27.89	18.24
Pain med with any condition	4.78	24.74	25.04	5.85
other	23.14	29.38	29.38	21.74

Table B.2: Life Expectancy at Age 65 — Male.

state	No transitions	With transitions (overall)	With transitions (age-binned)	Test data (smoothed)
A	15.06	21.31	21.67	14.70
C	22.15	23.69	23.84	23.02
C+A	19.10	21.72	22.03	19.19
C+A+D	21.79	22.31	22.71	19.22
C+A+MH	19.47	19.47	19.47	17.26
C+D	24.07	24.07	24.07	23.84
C+MH	20.87	20.87	20.87	20.36
D	32.23	32.23	32.23	19.22
Healthy	20.35	22.19	22.48	20.48
MH	32.23	22.68	22.80	19.22
P	13.71	21.18	21.79	13.42
P+C	17.06	21.54	21.73	16.98
P+C+A	14.19	22.70	22.73	14.64
Pain med with any condition	3.13	21.28	22.02	3.85
other	20.24	23.74	23.74	19.61

high feature cardinality across procedure and prescription groupings and (ii) complex, non-linear interactions that are better expressed via many small partitions than by a few global effects.

There are signals consistent with potential information leakage (see Chapter 3), but because mortality is an aggregated endpoint and some inputs may indirectly encode timing or treatment pathways, there is no clear way to *prove or disprove* leakage in this setting. In light of this, the survival tree should be read mainly as an *exploratory risk–stratification* tool that surfaces clinically interpretable differences across subgroups, rather than as a materially more accurate population-level predictor: it separates risk more finely, but its overall accuracy is not markedly better. If used operationally, safeguards are needed to avoid using information that would not be available at decision time, including building features strictly from pre-available data.

Listing B.1: Fitted tree for the entire raw dataset

```
[1] root
| [2] C09_scripts <= 10
| | [3] total_services <= 88
| | | [4] C10_scripts <= 8
| | | | [5] C03_qty <= 3
| | | | | [6] C08_scripts <= 10
| | | | | | [7] C09_qty <= 240: 96.000 (n = 25225)
| | | | | | [8] C09_qty > 240
| | | | | | | [9] C01D_qty <= 0: 107.000 (n = 3855)
| | | | | | | [10] C01D_qty > 0: Inf (n = 93)
| | | | | [11] C08_scripts > 10
| | | | | | [12] C09_scripts <= 8: Inf (n = 1209)
| | | | | | [13] C09_scripts > 8: Inf (n = 98)
| | | | [14] C03_qty > 3
| | | | | [15] C10_scripts <= 0
| | | | | | [16] C03_scripts <= 2: 95.000 (n = 1989)
| | | | | | [17] C03_scripts > 2: 99.000 (n = 2135)
| | | | | [18] C10_scripts > 0
| | | | | | [19] C08_scripts <= 7: 89.000 (n = 1147)
| | | | | | [20] C08_scripts > 7: Inf (n = 153)
| | | [21] C10_scripts > 8
| | | | [22] total_qty <= 1034
| | | | | [23] C08_scripts <= 9
| | | | | | [24] C10_scripts <= 10
| | | | | | | [25] C08_scripts <= 4: Inf (n = 4553)
| | | | | | | [26] C08_scripts > 4
| | | | | | | | [27] C01D_scripts <= 0: 92.000 (n = 170)
| | | | | | | | [28] C01D_scripts > 0: Inf (n = 12)
| | | | | | [29] C10_scripts > 10
| | | | | | | [30] total_services <= 21: Inf (n = 8002)
| | | | | | | [31] total_services > 21: Inf (n = 5511)
| | | | | [32] C08_scripts > 9
| | | | | | [33] B01A_qty <= 101: Inf (n = 794)
| | | | | | [34] B01A_qty > 101: Inf (n = 63)
| | | | [35] total_qty > 1034
| | | | | [36] C09_scripts <= 7: Inf (n = 3427)
| | | | | [37] C09_scripts > 7
```

```
| | | | | | [38] C02A_qty <= 330: Inf (n = 1504)
| | | | | | [39] C02A_qty > 330
| | | | | | | [40] C02A_qty <= 510: Inf (n = 17)
| | | | | | | [41] C02A_qty > 510: Inf (n = 13)
| | [42] total_services > 88
| | | [43] pain_med_num <= 1
| | | | [44] C10_scripts <= 10
| | | | | [45] total_services <= 259
| | | | | | [46] B01A_qty <= 300: 91.000 (n = 1175)
| | | | | | [47] B01A_qty > 300: Inf (n = 470)
| | | | | [48] total_services > 259: 87.000 (n = 106)
| | | | [49] C10_scripts > 10: Inf (n = 750)
| | | [50] pain_med_num > 1: 79.000 (n = 62)
| [51] C09_scripts > 10
| | [52] B01A_qty <= 222
| | | [53] C01D_scripts <= 1
| | | | [54] C08_scripts <= 11: Inf (n = 21520)
| | | | [55] C08_scripts > 11
| | | | | [56] C10_qty <= 0: Inf (n = 1673)
| | | | | [57] C10_qty > 0
| | | | | | [58] C03_scripts <= 2: Inf (n = 2297)
| | | | | | [59] C03_scripts > 2: Inf (n = 307)
| | | [60] C01D_scripts > 1: Inf (n = 861)
| | [61] B01A_qty > 222
| | | [62] C10_scripts <= 0
| | | | [63] C07_scripts <= 8
| | | | | [64] C03_scripts <= 3: Inf (n = 1549)
| | | | | [65] C03_scripts > 3: Inf (n = 299)
| | | | [66] C07_scripts > 8: Inf (n = 451)
| | | [67] C10_scripts > 0: Inf (n = 8510)

Number of inner nodes: 33
Number of terminal nodes: 34
```

The mortality per leaf node of the survival tree is shown in Figure B.7.

