POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Multi-state models with intermittent observation scheme in the aging research field

Tesi di Laurea Magistrale in
Mathematical Engineering - Ingegneria Matematica

Author: **Alessandra Pescina**

Student ID: 225038
Advisor: Prof. Francesca Ieva
Co-advisors: Caterina Gregorio
Academic Year: 2023-24

Abstract

Multi-state models are powerful statistical tools for analyzing life history processes, where individuals transition between distinct states over time. They are widely used in medical and epidemiological research as they naturally capture biological mechanisms and disease progression, representing categorical state transitions as time-dependent processes. However, a major challenge arises when dealing with panel data, where observations of a continuous-time process are collected at discrete time points, leading to uncertainty in transition timing and disease trajectory reconstruction. This issue is particularly relevant in the study of chronic diseases such as dementia, where patient assessments are limited to scheduled clinical visits.

This thesis addresses the limitations of existing multi-state modeling approaches under such uncertainty and proposes a novel strategy to overcome them. We first review traditional semi-parametric and parametric multi-state models that assume exact transition times, assessing their strengths and weaknesses when applied to panel data. The study then explores time-homogeneous and time-inhomogeneous multi-state models that explicitly account for interval censoring and unobserved states, providing a theoretical and computational evaluation of their performance.

A key contribution of this work is the development of Multiple Imputation for Panel Data (MIPD), designed for an illness-death model where dementia represents the intermediate disease state. This innovative methodology leverages multiple imputation to reconstruct missing information by generating the exact onset time for patients diagnosed with dementia and imputing the disease status for those whose progression remains unobserved. By doing so, MIPD eliminates the need to explicitly model uncertainty, effectively bridging the gap between traditional methods that assume exact transition times and more complex models for interval-censored data, while preserving the tractability of the likelihood function.

To evaluate the performance and reliability of different modeling strategies, we conduct an extensive simulation study, assessing their accuracy across varying observational schemes and data-generating processes. Results demonstrate that MIPD outperforms traditional

methods in settings with irregular observation schedules and wide observation intervals. However, our findings highlight that no single strategy is universally optimal since the impact of ignoring transition-time uncertainty varies across scenarios. Therefore, we provide practical recommendations for researchers working with multi-state applied to panel data, guiding them in selecting the most appropriate modeling approach based on their study design and data structure.

Keywords: Multi-state models, panel data, interval-censoring, illness-death model, Markov and Semi-Markov models, multiple imputation, chronic disease modeling, dementia, simulation study.

Abstract in lingua italiana

I modelli multi-stato sono strumenti statistici fondamentali per l'analisi di processi in cui gli individui possono transitare tra diversi stati nel tempo. Ampiamente utilizzati in ambito medico ed epidemiologico, permettono di rappresentare la progressione delle malattie e le transizioni tra stati di salute come processi dipendenti dal tempo. Tuttavia, quando si lavora con dati panel, in cui le osservazioni di un processo continuo avvengono a intervalli discreti, si introduce un'incertezza significativa relativa al momento esatto delle transizioni e alla conoscenza della traiettoria complessiva del fenomeno. Questo problema è particolarmente rilevante nello studio delle malattie croniche, come la demenza, dove le valutazioni dei pazienti si basano esclusivamente su visite programmate.

Questa tesi affronta i limiti dei modelli multi-stato tradizionali in presenza di tali incertezze e propone una strategia innovativa per superare queste difficoltà. Inizialmente, vengono esaminati i modelli semi-parametrici e parametrici, che assumono tempi di transizione esatti, valutandone vantaggi e criticità nei dati panel. Successivamente, vengono analizzati i modelli tempo-omogenei e tempo-inomogenei che tengono esplicitamente conto della censura intervallare e degli stati non osservati, fornendo un'analisi teorica e computazionale delle loro prestazioni.

Il principale contributo di questo lavoro è lo sviluppo di Multiple Imputation for Panel Data (MIPD), una metodologia innovativa applicata a un modello illness-death, in cui la demenza rappresenta lo stato di malattia. MIPD utilizza tecniche di imputazione multipla per stimare il momento esatto di insorgenza della demenza nei pazienti diagnosticati e per imputare lo stato di malattia nei soggetti per i quali l'evoluzione della demenza non è stata direttamente osservata. In questo modo, si evita la necessità di modellare esplicitamente l'incertezza, combinando i vantaggi dei modelli che assumono transizioni esatte con quelli che gestiscono la censura intervallare e preservando al contempo la trattabilità della funzione di verosimiglianza.

L'affidabilità e le prestazioni delle diverse strategie di modellazione vengono valutate attraverso un ampio studio di simulazione, che analizza l'accuratezza delle stime in scenari con differenti schemi di osservazione e processi generativi dei dati. I risultati mostrano che

MIPD migliora le prestazioni rispetto ai metodi tradizionali in contesti con osservazioni irregolari o intervalli di osservazione ampi. Tuttavia, emerge che non esiste una strategia universalmente ottimale, poiché l'impatto dell'incertezza sui tempi di transizione dipende dal contesto di studio. Per questo motivo, la tesi fornisce raccomandazioni pratiche per i ricercatori che applicano modelli multi-stato a dati panel, offrendo linee guida per la scelta del metodo più appropriato in base alla struttura dei dati e agli obiettivi dell'analisi.

Parole chiave: Modelli multi-stato, dati panel, censura intervallare, modello illness-death, modelli di Markov e Semi-Markov, imputazione multipla, modellazione delle malattie croniche, demenzia, studio di simulazione.

Contents

Introduction

Multi-state models are widely used to analyze life history processes in which each individual is assumed to occupy one of a finite number of states at any given point in time. Common applications include post-transplantation chronic disease studies, infectious disease modeling, and cancer research and treatment response [14, 34]. In recent years, the use of multi-state models in aging research has gained increasing attention. Notable examples include studies on frailty [35], which focus on the gradual decline of physical resilience; research on sarcopenia [25], aimed at understanding the progressive loss of muscle mass and strength associated with aging; and cognitive aging studies [3, 4], where transitions between different stages of dementia are modeled. Multi-state models are regarded as the preferred analytical approach in those applications because they align with underlying biological mechanisms or disease dynamics, which conceptualize categorical state transitions as a function of time. However, obtaining complete longitudinal data on an individual's transitions through different states is feasible only in certain cases. For conditions with a sudden onset, such as post-operative recovery or intensive care monitoring, patients are continuously observed, allowing precise tracking of state changes. In contrast, for chronic diseases like Alzheimer's or diabetes, individuals are typically assessed at discrete time points, such as routine medical visits or scheduled health surveys, making transitions between states inferred from periodic observations. The data that arise from this intermittent observation process are referred to as **panel data**. This type of data is characterized by an **interval censoring mechanism**, meaning that while a transition between states is known to have occurred between two observed time points, the exact timing remains uncertain. Unlike continuous monitoring, where every state transition is precisely recorded, panel data provide only snapshots of an individual's status at specific moments. As a consequence, the full sequence of states an individual occupies within an observation period is often unknown. This introduces hidden transitions, as it cannot be assumed that every intermediate state will be observed, potentially leading to additional complexities in handling panel data.

Methods for multi-state models are commonly formulated in a Markov process framework. The Markov property states that the future of the process depends only on the current

state.

The transition probabilities in continuous time Markov processes can be written in terms of the Kolmogorov Forward Equations (KFE), a first-order system of ordinary differential equations. To facilitate estimation, a time-homogeneous Markov process is usually assumed [8, 10], meaning the instantaneous risks of moving across states is considered to be constant over time (i.e., the hazard functions are constant). Under this time-homogeneous assumption, the KFE become linear leading to closed-form solutions.

Several existing methods of allowing time non-homogeneity have mostly focused on special cases where analytic solutions can be maintained, either via piecewise constant intensities or by time transformation models [6]. Titman [27], instead, developed methods for fitting non-homogeneous parametric Markov models based on direct numerical solution of the initial value problem defined by the KFE. This approach allows a much wider range of non-homogeneous models to be fitted. However, the functional forms underlying the hazards are often unknown and parametric models can be too restrictive. Flexible multi-state models can be obtained with smooth nonparametric hazard functions specification. In this case, the hazards are specified in terms of splines function basis and penalized maximum likelihood is used to estimate the models. Joly et al. [9] developed a penalized maximum likelihood estimation method for an illness-death model, where smoothing parameters are determined using a grid search combined with cross-validation. A similar model that also accommodates backward transitions was introduced later by Machado et al. [16]. Although this method is general it can become computationally intensive when multiple penalties are involved. To address this limitation, a new automatic method for estimating multiple smoothing parameters was developed [17] .

In many applications, semi-Markov models are preferable or even necessary to effectively model transition intensities over time. They relax the Markov assumption by letting the transition intensities depend on the time elapsed in the current state. For instance, in studies of human papilloma virus [11], semi-Markov models are crucial to reflect the strong link between the duration of infection and the progression to cervical abnormalities. Similarly, in aging research [30], they address the fact that the probability of functional improvement decreases as the duration of impairment increases. Nevertheless, due to the intractability of the likelihood function, approaches for estimating intermittently observed semi-Markov multi-state models within the most general framework have not been thoroughly explored in the literature.

Some strategies have been developed for a three-state progressive semi-Markov model with interval censoring, including the previously mentioned approach based on the penalized

likelihood function [9]. Wei and Kryscio [30] employed a quasi-Monte Carlo method to address the higher-order integration needed to account for uncertainties arising from unobserved transition times in a semi-Markov model that accommodates backward transitions. However, their method assumed that the entire sequence of states for each individual was fully observed, with no unobserved states. A significant advancement in relaxing this assumption was made by Aralis and Brookmeyer [2], who developed an iterative stochastic expectation-maximization algorithm. This approach uses a simulation-based approximation of the likelihood function and implements the algorithm via rejection sampling. Aastveit et al [1]. present an alternative and more recent approach, introducing a general framework for parametric inference with interval-censored multi-state data under a semi-Markov assumption. Their method accommodates any parametric model for transition times, incorporates covariates, and can handle any acyclic graph with up to six edges between the initial and absorbing states.

A completely different approach from those previously presented is that of Yu et al. [33], who propose using a multiple imputation method to create a complete disease history for the patients with exact or right-censored event times. Standard statistical methods for multi-state models are then applied to obtain parameter estimates. Covariates can be included in the regression, and the models for transition intensities can either be Markov or semi-Markov.

Importantly, only a few of the aforementioned methods offer a reproducible and generalized implementation suitable for fitting multi-state models with panel data. Among the existing approaches, we analyzed those implemented in R [23] packages. In particular:

- The *msm* package [8] fits Markov multi-state models to panel data where the subjects' statuses are observed at a finite series of inspection times. However, it assumes time-homogeneity, where transition intensities are constant or piecewise-constant between successive observation times, to maintain a tractable likelihood function.
- The *nhm* package [27, 28] supports non-homogeneous Markov models with smoothly varying transition intensities. It handles intermittently observed data by directly solving differential equations defining transition probabilities. The package allows models with log-linear time trends, Weibull intensities, or B-spline intensity functions to be specified. Additionally, users can provide custom functions for the generator matrix and its derivatives, enabling the fitting of bespoke models.
- The *SmoothHazard* package [29] is designed for illness-death Markov models, where entry into the intermediate state can be interval-censored. It produces smooth estimates of transition intensity functions by assuming Weibull or M-spline baseline

functions.

- The *smms* package [1] is the only R package specifically designed for semi-Markov multi-state models with panel data to the best of our knowledge. It constructs and optimizes the likelihood for arbitrary multi-state models, where possible state transitions are represented by an acyclic graph with one or more initial states and one or more absorbing states. The framework supports a wide range of parametric models for transition times and allows covariates to influence these times in various and flexible ways.

As highlighted, the methods developed for fitting multi-state models with intermittent observation schemes are numerous and highly diverse, differing in their underlying theoretical assumptions, the algorithms employed, and the historical context in which they were conceived and applied. The computational challenges associated with maximizing the likelihood function accounting for unobserved transition times become increasingly complex as the number of states in the model grows and as transitions become reversible. This has led to a spontaneous evolution in the literature, starting from simpler models with a single intermediate state to increasingly sophisticated frameworks capable of accommodating multiple plausible pathways and re-visitable states.

The fragmentation of knowledge highlighted in the previous paragraphs has not allowed for the creation of a unified notation or strategy. As a result, those approaching this field face considerable difficulties in determining which strategies are most suitable for specific scenarios or in understanding the motivations behind the use of particular algorithms. The lack of a cohesive framework further complicates the ability to systematically grasp the available techniques and their practical limitations.

The **primary aim** of this thesis is then to bridge this gap by conducting a **comprehensive overview of existing strategies** for defining multi-state models in the context of panel data. This effort seeks to collect, analyze, and standardize these approaches in a unified framework, providing both a theoretical foundation and practical guidance to researchers.

Given that such a synthesis and standardization has never been attempted before, it is essential to proceed incrementally. For this reason, the thesis focuses on a progressive illness-death model, which involves a single intermediate state and an irreversible transition from this state to the absorbing state. This model serves as a practical starting point for examining methodologies applicable to multi-state models with panel data, enabling a focused and accessible exploration.

Within this framework, we apply and systematically evaluate the most promising techniques for the illness-death scenario, detailing their assumptions, underlying algorithms, and performance. By maintaining a consistent and intuitive nomenclature throughout the analysis, this thesis lays the groundwork for a standardized approach to the multi-state modeling with intermittent observation field.

Beyond this comparative analysis, a key contribution of this thesis is the proposal and **implementation of an innovative method for handling multi-state models with panel data**. This approach offers great flexibility in the choice of the hazard functions and by taking a novel perspective on the treatment of intermittently observed transitions distinguish itself from existing techniques.

The **second objective** of this thesis is then to conduct a **simulation study** aimed at addressing key questions frequently encountered by researchers employing multi-state models with panel data. These issues include:

- Comparing different strategies and evaluating their strengths and limitations based on the analysis context.
- Determining the role of the sample size in ensuring robust and reliable parameter estimates across different modeling scenarios.
- Investigating the impact of different observation schemes, including varying grid densities that lead to wider or narrower observation intervals.
- Exploring the effects of incorrectly specifying the functional form of the transition intensities.
- Assessing the implications of neglecting interval censoring by treating observation times as if they were exact transition times.

As the ultimate goal, the thesis aims to synthesize the findings from the simulation study into a cohesive **set of practical recommendations**. These recommendations will assist researchers in selecting appropriate modeling strategies for analyzing multi-state models with panel data. By consolidating the conclusions in this manner, the thesis seeks to bridge the gap between theoretical advancements in this topic and their practical application, offering a valuable resource to researchers.

Structure of the thesis

This thesis is structured into six chapters. After a brief overview of the state of the art and an introduction to the objectives of the study, **Chapter 1** outlines the reasons why we

chose chronic diseases as central area of investigation. It also provides an overview of the theoretical background needed to understand multi-state models applied to panel data, along with a population-based study that illustrates how such data are generated. The chapter concludes by discussing the mathematical challenges associated with estimating multi-state models with this type of data.

In **Chapter 2**, the reasons for exploring specific modeling strategies are clarified, and a systematic overview of these approaches is presented. **Chapter 3** introduces an innovative method designed to address the limitations of the strategies discussed in the previous chapter.

Chapter 4 presents a simulation study aimed at measuring and comparing the performance of different methods under various data-generating scenarios. **Chapter 5** summaries the findings in practical recommendations to address key questions frequently encountered by researchers employing multi-state models with panel data.