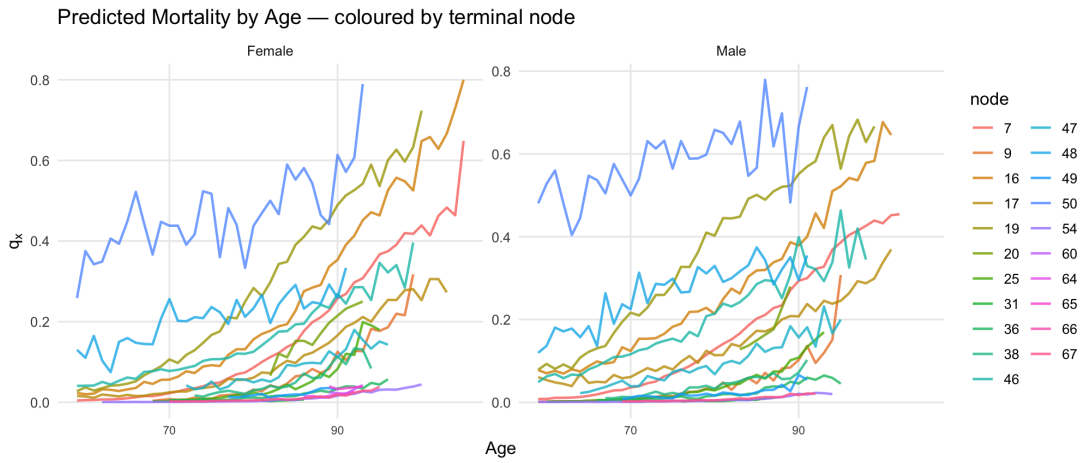


Figure B.7: Mortality for leaf nodes in the survival tree trained using raw data.

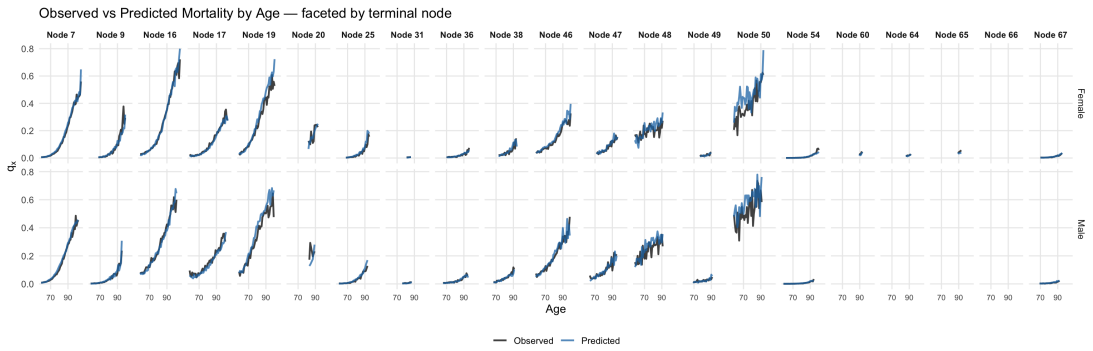


Figure B.8: The performance of the Survival Trees, for the disease only dataset.

Appendix C

Additional Related Work

Research Done on Common Causes for Death in Australia

Data from the Australian Institute of Health and Welfare [Australian Institute of Health and Welfare, 2021] shows that the top five leading causes of death in Australia have remained consistent from 2013 to 2023, with chronic conditions continuing to dominate. These causes include:

1. Coronary heart disease
2. Dementia, including Alzheimer's disease
3. Cerebrovascular disease
4. Lung cancer
5. Chronic obstructive pulmonary disease (COPD)

These findings suggest that cardiovascular diseases, respiratory illnesses, and neurological conditions remain the primary contributors to mortality in Australia. While other causes of death have shifted in rank, the consistency at the top of the list emphasises the impact of these diseases even with medical advancements [Australian Institute of Health and Welfare, 2021]. These are also listed within World Health Organisation [2021], thus these five diseases and their impact on mortality have been studied in detail.

Coronary heart disease

There are numerous research conducted on the impact of coronary heart disease on mortality [European Heart Journal, 2023; Goldberg et al., 2006a,b]. The papers mostly focus on the change in the impact coronary heart disease on mortality from 1990's to the first decade of the 2000's. The change in the impact can be due to the drastic development in the medical technology in the start of the 21st century. There are further results conducted on the impact of the disease on a sub population, for instance, McAlpine and Jackson [1990] studies the impact of coronary heart disease for middle aged diabetic patients.

The study investigates the relationship between social class and mortality in patients with coronary artery disease (CAD). It found a significant correlation between social class and death rates due to CAD, with those from lower social classes experiencing higher mortality rates. This study highlights the importance of socio-economic factors, such as income, education, and occupation, in influencing outcomes for individuals with heart disease.

Key findings of this paper include:

- Higher mortality rates in lower social classes: The study demonstrated that individuals in lower socio-economic groups were more likely to die from CAD compared to those in higher social classes.
- Impact of lifestyle factors: Lifestyle factors associated with lower social class, such as smoking, poor diet, and limited access to healthcare, contributed to the higher death rates.
- Socio-economic disparity in care access: There was also evidence suggesting that individuals in lower social classes had less access to medical care and preventive measures, which may contribute to the disparity in mortality rates.

The study concluded that addressing socio-economic disparities and improving access to healthcare could help reduce the mortality gap in patients with coronary artery disease (CAD). However, these findings have not been explored at the population level or within the Australian context. Therefore, further analysis is needed to assess whether these results are applicable to Australian data. This underscores the importance of considering socio-economic variables in mortality studies, as the impact of social class and lifestyle factors may vary depending on the specific disease. Consequently, it is crucial to account for these variables, as their effect can differ in magnitude based on the disease under consideration.

Dementia, including Alzheimer's disease

The findings from these studies highlight the importance of a comprehensive approach to mortality predictions, particularly for retirees. Below, we discuss the findings of relevant studies and their implications for improving mortality models.

The study published on PubMed Central [Taudorf et al., 2020] emphasises the relationship between mortality and chronic diseases such as dementia, noting that while it is well established that chronic diseases increase mortality risk, the precise mechanisms linking these diseases to mortality remain unclear. This gap in understanding is particularly relevant to this research, which aims to enhance mortality prediction models for retirees by considering a broader range of health factors. By accounting for the complexities of chronic diseases, especially dementia, into mortality prediction models, a more nuanced understanding of retirement mortality can be achieved. This reinforces the need to integrate such health conditions into the framework of mortality predictions, ensuring that models account for the increased mortality risk posed by specific diseases.

The study published in The Lancet [Liang et al., 2021] delves into global health challenges, with a particular focus on dementia as a growing health concern. The findings underscore dementia's substantial contribution to mortality, especially in ageing populations. As dementia-related mortality continues to rise, particularly in retirees, it is crucial to incorporate data on dementia and Alzheimer's disease into mortality prediction models. This aligns with the goal of this research to integrate comprehensive health data, offering a more accurate and personalised mortality forecast for retirees. By considering dementia as a specific risk factor, mortality models can better reflect the unique needs of older populations, providing insights into more sustainable annuity payments and retirement planning strategies.