Finally, the thesis concludes with a **Discussion** of the results obtained and proposals for future developments.

All analyses were conducted using R, and the complete code files are available in the GitHub repository: https://github.com/Alepescinaa/Tesi-KI.

1 | Technical background

This chapter aims to clarify the motivations behind the inception of this thesis, specifically addressing why the issue of cognitive decline has been chosen as the central area of investigation. While the introduction has already established the fundamental importance of conducting a comprehensive review of methods for fitting multi-state models with panel data, as well as the objectives of such an effort, the rationale behind the choice of aging research as the focal point of this study remains to be explained. Additionally, this chapter provides an overview of the theoretical background necessary to understand the application of multi-state models to panel data. The concept of panel data will be introduced, starting with how such data are generated, followed by an explanation of how they are processed and the complications associated with this type of data. The discussion will cover the unique challenges posed by panel data in the context of chronic disease research, highlighting issues such as irregular observation times, incomplete data, and the presence of interval censoring.

1.1. An introduction to dementia

The global population is aging at an unprecedented pace, with nearly two billion people projected to be over the age of 65 by 2050, according to the World Health Organization [32]. In high-income countries, this demographic shift is largely attributed to declining fertility rates and significant increases in life expectancy. Nevertheless, these longer lifespans are often accompanied by the progressive decline in physical and mental health among older adults. This decline places a growing strain on healthcare systems and increases the demand for both medical and social care services. Among the health challenges faced by aging populations, neurological disorders play an increasingly prominent role, thus addressing their impact becomes a pressing public health priority.

Dementia stands out as one of the most devastating, not only because of its profound effects on the individuals who experience cognitive decline but also due to its far-reaching impact on families, society, and economies. The World Health Organization estimates that the global cost of dementia surpasses 1 trillion USD annually [31], a figure expected to rise

in the coming decades as the aging population grows. Dementia is clinically featured by a progressive deterioration in multiple cognitive domains, severe enough to interfere with daily activities and social functioning. Alzheimer's disease is the most common cause of dementia, with 60–80% of dementia cases caused by the neuropathology of Alzheimer's disease [19]. Although reversible forms of dementia exist, they represent a minority of cases and will not be the focus of this thesis.

The progression of dementia is characterized by a gradual and persistent decline in both cognitive and physical abilities. It is generally divided into three main stages, each distinguished by specific symptoms, degrees of impairment, and corresponding care needs [19]. It is important to note that the course of dementia varies significantly between individuals, shaped by factors such as the type of dementia, accompanying health conditions, and the quality of care and support received.

Early Stage (Mild Dementia)
In the early stage, the symptoms of dementia are often subtle and may be attributed to normal aging. Individuals typically experience mild memory lapses and may have difficulty finding the right words or maintaining focus. Organizational challenges and a decline in problem-solving abilities may also emerge. Despite these impairments, most individuals in this stage remain largely independent and capable of managing their daily activities, although occasional reminders or minimal assistance may be required.

Middle Stage (Moderate Dementia)
The middle stage marks a significant escalation in cognitive and functional decline. Memory impairments become more pronounced, often including the inability to recognize familiar people, places, or events. Behavioral and psychological symptoms, such as agitation, paranoia, or depression, frequently develop, compounding the challenges of care. Individuals in this stage struggle with activities of daily living.

Late Stage (Severe Dementia)
The late stage of dementia is characterized by profound cognitive and physical deterioration. Individuals typically lose the ability to communicate coherently and may become nonverbal. Physical abilities, such as walking or eating, are significantly impaired, often resulting in complete dependency on caregivers for basic needs.

After briefly outlining the nature of dementia, we can now structure the rationale for selecting it as the focus of our study on multi-state models with panel data in the aging research field:

- **A widespread and growing public, social, and economic issue.**
  Dementia is a condition that significantly impacts public health, social structures, and the economy. According to the World Health Organization, in 2018, approximately 50 million people worldwide were affected by Alzheimer's disease and other forms of dementia, with this number projected to double approximately every 20 years. As the global population ages, the prevalence of dementia continues to rise, making it a critical challenge for healthcare systems and policy planning in the coming decades.

- **Multi-state models as the predominant approach to study dementia.**
  Dementia is a progressive condition that can be effectively represented using multi-state models, where individuals transition through distinct disease stages over time. These models enable researchers to track disease progression, from early stages to more advanced forms and eventual death, providing valuable insights into the dynamics of dementia. Numerous clinical trials and cohort studies leverage multi-state models to analyze disease evolution and assess the impact of various risk factors and interventions [3, 4, 20].

- **A perfect example of panel data.**
  As a chronic condition, dementia is typically diagnosed and monitored during scheduled medical visits, which only capture the disease status at discrete time points. This limitation prevents the precise observation of when individuals transition between states, leading to *interval censoring*. Interval censoring occurs when the exact timing of a transition is unknown, but it is known to have occurred between two observation points. This issue is particularly relevant in dementia research, where disease progression is rarely observed continuously. This topic will be explored further in the next subsection.

- **Expandable to more complex scenarios.**
  The progression of dementia can be modeled with increasing complexity to better capture the disease's natural history. For instance, introducing a pre-dementia state, such as mild cognitive impairment (MCI), allows researchers to study the early stages of cognitive decline and its potential reversibility. MCI is characterized by noticeable cognitive impairment beyond normal aging but not severe enough to interfere significantly with daily life. Importantly, it is considered a precursor to dementia, with some individuals eventually progressing to Alzheimer's disease or other types of dementia. By including MCI as an additional state in multi-state models, researchers can explore the probability of progression or reversion, as well as identify factors influencing the transition from MCI to dementia. This expansion

enhances the model's applicability to preventive strategies and early interventions.

From this point forward, we will focus on a specific instance of a multi-state model: the progressive illness-death model, in which the illness would be represented by the dementia condition. In its simplest form, the model includes three states:
Healthy (or disease-free): the individual has not yet developed the condition of interest.
Illness (or diseased): the individual has been diagnosed with the condition but is still alive.
Death (absorbing state): the individual has passed away, marking the end of the process.
Transitions occur progressively between these states without the possibility of reversibility or skipping states in accordance with biological and medical knowledge.

In this thesis, we will model dementia as a binary condition—either present or absent—ignoring its potential stages. This allows us to work within a reduced framework, as discussed in the introduction, facilitating the development and application of multi-state modeling techniques with panel data.

1.2. Mathematical framework

A multi-state model is defined as a model for a stochastic process $X = X(t)$, which at any time t occupies one of a set of n discrete states, S. These models are particularly useful in research where the change of an individual's state over time is of interest. The simplest case is represented by survival models, which are two-state models where the only allowed transition is from the state "alive" to "death."

Continuous-time multi-state models allow changes of state at any time, rather than at fixed intervals. These models are particularly well-suited for chronic diseases, where the disease progression occurs continuously. For this reason, we focus our research on continuous-time multi-state models.

A multi-state model is completely specified by a set of **transition intensities** (or hazard functions), which describe the instantaneous risk of transitioning from a state to another. They are typically arranged in a square matrix $\mathbf{Q}$, known as the generator matrix, with dimensions $n \times n$. The entry in the r-th row and s-th column, h_{rs}, describes the risk of transitioning from state r to state s. The rows of this matrix sum to zero, ensuring that the diagonal entries are defined as $h_{rr} = -\sum_{s \neq r} h_{rs}$. The hazards functions may depend on the time t of the process or, more generally, on a set of individual-specific or time-varying explanatory variables $\mathbf{Z}(t)$.

A continuous-time multi-state model is a **Markov** chain on the state space S if:

$$\mathbf{P}(X(t+\Delta t)=s \mid X(t)=r, \{X(v): 0 \leq v < t\}) = \mathbf{P}(X(t+\Delta t)=s \mid X(t)=r),$$

where the future depends only on the present state, not on the history of the process. This property is called the Markov property.

If, however, the sojourn times in a certain state depend on the history of the process (i.e., the time since entering the state), the resulting model is a **Semi-Markov** chain. In this case, the time scale must be reset each time a subject enters a new state. The Semi-Markov formulation is particularly useful for capturing the duration of time spent in specific states, which can be essential for chronic disease modeling.

A **time-homogeneous** multi-state model is one in which the transition intensities are constant over time. This means that the risk of transitioning from one state to another does not vary with time.

The **transition probabilities** $p_{rs}(t,u)$ represent the probability of being in state s at time u, given that the process was in state r at time t. These probabilities are obtained by solving the **Kolmogorov Forward Equations** (KFE):

$$\frac{d\mathbf{P}(t,u)}{du} = \mathbf{P}(t,u)\mathbf{Q}(u), \tag{1.1}$$

where:

- $\mathbf{P}(t,u)$ is the $n \times n$ matrix of transition probabilities, and
- $\mathbf{Q}(u)$ is the generator matrix at time u.

The initial condition for the system is:

$$\mathbf{P}(t_0,t_0) = \mathbf{I},$$

where $\mathbf{I}$ is the $n \times n$ identity matrix, indicating that the probability of staying in the same state at the same time is 1.

For most forms of $\mathbf{Q}(t)$ the expression 1.1 is intractable.
While, for time-homogeneous models (where $\mathbf{Q}(u) = \mathbf{Q}$), the solution simplifies to:

$$\mathbf{P}(t,u) = \exp\left(\mathbf{Q} \cdot (u-t)\right) \tag{1.2}$$

The transition probabilities are essential for calculating quantities of interest, such as the expected time spent in each state and cumulative transition probabilities.

In longitudinal studies, the observation window during which data are collected may result in parts of the disease progression being unobserved. The duration and structure of this observation window are influenced by the choice of **time scale** used in the analysis. The time scale defines the units of time in which the process is observed to evolve. One possible choice for the time scale in survival analysis is the time since entry into the study. In this case, time is measured from the moment an individual joins the study, and the observation window is defined relative to this entry point. However, an alternative time scale is the individual's age. Using age as the time scale means that we must account for delayed entry with respect to the origin of the observation window. This implies that individuals who enter the study at different ages will have varying lengths of follow-up. Additionally, some disease processes may not be observed if they occur outside the study's observation window, leading to potential biases. From this point forward, we adopt **age as time scale**, explicitly accounting for delayed entry into the study. This choice ensures that individuals contribute to the risk set only from the age at which they enter the study.

Different types of censoring can occur in multi-state models, depending on the data collection process:

- **Right censoring** occurs when the endpoint of interest has not occurred yet by the end of the observation window.
- **Left censoring** occurs when the time origin is unknown, i.e., the individual has already experienced the event before the study began.
- **Left truncation** occurs when individuals who have already experienced the event at the time of study recruitment are excluded from the study.
- **Interval censoring** occurs when the event of interest happens between two observation times, and the exact time of the transition is not observed.

These types of censoring arise from the observation scheme employed in the study. Common observation schemes include:

- *Fixed visits*: Each subject is observed at fixed intervals specified in advance.
- *Random visits*: Sampling times vary randomly, independently of the current state of the disease.
- *Doctor's care*: More severely ill patients are monitored more frequently, with the next sampling time determined by the current disease state.

- *Patient self − selection*: Patients may choose to visit the doctor when they experience severe symptoms.

The observation scheme is considered **informative** (e.g. doctor's care and patient self-selection) if the distribution of sampling times affects the distribution of the stochastic process. In such cases, the sampling times must be accounted for in the analysis to avoid biased estimations. Conversely, if the observation scheme is **non-informative** (e.g. fixed and random visits), it can be ignored when using maximum likelihood estimation to estimate the parameters of the multi-state process.

For the purposes of this study, we assume that the observation scheme is non-informative, and we assume independent censoring mechanisms.

1.3. Panel data generation

Panel data consist of observations of a continuous-time stochastic process at arbitrary, discrete time points. In the context of chronic disease studies, the stochastic process is represented by the transitions of individuals between different states of interest during the study period.

To properly analyze such data, we account for two phenomena often encountered in longitudinal studies: right censoring and left truncation. In the context of multi-state models, right censoring applies to the transition to death, referring to the situation where an event of interest has not occurred by the end of the observation period. Additionally, we consider delayed entry into the study, with participants entering at different ages in state 0—indicating they are free of dementia at baseline. This selection criterion leads to left truncation, as we only include individuals who have not yet developed dementia before the study begins.

Participants are then observed at scheduled intervals, which represent medical follow-ups, during which their state is assessed through appropriate diagnostic tests. When a test indicates a positive diagnosis, the individual is declared to have dementia as of that specific time point. However, the actual onset of the condition may have occurred earlier but remained unobserved until the scheduled assessment. This delay in observing the true onset leads to the phenomenon of *interval censoring*, since the exact time of a transition between states is unknown, but it is known to have occurred within two consecutive follow-ups.

Many epidemiological and medical studies exhibit this feature, particularly in the study of chronic diseases where the onset of a condition is gradual rather than acute. While

the precise time of onset may be difficult or impossible to ascertain, it is often feasible to identify the interval in which the transition occurred.

When dealing with panel data, events themselves may also be censored. For instance, consider a patient who is observed to be disease-free at their last follow-up but dies or withdraws from the study before the next scheduled visit. In such cases, it is possible that the patient developed dementia during the interval between these events, but this will remain unobserved. This scenario introduces *event censoring*, as it is impossible to confirm the patient's state or the trajectory they followed before leaving the study.

Thus, two main sources of uncertainty arise when working with panel data in chronic disease studies:

- Time censoring: uncertainty in the exact time of transitions due to interval censoring.
- Event censoring: uncertainty in the observed trajectory or state of participants.

These sources of uncertainty significantly complicate the statistical calculations required for likelihood maximization, which is essential for estimating the parameters needed to fit the model.

Real-World Application: a population-based study

We will now refer to a population-based study in the field of aging research as an example of how panel data are generated and collected in longitudinal studies.

The Swedish National Study on Aging and Care (SNAC) is a comprehensive, longitudinal study designed to examine the aging process and care provision for the elderly population across Sweden [5]. One of the unique aspects of the SNAC study is its dual focus, this approach allows for a detailed exploration of the development of care needs in the elderly population, as well as how well these needs are met by the public healthcare system and other service providers.

To ensure that the study accurately represents the entire Swedish population, SNAC is conducted across four research centers strategically distributed across different regions of the country, focusing on areas with high population density. In this work, we specifically refer to the population part of the study conducted in the Kungsholmen area, known as SNAC-K. A key consideration in the design of the SNAC-K study was the choice of age groups and follow-up intervals. To capture the full spectrum of the aging process, including the transition from work to retirement, SNAC-K includes individuals starting at

the age of 60. Participants are thus taken from 10 different age cohorts beginning at the age of 60 up to the age of 96 years with a 6-year interval between the younger cohorts (up to 78 years) and 3-year intervals thereafter, to account for the more rapid aging process of older individuals. A random sample stratified for age groups was selected from the whole 60+ old population of the considered area, including subjects living in institutions. The basic study design involves following up data collection for each cohort whenever they reach the age of an older cohort at baseline. Every six years, a new cohort of 60-year-old is added to the study.

We will briefly summarize this data collection process for the area of Kungsholmen, Stockholm in figure 1.1.