The article from Psychosocial Gerontology [Zissimopoulos et al., 2021] highlights the significant role of psychosocial factors in shaping mortality outcomes. Factors such as mental health, social support, and lifestyle behaviours, which are often influenced by socio-economic status, are critical in predicting life expectancy, especially in retirees. The findings suggest that socio-economic variables should be integrated into mortality models, a perspective directly relevant to this research. Given the growing body of evidence that links socio-economic factors with longevity, the inclusion of these variables in mortality predictions can help to create more equitable and accurate models. Such models would not only improve predictions but also assist in tailoring retirement income products that are better aligned with the specific needs of different socio-economic groups.

The study published in Neurology [Seshadri et al., 1997] investigates the impact of dementia on mortality rates, revealing that dementia significantly contributes to early mortality, particularly in older adults. The findings underscore the relevance of dementia as a critical factor in mortality predictions, particularly for retirees who are at greater risk due to age. Incorporating dementia into mortality prediction models is crucial for improving the accuracy of life expectancy forecasts for retirees, especially in light of the increasing prevalence of dementia-related conditions. This reinforces the need for more targeted mortality models that account for age-related health conditions, ensuring better financial planning and healthcare outcomes for retirees.

These findings collectively highlight the growing importance of incorporating comprehensive health data, including chronic diseases and dementia into mortality prediction models. As the ageing population grows, particularly in Australia, these factors will become increasingly significant in forecasting mortality risk. By integrating these diverse factors into mortality models, this research aims to improve prediction accuracy, offering more tailored insights for the retirement and insurance industries. Ultimately, the goal is to create a framework that provides more equitable and personalised mortality predictions.

Cerebrovascular disease

Cerebrovascular disease encompasses any disorder of the blood vessels supplying the brain or its covering membranes. The most common manifestation is stroke, either ischaemic (vessel blockage) or hemorrhagic (vessel rupture), but it also includes transient ischaemic attacks, aneurysms and vascular malformations [Australian In-

stitute of Health and Welfare, 2015].

Cerebrovascular disease has been a major driver of premature mortality in Australia. According to the [Australian Institute of Health and Welfare, 2015], its premature death rate peaked at 94 per 100,000 in 1952, then fell by 69 % between 1970 and 1990 (from 72 to 22 per 100,000) and continued to decline into the early 21st century (e.g. from 12 per 100,000 in 2003), reaching sex,specific rates of 8.8 per 100,000 for males and 6.5 per 100,000 for females by 2012, a narrowing of the male–female gap to 2.3 per 100,000 population.

In the United States, an analysis of cerebrovascular-related deaths from 1999 to 2020 reported decline from 1038 per 100,000 in 1999 to 709 per 100,000 in 2020 (average annual percentage change –1.9 %).

Focusing on older adults (75 years and above), a U.S. study found pronounced geographic heterogeneity: state-level AAMRs ranged from 609.7 (95 % CI 606.9–612.6) per 100,000 in New York to 1076.3 (1069.2–1083.3) per 100,000 in Tennessee over the same period, highlighting regional variations in risk factor profiles and healthcare access.

Though the number of deaths overall have reduced significantly over time, it is still one of the main causes of deaths in the older population. One important takeaway from the above papers is the importance of further investigating the mortality change split by different locations.

Lung cancer

Lung cancer remains the leading cause of cancer death in Australia and globally. In 2022 there were 9 048 deaths from lung cancer in Australia, and age-standardised mortality rates declined from 54.2 to 35.0 per 100 000 persons between 1982 and 2022; in 2024 lung cancer is estimated to account for 16.9 % of all cancer deaths in Australia [Cancer Australia, 2025].

Globally, the GLOBOCAN 2022 estimates report 2.21 million new lung cancer cases (12.4 % of all cancers) and 1.80 million deaths (18.7 % of all cancer deaths), making lung cancer the most lethal malignancy worldwide [Bray et al., 2024; World Health Organization, 2023]. Tobacco use remains the predominant risk factor responsible for approximately 85 % of cases [World Health Organization, 2023].

In the United States, lung and bronchus cancer accounted for the third highest number of new cancer diagnoses in 2025 and contributed substantially to an estimated 618 120 cancer deaths; age-adjusted mortality rates for lung cancer have declined more slowly than for breast and colorectal cancers [National Cancer Institute, 2025].

Population-based registry data from Japan reveal that, for distant-stage lung tumours, over 80 % of five-year mortality in lung cancer patients is directly attributable to the disease itself, whereas localised tumours exhibit a higher proportion of competing-cause deaths; this underscores the importance of stage-specific risk adjustment when modelling lung cancer mortality [Charvat et al., 2023].

Incorporating lung cancer status and stage into cohort-based mortality models can therefore enhance risk-pooling accuracy for retirement insurance and annuity

products by capturing the substantial excess hazard associated with this disease.

Chronic obstructive pulmonary disease (COPD)

COPD is a major contributor to premature mortality both in Australia and worldwide. Globally, it was the fourth leading cause of death in 2021, causing 3.5 million deaths (5 % of all deaths), with nearly 90 % of COPD deaths under age 70 occurring in low- and middle-income countries [World Health Organization, 2024]. In Australia, an estimated 638 000 people (2.5 % of the population) were living with COPD in 2022, and COPD accounted for 4.0 % of all deaths that year [Australian Institute of Health and Welfare, 2023]. Prevalence increases steeply with age (rising to 8.3 % in men and 5.4 % in women aged 75 +) and is higher in outer regional/remote areas and amongst socioeconomically disadvantaged groups [Australian Institute of Health and Welfare, 2023].

Workplace and industry also shape COPD mortality: in 2020, 10.3 % of deaths among ever-employed U.S. adults listed COPD as an underlying or contributing cause, with the highest proportionate mortality ratios in mining (PMR = 1.33) and food preparation and serving occupations (PMR = 1.30) [Syamlal et al., 2022]. Pharmacologic interventions can reduce risk recent IMPACT and ETHOS trials showed that triple inhaled therapy (ICS/LAMA/LABA) lowered all-cause mortality (HR 0.72 and 0.51, respectively) in high-risk patients with moderate-to-severe COPD [Mintz et al., 2023]. In a cohort of 339 647 English patients, 97 882 died over 2010–2020; frequent ($\geq$2) or severe exacerbation were strongly associated with COPD-related death (HR 1.64 and 2.17), while cardiovascular disease accounted for 23.3 % of deaths [Whittaker et al., 2024].

These findings underscore the importance of incorporating COPD status, exacerbation history and comorbidity profiles into pooled mortality models to more accurately price retirement insurance and annuity products while guarding against adverse selection.