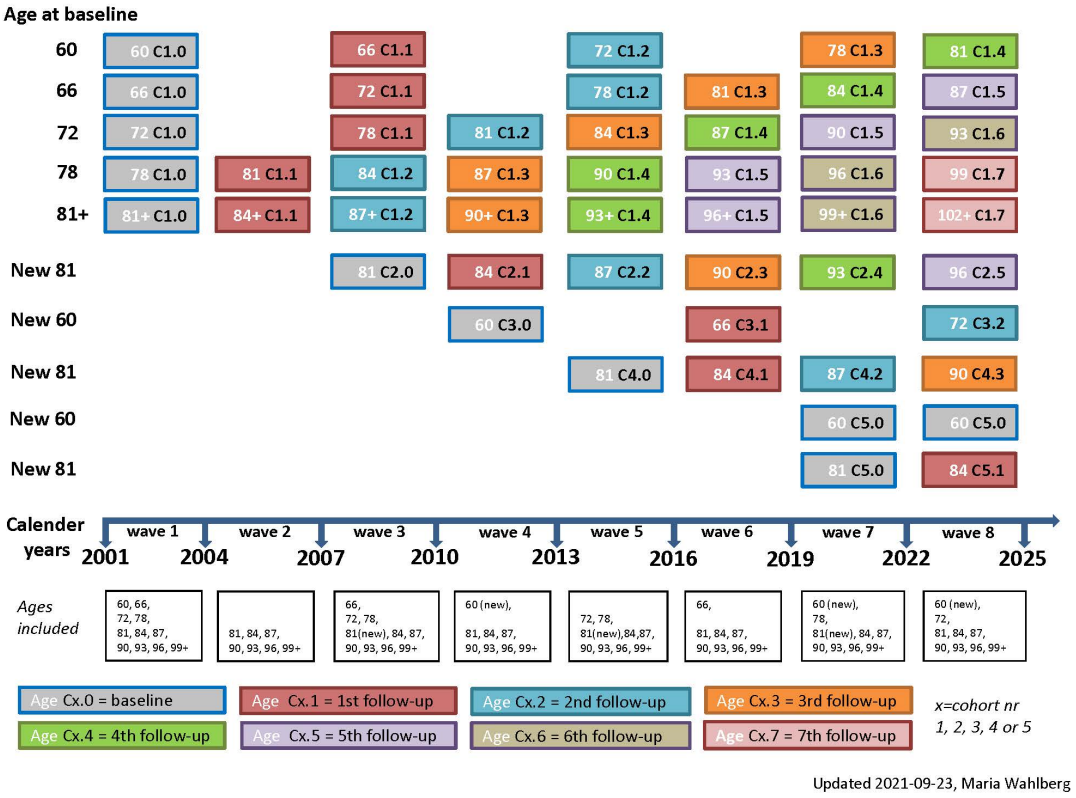


Figure 1.1: Data collection in SNAC-K study

The study's longitudinal design allows researchers to track and compare changes across different age groups over time. The core protocol includes medical, social, and psychological assessments. Interviews and clinical examinations form the backbone of the data collection process. While interviews are generally more costly, they help ensure higher-quality data compared to self-administered questionnaires, which may be difficult for older individuals to complete accurately. Examinations, lasting four to six hours, take place at medical centers or in participants' homes or institutions if needed. The study also links to

national hospital discharge and mortality registers, enhancing its ability to track chronic conditions like dementia and explore factors influencing disease progression.

As anticipated, the two primary challenges associated with working with panel data emerge from this study design.

Interval censoring arises because participants in the SNAC-K study are assessed only at specific follow-up intervals (3 or 6 years, depending on their cohort). This means that for certain events—such as the onset of a disease—the exact timing cannot be determined. Researchers can only ascertain that the event occurred sometime between two successive assessments. Therefore, the only way to address this issue is to incorporate interval censoring appropriately in the statistical analysis, particularly during the likelihood computation.

Event censoring, on the other hand, occurs when participants leave the study prematurely, either because they drop out, die, or become unavailable for follow-up before experiencing the event of interest. This results in incomplete data for those individuals beyond their last recorded observation. Their subsequent outcomes, including whether they eventually experienced the event of interest, remain unknown.

The SNAC-K study employs two strategies to **mitigate the impact of event censoring:**

- Preventing dropout: this includes making the study well-known among elderly residents in the target area through media outreach and public engagement events, such as open meetings. These meetings present the study's design, share findings from earlier research, and emphasize the importance of the participants' contributions to advancing knowledge on aging and care. Such measures help maintain a strong connection with participants, reducing the likelihood of dropout.
- Recovering data post-event: for some conditions, it is possible to gather additional information after the participant's death. For example, in cases where the onset of a disease is of interest (e.g., dementia), an autopsy conducted after death can determine whether the individual had developed the disease between their last recorded visit and the time of death. This approach allows researchers to fill in some of the missing information, enhancing the dataset and providing valuable insights for inference about disease progression and other outcomes.

1.4. Parameters estimation in multi-state models

To fit a multi-state model, the first essential step involves constructing the model's likelihood function. This function encapsulates the probability of observing the data given the parameters of the model. The parameters are then estimated by maximizing the likelihood function, a process that yields the estimators for the unknown model parameters.

In this section, we will first describe the likelihood construction and maximization in its simplest form, where exact transition times between states are known. We will then delve into the additional complexities introduced when working with panel data. Finally, we will illustrate these concepts in the specific framework of interest for this study: a progressive illness-death model.

Given the overall complexity of the likelihood formulation and the primary goal of illustrating the additional challenges introduced by the presence of panel data, we will assume a time-homogeneous model, hence the transition intensities will be constant functions of time [8].

Suppose i indexes N individuals. The data for individual i consist of a series of times $(t_{i1}, ..., t_{in_i})$ and corresponding states $(S(t_{i1}), ..., S(t_{in_i}))$, where n_i denotes the number of observation times for individual i. Consider a pair of successive observed disease states $S(t_j)$, $S(t_{j+1})$ at successive times. The contribution to the likelihood for this pair of states is determined by the nature of the time series.

When the times $(t_{i1}, ..., t_{in_i})$ represent the **exact transition times** between the states, the likelihood function for a multi-state model can be directly constructed using the transition intensities. The contribution associated to individual i can be written as:

$$\mathcal{L}_{i,j} = h_{S(t_{j-1})S(t_j)} \cdot e^{h_{S(t_j)S(t_{j+1})} \cdot (t_{j+1} - t_j)}$$

The full likelihood $\mathcal{L}$ is the product of all such terms $\mathcal{L}_{i,j}$ over all individuals and all transitions. It depends on the unknown transition matrix $\mathbf{Q}$.

When working with **interval-censored data** the times $(t_{i1}, ..., t_{in_i})$ represent the instants in which the subject was observed. The exact transition times are ignored, but the transition is known to have occurred within the interval $[t_j, t_{j-1}]$. As a result the KFE must be solved while maximizing the likelihood, making it impossible to explicitly express the likelihood as a function of the transition intensities. The contribution associated to

individual i for the pair of state is written as:

$$\mathcal{L}_{i,j} = p_{S(t_j)S(t_{j+1})}(t_{j+1} - t_j)$$

where $p_{S(t_j)S(t_{j+1})}$ represents the entry of the transition matrix $\mathbf{P}(t)$ at the $S(t_j)$ row and $S(t_{j+1})$ column, evaluated at $t = t_{j+1} - t_j$

In observational studies of chronic diseases, it is common that the time of death is known, but the state on the previous instant before death is unknown. If $S(t_{j+1}) = D$ is such a death state, then the contribution to the likelihood is summed over the unknown state m on the instant before death:

$$\mathcal{L}_{i,j} = \sum_{m \neq D} p_{S(t_j)m}(t_{j+1} - t_j) \cdot h_{S(t_j)D}$$

The sum is taken over all possible states m which can be visited between $S(t_j)$ and D.