Bibliography

Actuaries Institute, 2018. Exploring Retiree Mortality. Technical report, Actuaries Institute, Australia. https://actuaries.asn.au/Library/Opinion/SuperannuationRetirementIncomes/2018/AIExploringRetireeMortalityFINAL.pdf. Prepared by Rice Warner; accessed 11 May 2025. (cited on page 18)

Aggarwal, C. C.; Hinneburg, A.; and Keim, D. A., 2001. On the surprising behavior of distance metrics in high dimensional space. In *Database Theory — ICDT 2001*, vol. 1973 of *Lecture Notes in Computer Science*, 420–434. Springer, Berlin, Heidelberg. doi:10.1007/3-540-44503-X_27. https://link.springer.com/chapter/10.1007/3-540-44503-X_27. PDF available at https://bib.dbvis.de/uploadedFiles/155.pdf. (cited on page 58)

Aliverti, E.; Mazzuco, S.; and Scarpa, B., 2022. Dynamic modelling of mortality via mixtures of skewed distribution functions. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 185, 3 (2022), 1030–1054. doi:10.1111/rssa.12855. https://academic.oup.com/jrsssa/article/185/3/1030/7068920. (cited on page 23)

Alvarez-Garcia, M. H.; Ibar-Alonso, R.; and Arenas-Parra, M., 2024. A comprehensive framework for explainable cluster analysis. *Information Sciences*, 663 (2024), 120282. doi:10.1016/J.INS.2024.120282. (cited on pages 26 and 29)

Australian Bureau of Statistics, 2021. Safe researcher training. https://www.abs.gov.au/statistics/microdata-tablebuilder/datalab/safe-researcher-training. Released 19 November 2021; Accessed: 2025-09-25. (cited on page 31)

Australian Bureau of Statistics, 2023. Associated causes of death in mortality. https://www.abs.gov.au/articles/associated-causes-death-mortality. Accessed 2025-10-17. (cited on page 15)

Australian Bureau of Statistics, 2023. Population projections, australia, 2022 (base) – 2071. https://www.abs.gov.au/statistics/people/population/population-projections-australia/latest-release. Released 23 November 2023; accessed YYYY-MM-DD. (cited on page 10)

Australian Bureau of Statistics, 2024. Causes of Death, Australia, 2023. Web report, Australian Bureau of Statistics. https://www.abs.gov.au/statistics/health/causes-death/causes-death-australia/latest-release. Accessed: 11 May 2025. (cited on pages 18 and 19)

Australian Bureau of Statistics, 2024. Retirement and Retirement Intentions, Australia. https://www.abs.gov.au/statistics/labour/employment-and-unemployment/retirement-and-retirement-intentions-australia/latest-release. Reference period: 2022–23 financial year. (cited on pages 2 and 114)

Australian Bureau of Statistics, 2025. Person Level Integrated Data Asset (PLIDA). https://www.abs.gov.au/about/data-services/data-integration/integrated-data/person-level-integrated-data-asset-plida. Release date: 29 Aug 2025. Accessed 21 Sep 2025. (cited on pages 4 and 38)

Australian Commission on Safety and Quality in Health Care, 2025. Mortality — Indicators, Measurement and Reporting. https://www.safetyandquality.gov.au/our-work/indicators-measurement-and-reporting/mortality. Accessed 2025-10-17. (cited on page 9)

Australian Government, 2022. Retirement income covenant. https://treasury.gov.au/policy-topics/superannuation/retirement-framework. Accessed 2025-10-17. (cited on page 11)

Australian Government Actuary, 2014. Australian life tables 2010–12. Technical report, Australian Government Actuary. https://aga.gov.au/sites/aga.gov.au/files/publications/life_table_2010-12/downloads/Australian_Life_Tables_2010-12_Final_V2.pdf. Accessed: 2025-09-06. (cited on page 91)

Australian Government Actuary, 2021a. Life tables by birthplace: A microdata approach to resident sub-group life tables. Technical report, Centre for Population, Australian Government. https://population.gov.au/sites/population.gov.au/files/2021-12/sgm-paper2.pdf. Accessed: 10 May 2025. (cited on page 18)

Australian Government Actuary, 2021b. Life tables by relative socio-economic advantage and disadvantage: A microdata approach to resident sub-group life tables. Technical report, Centre for Population, Australian Government. https://population.gov.au/sites/population.gov.au/files/2021-12/sgm-paper3.pdf. Accessed: 10 May 2025. (cited on page 18)

Australian Government Actuary, 2021c. Life tables by state and territory: A microdata approach to resident sub-group life tables. Technical report, Centre for Population, Australian Government. https://population.gov.au/sites/population.gov.au/files/2021-12/sgm-paper1.pdf. Accessed: 10 May 2025. (cited on page 18)

Australian Government Actuary, 2021d. Sub-group mortality: A microdata approach to resident sub-group life tables. Centre for Population, Australian Government. https://population.gov.au/publications/research/sub-group-mortality-microdata-approach-resident-sub-group-life-tables. Accessed: 10 May 2025. (cited on page 18)

Australian Government Actuary, 2024. Australian Life Tables 2020-2022. Technical report, Australian Government Actuary. https://aga.gov.au/sites/aga.gov.au/

files/sites/aga.gov.au/files/publications/2024-12/australian-life-tables-2020-22_1.pdf. Accessed: 2024-05-08. (cited on pages xix, 1, 8, 17, and 23)

Australian Government Treasury, 2022. Retirement income covenant: Supporting guidance. https://treasury.gov.au/policy-topics/superannuation/retirement-framework. Australian Government Treasury, Canberra. (cited on page 2)

Australian Government Treasury, 2025. Best practice principles for superannuation retirement income solutions: Treasury guidance. https://consult.treasury.gov.au/c2025-685228. Australian Government Treasury, Canberra. (cited on page 2)

Australian Institute of Health and Welfare, 2014. *Australia's Hospitals 2012–13: At a Glance*. AIHW, Canberra. https://www.aihw.gov.au/reports/hospitals/australias-hospitals-2012-13-at-a-glance. Includes discussion on hospital mortality indicators and outcomes. (cited on page 15)

Australian Institute of Health and Welfare, 2015. Premature mortality due to cerebrovascular disease. Fact sheet PHE 195, AIHW. https://www.aihw.gov.au/getmedia/a43dd111-1304-4abf-9946-b8513a57981b/phe195-cerebrovascular.pdf.aspx. Accessed: 12 May 2025. (cited on pages 149 and 150)

Australian Institute of Health and Welfare, 2018. *Causes of Death, Australia 2018*. AIHW, Canberra. https://www.aihw.gov.au/reports/life-expectancy-deaths/deaths-in-australia/contents/life-expectancy. Analysis of disease-specific mortality trends in Australia. (cited on page 15)

Australian Institute of Health and Welfare, 2021. Leading causes of death. https://www.aihw.gov.au/reports/life-expectancy-deaths/deaths-in-australia/contents/leading-causes-of-death. Accessed: 2024-05-08. (cited on pages 19 and 147)

Australian Institute of Health and Welfare, 2023. Chronic obstructive pulmonary disease. Report, AIHW. https://www.aihw.gov.au/reports/chronic-respiratory-conditions/copd. Accessed: 12 May 2025. (cited on page 151)

Australian Institute of Health and Welfare, 2025. Deaths in Australia: About. Web report, Australian Institute of Health and Welfare. https://www.aihw.gov.au/reports/life-expectancy-deaths/deaths-in-australia/contents/about. Accessed: 11 May 2025. (cited on page 18)

Australian Institute of Health and Welfare (AIHW), 2021. Opioid harm in australia. https://www.aihw.gov.au/reports/illicit-use-of-drugs/opioid-harm-in-australia/summary. Accessed 2025-09-11. (cited on pages 20 and 45)

Australian Prudential Regulation Authority, 2019. Apra insights: Superannuation system statistics and policy trends. *APRA Reports*, (2019). https://www.apra.gov.au/sites/default/files/quarterly_superannuation_performance_statistics_june_2019.pdf. Accessed 2025-10-17. (cited on pages 10 and 11)

Australian Retirement Trust, 2025. How much super should i have? https://www.australianretirementtrust.com.au/superannuation/how-much-super-should-i-have. Accessed: 2025-09-15. (cited on page 111)

Australian Securities & Investments Commission, 2025. Superannuation. https://www.asic.gov.au/for-consumers/superannuation/. Accessed: 2025-10-21. (cited on page 11)

Australian Taxation Office, 2025. Accessing your super to retire. https://www.ato.gov.au/individuals-and-families/jobs-and-employment-types/working-as-an-employee/leaving-the-workforce/accessing-your-super-to-retire. Accessed: 2025-09-15. (cited on page 110)

Australian Treasury, 2020a. Retirement income review: Final report. https://treasury.gov.au/publication/p2020-100554. Accessed 2025-10-17. (cited on page 11)

Australian Treasury, 2020b. Towards higher retirement incomes for australians: a history of the australian retirement income system since federation. https://treasury.gov.au/sites/default/files/2019-03/round4.pdf. Accessed 2025-10-17. (cited on page 10)

Bray, F.; Laversanne, M.; Sung, H.; Ferlay, J.; Siegel, R. L.; Soerjomataram, I.; and Jemal, A., 2024. Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 74, 3 (2024), 229–263. doi:10.3322/caac.21834. https://doi.org/10.3322/caac.21834. (cited on page 150)

C3.ai, 2025. Lime: Local interpretable model-agnostic explanations. https://c3.ai/glossary/data-science/lime-local-interpretable-model-agnostic-explanations/. (cited on page 55)

Cancer Australia, 2025. Lung cancer in australia statistics. Web report, Cancer Australia. https://www.canceraustralia.gov.au/cancer-types/lung-cancer/lung-cancer-australia-statistics. Accessed: 12 May 2025. (cited on pages 9 and 150)

Charvat, H.; Fukui, K.; Matsuda, T.; Katanoda, K.; and Ito, Y., 2023. Impact of cancer and other causes of death on mortality of cancer patients: A study based on japanese population-based registry data. *International Journal of Cancer*, 153, 6 (2023), 1162–1171. doi:10.1002/ijc.34610. https://doi.org/10.1002/ijc.34610. Epub 2023 Jun 6; accessed: 12 May 2025. (cited on page 150)

Chiu, Y. L.; Jhou, M. J.; Lee, T. S.; Lu, C. J.; and Chen, M. S., 2021. Health data-driven machine learning algorithms applied to risk indicators assessment for chronic kidney disease. *Risk Management and Healthcare Policy*, 14 (2021), 4401–4412. doi:10.2147/RMHP.S319405. https://doi.org/10.2147/RMHP.S319405. (cited on page 21)

Clinical Practice Research Datalink, 2024. Search results for "mortality". https://www.cprd.com/search/content?keys=mortality&op=Search. Accessed: 2024-11-09. (cited on page 16)

Commonwealth Superannuation Corporation, 2025. Join CSCri: Retirement income stream (CSCri). https://www.csc.gov.au/Members/Retirement/Retirement-income-CSCri/Join-CSCri. Accessed: YYYY-MM-DD. (cited on page 1)

Crimmins, E.; Hayward, M.; and Saito, Y., 2016. Differential trends in mortality and life expectancy: the significance of health disparities. *The Journals of Gerontology: Series B*, 71, 5 (2016), 826–834. doi:10.1093/geronb/gbw054. (cited on page 16)

Crimmins, E. M. and Beltrán-Sánchez, H., 2018. Mortality and morbidity trends: Is there compression of morbidity? *PLOS Biology*, 16, 10 (2018), e3000148. doi:10.1371/journal.pbio.3000148. https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3000148. (cited on pages 16 and 39)

Crimmins, E. M.; Preston, S. H.; and Cohen, B. (Eds.), 2010. *International Differences in Mortality at Older Ages: Dimensions and Sources*. National Academies Press (US), Washington, DC. https://www.ncbi.nlm.nih.gov/books/NBK62597/. Panel on Understanding Divergent Trends in Longevity in High-Income Countries, National Research Council (US). (cited on page 22)

Currie, J. et al., 2023. Probabilistic modelling of pancreatic cancer survival: can markov chains predict survival in stage iv pancreatic cancer? *Annals of Pancreatic Cancer*, (2023). https://apc.amegroups.org/article/view/7769/html. Accessed: 2025-10-17. (cited on page 25)

DataCamp, 2023. An Introduction to SHAP Values and Machine Learning Interpretability. https://www.datacamp.com/tutorial/introduction-to-shap-values-machine-learning-interpretability. (cited on page 55)

De Moivre, A., 1725. *Annuities upon Lives: Or, the Valuation of Annuities upon Any Number of Lives*. William Pearson, London. First published in 1725. (cited on pages 12 and 14)

Dickson, D. C.; Hardy, M. R.; and Waters, H. R., 2013. *Actuarial Mathematics for Life Contingent Risks*. Cambridge University Press, 2nd edn. ISBN 978-1-107-04407-4. (cited on page 92)