In certain situations, both states and event times may be censored. This arises from the fact that, when individuals are observed at intermittent intervals, their exact trajectories between observations might remain unknown, as previously mentioned in the earlier section. This reflects on the likelihood formulation. For example, at the end of some chronic disease studies, patients are known to be alive but in an unknown state. For such a censored observation $S(t_{j+1})$, with $j + 1 = n_i$, known only to be a state in the set C, the equivalent contribution to the likelihood is

$$\mathcal{L}_{i,j} = \sum_{m \in C} p_{S(t_j)m}(t_{j+1} - t_j)$$

The likelihood formulation becomes even more complex and intractable when the generator matrix takes a time-dependent form $\mathbf{Q} = \mathbf{Q}(t)$. As a result, the analytical maximization of the likelihood is rarely feasible. To overcome this challenge, numerical approximation methods are commonly employed, such as quadrature methods or Monte Carlo approximations, which allow for efficient estimation of integrals. Regarding the maximization of the likelihood, iterative methods like the Expectation-Maximization (EM) algorithm or Newton-Raphson-based algorithms are used. These iterative techniques progressively refine parameter estimates, improving them until convergence is achieved. These approaches are essential for handling the complexity of time-dependent transition models and for obtaining robust estimates, especially in the presence of panel data.

Parameters estimation in a progressive illness-death model

We consider an irreversible illness-death process $X = (X(t), t \geq 0)$ which takes values in $\{0, 1, 2\}$. Subjects are initially disease-free ($X(0) = 0$) and may transition to a state representing dementia ($0 \to 1$) and die ($1 \to 2$), or directly die without disease ($0 \to 2$). Recovery from dementia is not considered, making the process irreversible.

X is assumed to be a non-homogeneous process, which means that the future evolution of the process $\{X(t + \Delta t), \Delta t > 0\}$ depends not only on the current state $X(t)$ but on the current time t as well.

The distribution of the process is fully characterized by the set of transition probabilities:

$$p_{rs}(t, t + \Delta t) = \mathbf{P}(X(t + \Delta t) = s | X(t) = r) \quad r, s \in \{0, 1, 2\}$$

The transition probabilities are related to instantaneous intensities by the relation:

$$h_k(t) = \lim_{\Delta t \to 0} \frac{p_k(t, t+\Delta t)}{\Delta t} \quad k \in \{01, 02, 12\}$$

The covariate effect is introduced separately for each transition to allow for more flexibility, even though for each subject i the covariate vector is the same across different transitions. Moreover, we assume proportional transition intensities models, i.e. constant effect of the covariates over time:

$$h_k(t) = h_{0,k}(t) \cdot \exp\left(\boldsymbol{\beta}_k^T \mathbf{Z}_i\right), \quad k \in \{01, 02, 12\}$$

where $h_{0,k}(t)$ represent the baseline intensity for each transition k, $\mathbf{Z}_i$ the covariate vectors associated to patient i and $\boldsymbol{\beta}_k$ the vector of regression parameters for the specific transition.

For each patient, we observe the vector $(L_{0i}, L_i, R_i, \delta_{1i}, \delta_{2i}, \tilde{T}_i)$ where $\tilde{T}_i = min(T_i, C_i)$ is the minimum between the transition time into the absorbing state T_i and the right censoring time C_i, and $\delta_{2i} = \mathbf{1}(T_i \leq C_i)$ indicates whether the patient is subject to right censoring. $L_{0i} \geq 0$ represents the delayed entry time at which the patient has joined the study. The left-truncation condition implies that $X(L_{0i} = 0)$. The variable δ_{1i} is set to 1 if the onset of the disease is observed, to 0 otherwise and according to it the visit times L_i and R_i are defined:

$$\begin{cases} L_{0i} \leq L_i \leq R_i \leq \tilde{T}_i & \delta_{1i} = 1 \\ L_{0i} \leq L_i, \leq \tilde{T}_i \quad R_i = \infty & \delta_{1i} = 0 \end{cases}$$

$L_i = R_i = I_i$ in the particular case in which the onset is exactly observed.

Since we are working with panel data the regression coefficients cannot be estimated by

partial likelihood maximization, but the estimation of the parameters must be done simultaneously. In order to properly express the likelihood formulation we need to introduce some quantities. For subject i, the disease-free survival function up to time t is defined by:

$S(t|\mathbf{Z}_i) = e^{-H_{01}(t|\mathbf{Z}_i)-H_{02}(t|\mathbf{Z}_i)}$

where $H_k(\cdot|\mathbf{Z}_i)$ represents the cumulative hazard function of transition k.

The likelihood contribution associated to each patient is determined in the special case $L_{0i} = 0$. Left-truncated event times are taken into account by simply dividing the likelihood contributions by the term $S(L_{0i}|\mathbf{Z}_i)$. The likelihood formulations for all possible scenario in an irreversible illness-death model are outlined below:

- **case 1**: exactly observed disease onset time, right censored alive
  $\delta_{1i} = 1, \delta_{2i} = 0, L_i = R_i = I_i$

  $\mathcal{L}_i = S(I_i|\mathbf{Z}_i) \cdot h_{01}(I_i|\mathbf{Z}_i) \cdot \frac{e^{-H_{12}(C_i|\mathbf{Z}_i)}}{e^{-H_{12}(I_i|\mathbf{Z}_i)}}$

- **case 2:** disease status at C_i unknown, right censored alive
  $\delta_{1i} = 0, \delta_{2i} = 0, L_i \leq R_i = \infty$

  $\mathcal{L}_i = S(C_i|\mathbf{Z}_i) + \int_{L_i}^{C_i} S(u|\mathbf{Z}_i) \cdot h_{01}(u|\mathbf{Z}_i) \cdot \frac{e^{-H_{12}(C_i|\mathbf{Z}_i)}}{e^{-H_{12}(u|\mathbf{Z}_i)}}\, du$

- **case 3**: interval censored disease onset time, right censored alive
  $\delta_{1i} = 1, \delta_{2i} = 0, L_i \leq R_i \leq \infty$

  $\mathcal{L}_i = \int_{L_i}^{R_i} S(u|\mathbf{Z}_i) \cdot h_{01}(u|\mathbf{Z}_i) \cdot \frac{e^{-H_{12}(C_i|\mathbf{Z}_i)}}{e^{-H_{12}(u|\mathbf{Z}_i)}}\, du$

- **case 4**: exactly observed disease onset time, death observed
  $\delta_{1i} = 1, \delta_{2i} = 1, L_i = R_i = I_i$

  $\mathcal{L}_i = S(I_i|\mathbf{Z}_i) \cdot h_{01}(I_i|\mathbf{Z}_i) \cdot \frac{e^{-H_{12}(T_i|\mathbf{Z}_i)}}{e^{-H_{12}(I_i|\mathbf{Z}_i)}} \cdot h_{12}(T_i|\mathbf{Z}_i)$

- **case 5**: disease status at death unknown, death observed
  $\delta_{1i} = 0, \delta_{2i} = 1, L_i \leq R_i = \infty$

  $\mathcal{L}_i = S(T_i|\mathbf{Z}_i) \cdot h_{02}(T_i|\mathbf{Z}_i) + \int_{L_i}^{T_i} S(u|\mathbf{Z}_i) \cdot h_{01}(u|\mathbf{Z}_i) \cdot \frac{e^{-H_{12}(T_i|\mathbf{Z}_i)}}{e^{-H_{12}(u|\mathbf{Z}_i)}} \cdot h_{12}(T_i|\mathbf{Z}_i)\, du$

- **case 6**: interval censored disease onset time, death observed
  $\delta_{1i} = 1, \delta_{2i} = 1, L_i \leq R_i \leq \infty$

$$\mathcal{L}_i = \int_{L_i}^{R_i} S(u|\mathbf{Z}_i) \cdot h_{01}(u|\mathbf{Z}_i) \cdot \frac{e^{-H_{12}(T_i|\mathbf{Z}_i)}}{e^{-H_{12}(u|\mathbf{Z}_i)}} \cdot h_{12}(T_i|\mathbf{Z}_i)\, du$$

We can observe a visual representation of all possible scenarios that generate the aforementioned likelihood formulations.