Dudel, C. and Myrskylä, M., 2020. Estimating the number and length of episodes in disability using a markov chain approach. *Population Health Metrics*, 18, 1 (2020), 15. doi:10.1186/s12963-020-00217-0. https://pophealthmetrics.biomedcentral.com/articles/10.1186/s12963-020-00217-0. Open access. (cited on page 24)

European Heart Journal, 2023. Trends in cardiovascular health and mortality in the eu. *European Heart Journal - Quality of Care and Clinical Outcomes*, 3, 1 (2023), 20–30. https://academic.oup.com/ehjqcco/article-abstract/3/1/20/2928168?redirectedFrom=PDF. (cited on page 147)

Fernández-Ballesteros, R.; Caprara, G. V.; Schettini, R.; Bustillos, A.; Molina, M. Á.; Orosa, T.; Losada, A.; and Spagnoli, L., 2022. Behavioral lifestyles and survival: A meta-analysis on the impact of health-related behaviors on mortality. *Frontiers in Psychology*, 12 (2022), 786491. doi:10.3389/fpsyg.2021.786491. https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2021.786491/full. (cited on page 15)

Gallager, R. G., 2011. 6.3: The Kolmogorov Differential Equations. https://www.math.cmu.edu/~gautam/teaching/2011-12/880-advanced-scalc/pdfs/kolmogorov-forward.pdf. (cited on page 24)

Goldberg, R. J.; Glatfelter, K.; Burbank-Schmidt, E.; Lessard, D.; and Gore, J. M., 2006a. Trends in community mortality due to coronary heart disease. *American Heart Journal*, 151, 2 (2006), 501–507. doi:10.1016/j.ahj.2005.04.024. https://doi.org/10.1016/j.ahj.2005.04.024. (cited on page 147)

Goldberg, R. J.; Glatfelter, K.; Burbank-Schmidt, E.; Lessard, D.; and Gore, J. M., 2006b. Trends in community mortality due to coronary heart disease. *American Heart Journal*, 151, 2 (2006), 501–507. doi:10.1016/j.ahj.2005.04.024. https://pdf.sciencedirectassets.com/272407/1-s2.0-S0002870305X09267/1-s2.0-S0002870305004436/main.pdf. (cited on page 147)

Gompertz, B., 1825. On the nature of the function expressive of the law of human mortality. *Philosophical Transactions of the Royal Society of London*, 115 (1825), 513–583. doi:10.1098/rstl.1825.0026. (cited on pages 12, 14, and 29)

Haberman, S. and Millossovich, P., 2022. Robust mortality forecasting in the presence of outliers. *British Actuarial Journal*, 27 (2022), 1–24. doi:10.1017/S1357321722000015. https://www.cambridge.org/core/journals/british-actuarial-journal/article/robust-mortality-forecasting-in-the-presence-of-outliers/94156DDA37912EC6F290EB7B02A26785. (cited on page 15)

Hansen, N. U.; Ergemen, Y. E.; and Kallestrup-Lamb, M., 2025. Individual health indices via register-based health records and machine learning. *European Actuarial Journal*, 15 (2025), 607–632. doi:10.1007/s13385-025-00417-8. https://link.springer.com/article/10.1007/s13385-025-00417-8. Open access. (cited on page 21)

Harrington, R. A. and of Encyclopaedia Britannica, T. E., 2020. Case fatality rate. https://www.britannica.com/science/case-fatality-rate. Accessed 2025-10-17. (cited on page 9)

Heligman, L. and Pollard, J. H., 1980. The age pattern of mortality. *Journal of the Institute of Actuaries*, 107, 1 (1980), 49–80. doi:10.1017/S0020268100040257. (cited on page 14)

Hothorn, T.; Hornik, K.; and Zeileis, A., 2006. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15, 3 (2006), 651–674. doi:10.1198/106186006X133933. https://www.zeileis.org/papers/Hothorn+Hornik+Zeileis-2006.pdf. (cited on page 62)

Huang, H.; Hui, V.; and Villegas, P., 2023. Australian retirement mortality and longevity: A socio-economic perspective. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5253598. (cited on pages 4, 15, 18, 21, 22, 23, 26, 55, 111, and 123)

Hunt, A. and Villegas, A., 2017. Modelling mortality improvement rates: A practical guide. In *AFIR-ERM Colloquium*. International Actuarial Association. https://actuaries.org/app/uploads/2025/07/AFIRERM_Paper_Villegas_Panama2017.pdf. (cited on page 15)

Kuan, V.; Fraser, H. C.; Hingorani, M.; Denaxas, S.; Gonzalez-Izquierdo, A.; Direk, K.; Nitsch, D.; Mathur, R.; Parisinos, C. A.; Lumbers, R. T.; Sofat, R.; Wong, I. C. K.; Casas, J. P.; Thornton, J. M.; Hemingway, H.; Partridge, L.; and Hingorani, A. D., 2021. Data-driven identification of aging-related diseases from electronic health records. *Scientific Reports*, 11, 1 (2021), 82459. doi:10.1038/s41598-021-82459-y. https://www.nature.com/articles/s41598-021-82459-y. (cited on page 21)

Lee, R. and Carter, L., 1992a. Modeling and forecasting u.s. mortality. *Journal of the American Statistical Association*, 87, 419 (1992), 659–671. doi:10.2307/2290201. (cited on pages 12, 14, and 23)

Lee, R. and Carter, L., 1992b. Modeling and forecasting us mortality. *Journal of the American Statistical Association*, 87, 419 (1992), 659–671. doi:10.2307/2290603. (cited on page 29)

Li, J. S.-H. and Wong, T. W., 2018. Incorporating structural changes in mortality improvements for more accurate forecasting. *Annals of Actuarial Science*, 12, 1 (2018), 1–24. https://ideas.repec.org/a/taf/sactxx/v2020y2020i9p776-791.html. (cited on page 15)

Li, Y.; Zhang, T.; Wang, X.; et al., 2023. Impact of biological factors on human mortality: recent evidence and emerging perspectives. *Frontiers in Public Health*, 11 (2023), 10807323. https://pmc.ncbi.nlm.nih.gov/articles/PMC10807323/. (cited on page 12)

Liang, C.-S.; Li, D.-J.; Yang, F.-C.; Tseng, P.-T.; Carvalho, A. F.; Stubbs, B.; and et al., 2021. Mortality rates in alzheimer's disease and non-alzheimer's dementias: a systematic review and meta-analysis. *Journal of Alzheimer's Disease*, 73, 2 (2021), 1013–1026. doi:10.3233/JAD-200400. https://doi.org/10.3233/JAD-200400. (cited on page 149)