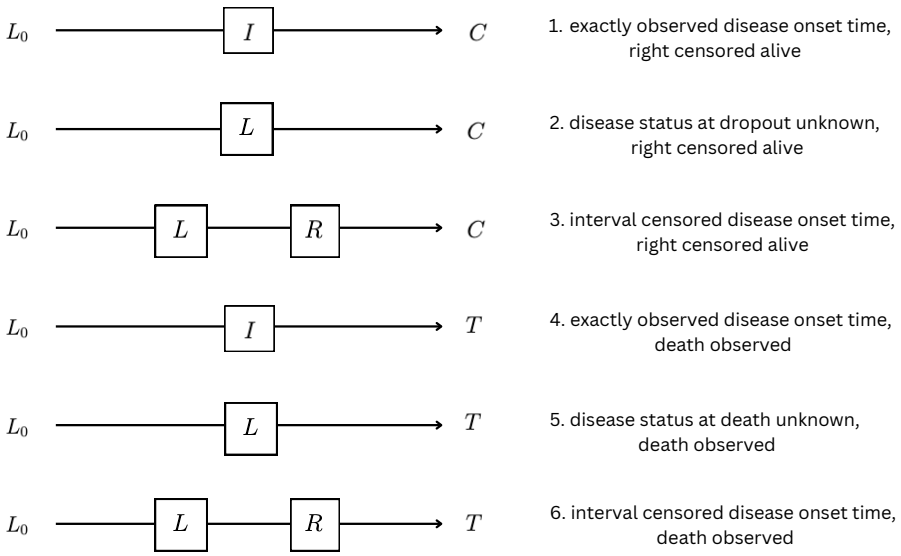


Figure 1.2: Observational patterns in an illness-death model. The letters I and T denote the transition times into the intermediate and absorbing state, respectively. The letters L_0 and C denote the start and end of follow-up, respectively, and the letters L and R the visit times between which the transition into the intermediate happened.

2 | Systematic overview of modeling strategies

2.1. Rationale for methodological choices

In this chapter, we aim to justify the choice of methodologies we have selected for investigation, before delving into their theoretical explanation and implementation.

Firstly, we prioritized methods that rely on already implemented `R` packages, motivated by two main reasons. Our ultimate goal is to support researchers using multi-state models with panel data by presenting innovative yet accessible strategies. Introducing methodologies that lack user-friendly code or require extensive expertise is not an ideal approach. Furthermore, in recent years, research in this field has progressed rapidly, yielding several new packages that remain relatively unknown. Thus, we aim to showcase those packages that we believe deserve recognition as state-of-the-art tools, making them accessible and practical for broader use.

Having clarified this aspect, we start exposing methods adopting simpler assumptions and gradually incorporating additional complexities. Specifically:

1. We begin with methods that do not account for the uncertainty associated with panel data. These methods assume that all visited states are observed and that transition times are exact. Within this category, we start with semi-parametric models and then move to the parametric framework. We investigated these strategies both under Markov and Semi-Markov assumption.
2. We then analyze methods that explicitly model the interval censoring mechanism and account for unobserved transitions in the Markovian framework:
    - Initially assuming time-homogeneous models,
    - Relaxing this assumption by investigating piecewise constant hazards,
    - Finally moving to time-inhomogeneous models with smoothly varying hazards.

For the methodologies in point 1, we deemed it reasonable to consider the two most commonly used, well-known, and user-friendly packages in this field: `survival` for the semi-parametric framework and `flexsurv` for the parametric framework. Our main objective is to make readers aware of the limitations of these packages when applied to multi-state models with panel data.

For the methodologies in point 2, we chose to use `msm`, specifically designed for implementing time-homogeneous models that account for interval censoring. Subsequently, we explored the functionalities of the `nhm` package, which extends `msm` by allowing flexible baseline hazard functions that vary smoothly with time. As an alternative to `nhm`, we considered `SmoothHazard`, which was developed for regression models applied to an illness-death model. In this framework, the transition times to the intermediate state may be interval-censored, and all event times can be right-censored. However, since this package does not support general multi-state models and offers a limited choice of baseline hazards (Weibull distributions or M-splines in a semi-parametric approach), we decided to focus on `nhm`.

Finally, we did not present any method involving penalized maximum likelihood estimation. These methods are generally computationally intensive, making them less practical, and to the best of our knowledge, they lack accessible implementations.

The literature associated to methods modeling interval censoring mechanism in in the Semi-Markovian framework is limited and the methods are highly complex. We intended to present the `smms` package, which optimizes the likelihood for arbitrary multi-state models and supports a wide range of parametric models for sojourn times. However, we found that this package is not yet well-defined for practical use due to the following limitations:

- Delayed entry for age as a time scale is not supported, requiring the assumption that all patients are at risk of the event of interest from time $t = 0$, which is unrealistic for most applications.
- Issues in solving the integrals involved in the likelihood formulation when the package is used with models other than those presented in its vignette.

2.2. Semi-parametric multi-state model assuming exact transition times

In the literature, semi-parametric multi-state models assuming exact transition times are widely used due to their extensive documentation and ease of implementation, making them a popular choice for many researchers. This approach estimates the hazard of transitioning from one state to another by fitting separate Cox proportional hazards models for each transition. One of the key reasons for its popularity is that it does not require the specification of the baseline hazard function, making it flexible and less reliant on distributional assumptions.

By analyzing its performance in multi-state models with panel data, this work aims to quantify the limitations of this method when applied to interval-censored data. This evaluation will provide insights into the suitability of this approach in such scenarios and highlight areas where alternative approaches or complementary tools may be necessary.

Theoretical details

When fitting a semi-parametric multi-state model, the hazard of transitioning from one state to another is modeled using transition-specific Cox proportional hazards models [22]. For a transition from state r to state s, the hazard function is given by:

$$h_{rs}(t \mid Z) = h_{rs}^0(t) \exp(\boldsymbol{\beta}_{rs}^\top \boldsymbol{Z}),$$

where $h_{rs}^0(t)$ is the nonparametric baseline hazard, $\boldsymbol{Z}$ represents the vector of covariates, and $\boldsymbol{\beta}_{rs}$ are the associated regression coefficients.

The main advantage of this approach is its ability to model transition rates without assuming a particular parametric form for the baseline hazard. However, there are some key drawbacks, particularly when applied to more complex multi-state processes, such as those involving many intermediate states or panel data:

- **Independent hazards estimation**: transition-specific hazards are estimated separately for each transition, without considering the sequential dependencies between transitions, which may overlook important relationships in multi-state processes.
- **Lack of interval-censoring support**: this method does not accommodate interval-censored data, which is a limitation when dealing with panel data or situations where the exact timing of transitions is unknown but falls within a known time window.

These limitations can impact the accuracy and applicability of the model in cases where transition dependencies or uncertainty in event timings are crucial.

To mitigate biases arising from interval-censored data, one common heuristic method is *midpoint imputation* [13]. This method imputes transition times at the midpoint of the interval during which the transition is known to have occurred. The interval is defined by the time of the last observation before the event occurs and the time of the diagnosis. This approach provides a simple yet practical solution in the absence of more precise timing information.

Implementation details

In `survival` package survival data is typically represented as a pair (t_i, δ_i) , where t_i denotes the time to the event (or last follow-up) and δ_i is a binary variable indicating whether the event occurred ($\delta_i = 1$) or the subject was censored ($\delta_i = 0$) [26]. In `R`, this is encoded as a $Surv(time, status)$ object.