LLP, C. V. Cairns–blake–dowd (or cbd) model. https://www.clubvita.net/glossary/cairns-blake-dowd-or-cbd-model. Accessed 2025-10-17. (cited on page 12)

MacDonald, B.-J.; Jones, B.; Morrison, R. J.; Brown, R. L.; and Hardy, M., 2013. Research and reality: A literature review on drawing down retirement financial savings. *North American Actuarial Journal*, 17, 3 (2013), 181–215. doi:10.1080/10920277.2013.821938. (cited on page 115)

Makeham, W., 1867. On the law of mortality and the construction of annuity tables. *Journal of the Institute of Actuaries*, 13, 5 (1867), 325–358. https://www.jstor.org/stable/41134925. (cited on page 14)

McAlpine, R. G. and Jackson, T. J., 1990. Mortality in coronary artery disease and its relation to social class. *The BMJ*, 299, 6708 (1990), 1127–1130. doi:10.1136/bmj.299.6708.1127. https://www.bmj.com/content/299/6708/1127.short. (cited on page 147)

Migraine Australia, 2025. Beta blockers. https://www.migraine.org.au/beta_blockers. Accessed 2025-09-09. (cited on page 38)

Milan, V.; Fetzer, S.; and Hagist, C., 2021. Healing, surviving, or dying? – projecting the german future disease burden using a markov illness–death model with recovery. *BMC Public Health*, 21 (2021), 123. doi:10.1186/s12889-020-09941-6. https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-020-09941-6. Article number: 123. (cited on page 24)

Milinovic, M.; Santini, G.; Milinovic, D.; and Gelo, T., 2022. A multi-state markov model for mortality estimation and forecasting. *Mathematics*, 10, 7 (2022), 1162. doi:10.3390/math10071162. https://www.mdpi.com/2227-7390/10/7/1162. (cited on pages xix, 24, and 25)

Ministry of Health. *Health and Independence Report 2023*. Ministry of Health. ISBN 978-1-991075-83-3. https://www.health.govt.nz/system/files/2024-08/Health%20and%20Independence%20Report%202023_online_f3.pdf. Published July 2024; official annual overview of New Zealand's health system. (cited on page 16)

Mintz, M.; Barjaktarevic, I.; Mahler, D. A.; Make, B.; Skolnik, N.; Yawn, B.; Zeyzus-Johns, B.; and Hanania, N. A., 2023. Reducing the risk of mortality in chronic obstructive pulmonary disease with pharmacotherapy: A narrative review. *Mayo Clinic Proceedings*, 98, 2 (2023), 301–315. doi:10.1016/j.mayocp.2022.09.007. https://doi.org/10.1016/j.mayocp.2022.09.007. (cited on page 151)

Mullany, D. V. and et al., 2021. Associations between socioeconomic status, patient risk, and short-term intensive care outcomes. *Critical Care Medicine*, 49, 9 (2021), E849–E859. https://www.ccrg.org.au/all-publications/associations-between-socioeconomic-status-?utm_source=chatgpt.com. Accessed: 2025-10-18. (cited on page 23)

National Cancer Institute, 2025. Cancer statistics. Web report, National Cancer Institute, NIH. https://www.cancer.gov/about-cancer/understanding/statistics. Accessed: 12 May 2025. (cited on page 150)

National Health and Medical Research Council, 2025. National statement on ethical conduct in human research. https://www.nhmrc.gov.au/research-policy/ethics/national-statement-ethical-conduct-human-research. Accessed: 2025-09-25. (cited on page 4)

Orford, D. and Hennington, J., 2024. Period life expectancy vs cohort life expectancy: The difference is important. *Actuaries Digital*, (September 2024). https://www.actuaries.asn.au/research-analysis/period-life-expectancy-vs-cohort-life-expectancy-the-difference-is-important. Accessed 2025-10-23. (cited on page 92)

Organisation for Economic Co-operation and Development, 2022. Pensions at a glance 2022: Oecd and g20 indicators. https://www.oecd.org/pensions/. Accessed 2025-10-17. (cited on page 11)

Parliament of Australia, 1992. Superannuation guarantee (administration) act 1992. https://www.legislation.gov.au/Details/C2023C00342. Accessed 2025-10-17. (cited on page 10)

Pascariu, M. D. et al., 2018. Modelling and forecasting mortality: A comprehensive review. Technical report, SCOR. Accessed YYYY-MM-DD. (cited on pages xix, xxiii, and 14)

Perks, W., 1932. On some experiments in the graduation of mortality statistics. *Journal of the Institute of Actuaries*, 63 (1932), 12–57. (cited on page 14)

Rakshit, S. and McGough, M., 2025. How does u.s. life expectancy compare to other countries? https://www.healthsystemtracker.org/chart-collection/u-s-life-expectancy-compare-countries/. Accessed 2025-10-17. (cited on page 16)

Salisbury, C.; Johnson, L.; Purdy, S.; Valderas, J. M.; and Montgomery, A. A., 2014. Epidemiology and impact of multimorbidity in primary care: a retrospective cohort study. *British Journal of General Practice*, 61, 582 (2014), e12–e21. doi:10.3399/bjgp11X548929. https://pmc.ncbi.nlm.nih.gov/articles/PMC4103968/. (cited on page 15)

Services Australia, 2025. How much age pension you can get. https://www.servicesaustralia.gov.au/how-much-age-pension-you-can-get?context=22526. Accessed: 5 October 2025. (cited on page 115)

Seshadri, S.; Wolf, P. A.; Beiser, A.; Au, R.; McNulty, K.; White, R.; and D'Agostino, R. B., 1997. Dementia and its impact on neurological health: A study of mortality rates. *Neurology*, 49, 6 (1997), 1498. doi:10.1212/wnl.49.6.1498. https://www.neurology.org/doi/abs/10.1212/wnl.49.6.1498. (cited on page 149)

Siebert, U. et al., 2009. Bias in markov models of disease. *Mathematical Biosciences*, 221, 1 (2009), 33–41. doi:10.1016/j.mbs.2009.07.005. https://www.sciencedirect.com/science/article/pii/S0025556409001059. (cited on pages 25 and 128)

Siler, W., 1979. A competing-risk model for animal mortality. *Ecology*, 60, 4 (1979), 750–757. doi:10.2307/1936612. (cited on page 14)

SMC Australia, 2024. Australians' super savings on track to become second largest globally by the early 2030s. https://smcaustralia.com/news/australians-super-savings-on-track-to-become-second-largest-globally-by-the-early-2030s/. Accessed: 2025-10-17. (cited on page 11)

Stafford, M.; Knight, H.; Hughes, J.; Alarilla, A.; Mondor, L.; Kone, P.; Wodchis, W.; and Deeny, S., 2022. Associations between multiple long-term conditions and mortality in diverse ethnic groups. *PLoS One*, 17, 4 (2022), e0266418. doi:10.1371/journal.pone.0266418. 1932-6203 Stafford, Mai Orcid: 0000-0001-8411-1653 Knight, Hannah Hughes, Jay Alarilla, Anne Mondor, Luke Pefoyo Kone, Anna Orcid: 0000-0003-0231-2713 Wodchis, Walter P Orcid: 0000-0003-2494-7031 Deeny, Sarah R Journal Article PLoS One. 2022 Apr 1;17(4):e0266418. doi: 10.1371/journal.pone.0266418. eCollection 2022. (cited on page 16)

Statistics New Zealand, 2022. Integrated Data Infrastructure (IDI). https://www.stats.govt.nz/integrated-data/integrated-data-infrastructure/. Accessed 2025-10-17. (cited on page 16)

Steinsaltz, D. and Evans, S. N., 1969. Markov mortality models: Implications of quasistationarity and varying initial distributions. Technical Report Tech Report 636, Department of Statistics, University of California, Berkeley. https://statistics.berkeley.edu/sites/default/files/tech-reports/636.pdf. Accessed: 2025-10-18. (cited on page 23)

Steyerberg, E. and Bastiansen, A., 2019. Risk prediction models: Overview and applications in clinical medicine. *European Heart Journal*, 40, 7 (2019), 610–622. doi:10.1093/eurheartj/ehy713. (cited on page 26)

Sun, J.; Hu, J.; Luo, D.; Markatou, M.; Wang, F.; Edabollahi, S.; Steinhubl, S.; Daar, Z.; and Stewart, W., 2012. Combining knowledge and data driven insights for identifying risk factors using electronic health records. *AMIA Annu Symp Proc*, (2012), 901–910. (cited on page 20)

Swiss Re, 2023. Longevity predictions inaccurate because actuaries ignoring data. https://corporate-adviser.com/

longevity-predictions-inaccurate-because-actuaries-ignoring-data-swiss-re/. Accessed: 2024-11-10. (cited on page 2)

Syamlal, G.; Kurth, L. M.; Dodd, K. E.; Blackley, D. J.; Hall, N. B.; and Mazurek, J. M., 2022. Chronic obstructive pulmonary disease mortality by industry and occupation — united states, 2020. *Morbidity and Mortality Weekly Report*, 71, 49 (2022), 1550–1554. doi:10.15585/mmwr.mm7149a3. https://www.cdc.gov/mmwr/volumes/71/wr/mm7149a3.htm. Accessed: 12 May 2025. (cited on page 151)

Taudorf, L.; Nørgaard, A.; Brodaty, H.; Laursen, T. M.; and Waldemar, G., 2020. Mortality is known to be associated with chronic diseases such as dementia, though its precise relationship with dementia is not well clarified. *PubMed Central*, (2020). https://pmc.ncbi.nlm.nih.gov/articles/PMC8251545/. Accessed: 2024-05-08. (cited on page 148)

Thiele, T. N., 1871. On a mathematical formula to express the rate of mortality. *Transactions of the Actuarial Society of Denmark*, (1871). (cited on page 14)

UniSuper. How much super do you need to retire? https://www.unisuper.com.au/retirement/planning-your-retirement/how-much-super-do-you-need-to-retire. (cited on page 115)

van Leeuwen, P. J.; Kranse, R.; Hakulinen, T.; Roobol, M. J.; de Koning, H. J.; Bangma, C. H.; and Schröder, F. H., 2010. Disease-specific mortality may underestimate the total effect of prostate cancer screening. *Journal of Medical Screening*, 17, 4 (2010), 204–210. doi:10.1258/jms.2010.010074. https://pubmed.ncbi.nlm.nih.gov/21258131/. (cited on page 9)

Walsh, W. R. et al., 2014. Cardiac mortality and hospital outcomes in australia: A 10-year national study. *Medical Journal of Australia*, 201, 4 (2014), 203–209. doi:10.5694/mja13.10447. Investigates hospital mortality for specific diseases. (cited on page 15)

Wang, Y. and Sherris, M., 2022. A semi-parametric approach to multi-population mortality modelling using generalised additive models. *Risks*, 10, 3 (2022), 69. https://www.mdpi.com/2571-905X/7/4/69. (cited on page 15)

Wang, Y. and Zhou, J., 2023. A bayesian jump-diffusion approach to mortality modelling: Capturing pandemic effects in the lee–carter framework. *arXiv preprint arXiv:2311.04920*, (2023). https://arxiv.org/abs/2311.04920. (cited on page 15)

Weibull, W., 1951. A statistical distribution function of wide applicability. *Journal of Applied Mechanics*, 18 (1951), 293–297. (cited on page 14)

Whittaker, H.; Rothnie, K. J.; and Quint, J. K., 2024. Cause-specific mortality in copd subpopulations: A cohort study of 339,647 people in england. *Thorax*, 79, 3 (2024), 202–208. doi:10.1136/thorax-2022-219320. https://thorax.bmj.com/content/79/3/202. Accessed: 12 May 2025. (cited on page 151)

World Health Organisation, 2021. The top 10 causes of death. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death. Accessed: 2024-05-08. (cited on page 147)

World Health Organization, 2023. Lung cancer. Fact sheet, World Health Organization. https://www.who.int/news-room/fact-sheets/detail/lung-cancer. Accessed: 12 May 2025. (cited on page 150)

World Health Organization, 2024. Chronic obstructive pulmonary disease (copd). Fact sheet, World Health Organization. https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd). Accessed: 12 May 2025. (cited on page 151)

Yu, S.-H.; Su, E. C.-Y.; and Che, Y.-T., 2022. Data-driven approach to improving the risk assessment process of medical failures. *Journal of Medical Risk Assessment*, 15, 3 (2022), 200–215. https://pmc.ncbi.nlm.nih.gov/articles/PMC6209884/. (cited on page 20)

Zissimopoulos, J. M.; Tysinger, B. C.; St.Clair, P. A.; and Crimmins, E. M., 2021. Psychosocial factors in the elderly: Understanding the impacts on longevity. *Psychosocial Gerontology*, 73 (2021), S38–S39. doi:10.1093/psychsocgerontology/article/73/suppl_1/S38/4971567. https://academic.oup.com/psychsocgerontology/article/73/suppl_1/S38/4971567. (cited on pages 15 and 149)