In this study, the `survival` package is applied to multi-state data, where multiple outcomes and transitions per subject are defined. Here, the `time` variable in the survival object represents intervals $(t_1, t_2]$, and the `status` variable indicates the state transitioned into at time t_2 . An identifier variable tracks subjects, ensuring that transitions are ordered sequentially, with the endpoint t_2 of one row matching the starting point t_1 of the subsequent row for the same subject.

The `coxph()` function from the package is employed to estimate the hazards for each transition separately, assuming exact transition times as discussed in the theoretical section.

By ignoring intermediate states and the interval censoring inherent in the data, the use of the `survival` package for fitting multi-state models with panel data may introduce biases in the parameter estimates. These biases arise from oversimplified assumptions about the timing and structure of transitions that might not capture the complexity of the underlying processes.

2.3. Parametric multi-state model assuming exact transition times

Parametric multi-state models provide a flexible framework for analyzing complex event histories by explicitly defining transition-specific hazard functions using parametric distributions. Unlike semi-parametric approaches, which estimate baseline hazards nonpara-

metrically, parametric models allow for a more detailed and structured representation of transition dynamics. This is particularly beneficial for extrapolating beyond observed data and deriving key quantities such as transition probabilities and state occupation probabilities.

Theoretical details

In a parametric multi-state modeling framework, the probability density function for an event occurring at time t is generally expressed as:

$$f(t|\mu(\boldsymbol{Z}), \alpha(\boldsymbol{Z})), \quad t \geq 0$$

where μ represents the location (or central tendency) of the distribution, while $\alpha_r = (\alpha_1, \ldots, \alpha_R)$ are ancillary parameters that determine the shape or variance of the distribution.

From the density function, we can derive fundamental survival analysis functions:

- The cumulative distribution function:

$$F(t) = \int_0^t f(u)\, du$$

- The survivor function:

$$S(t) = 1 - F(t)$$

- The hazard function:

$$h(t) = \frac{f(t)}{S(t)}$$

- The cumulative hazard:

$$H(t) = \int_0^t h(u)\, du = -\log(S(t))$$

All parameters on which f depends may be modeled as functions of a vector of covariates $\boldsymbol{Z}$ through link-transformed linear models:

$$g_0(\mu(\boldsymbol{Z})) = \gamma_0 + \boldsymbol{\beta}_0^\top \boldsymbol{Z} \quad \text{and} \quad g_r(\alpha(\boldsymbol{Z})) = \gamma_r + \boldsymbol{\beta}_r^\top \boldsymbol{Z}$$

where the link function is typically $\log(\cdot)$ when parameters are constrained to be positive.

If the hazard function can be factorized as:

$$h(t|\mu(\boldsymbol{Z}), \alpha(\boldsymbol{Z})) = h_0(t|\alpha) \cdot \mu(\boldsymbol{Z}),$$

then the model represents a proportional hazards framework, where the hazard ratio between two groups remains constant over time.

The model is fitted by maximizing the likelihood over the observed data x. The likelihood function computed over each individual i, assuming exactly known transition times and accounting for right-censoring ($\delta_i = 0$), is given by:

$$\mathcal{L}(\boldsymbol{\theta} \mid \mathrm{t}, \delta) = \prod_{i:\delta_i=1} f_i(t_i) \cdot \prod_{i:\delta_i=0} S_i(t_i)$$

where the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{\beta})$ is estimated by maximizing the likelihood function.

Defining K as the number of all possible transitions, m_k the observations pertaining to transition k, the likelihood function can be expressed as:

$$\mathcal{L}(\boldsymbol{\theta} \mid x) = \prod_{k=1}^{K} \prod_{j=1}^{m_k} \mathcal{L}_{jk}(\boldsymbol{\theta} \mid x_{jk})$$

If the parameter vector can be partitioned as $\boldsymbol{\theta} = (\boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta_K})$ with independent components for each transition k, then the likelihood decomposes into K transition-specific likelihoods that can be maximized separately, reducing computational complexity.

However, when parameters or covariate effects are shared across transitions, joint likelihood maximization is necessary. Independent parameters alone do not guarantee the feasibility of maximizing partial likelihoods separately, but this is often possible under the assumption of continuously observed transition times.

Implementation Details

In this work, we use the `flexsurv` package in `R` to fit a parametric multi-state model assuming exact transition times. This package is particularly powerful because it provides extensive support for defining transition-specific hazard functions using parametric distributions, including exponential, Weibull, Gompertz, and more.

For multi-state models, `flexsurv` structures each individual's path over feasible transitions as described in Section 2.2. The model is fitted by maximizing the likelihood over

the observed data using the `optim` function in R. The default optimization method is the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm, which uses analytic derivatives when available, ensuring computational efficiency.

Some advantages of using this package include:

- **Flexibility in modeling transition hazards**: `flexsurv` allows for the use of a variety of parametric distributions to model the transition hazards over time, which can be adjusted to the specific characteristics of the data.
- **Extrapolation capabilities**: one of the key advantages of the package is its ability to generate predictions beyond the range of observed data, facilitating long-term projections or analysis in situations with limited follow-up.
- **Comprehensive estimation tools**: `flexsurv` directly computes key quantities such as transition probabilities, state occupation probabilities, and expected time spent in each state, derived from the estimated parametric hazard functions.
- **Independent estimation of transition-specific hazard functions**: the package offers a way to estimate hazards for each transition separately, which increases the efficiency of the computations.

However, there are some limitations to be aware of when using `flexsurv`:

- **Dependence on distribution selection**: the accuracy and interpretability of the model heavily depend on the appropriateness of the chosen parametric distribution. If the distribution does not accurately reflect the underlying hazard, the estimates may be biased or misleading.
- **Lack of interval-censoring support**: `flexsurv` does not inherently accommodate interval-censored data, which is a significant limitation when dealing with datasets where the exact timing of transitions is not known.
- **Bias from unobserved transitions**: the package assumes that all transitions are observed exactly, which may not be the case in real-world panel data settings.

In summary, `flexsurv` provides a robust and flexible framework for fitting parametric multi-state models but its lack of support for interval censoring remains a significant challenge when analyzing datasets with panel data or uncertain transition timings. In such cases, alternative methods might be necessary to properly account for this uncertainty.

2.4. Time-homogeneous multi-state model for interval-censored data

The `msm` package in `R` was specifically implemented to enable the fitting of general multi-state models to panel data, making it one of the most popular choices among researchers working with intermittent observation schemes. Its design and purpose make it intuitive and well-documented, allowing even non-experts to incorporate interval censoring into their analyses effectively. This accessibility has secure `msm`'s reputation as a go-to solution for addressing the complexities of multi-state modeling in panel data.

However, despite its strengths, the `msm` package comes with notable limitations. One significant restriction is its strict reliance on the Markovian assumption, where the future evolution of the process depends only on the current state and not on its history. Moreover, the package enforces, by default, an exponential baseline for the transition intensities, which assumes that the hazard rate for a transition remains constant over time. While this assumption facilitates computational efficiency by allowing a closed-form likelihood function for estimation, it can often be unrealistic. For instance, in an illness-death model, the assumption implies that once an individual develops dementia, the risk of death remains constant over time. This simplification is unlikely to hold in real-world scenarios, where the progression of dementia might lead to an increased hazard of death over time.

Despite these limitations, the `msm` package offers a range of useful features. It can be used to fit Markov models with any number of states and arbitrary transition patterns to panel data. Additionally, it includes several extensions, such as hidden Markov models (HMMs) and models where transition intensities vary with individual-specific or time-varying covariates.

While `msm` may often be the most obvious choice due to its ease of use and computational efficiency, it is important to critically assess whether it is also the best choice for accurately capturing the complexities of the data at hand. In fact, enforcing exponential form of the baseline might be harmful when does not align with the true underlying hazard structure.

Theoretical details

In a time-homogeneous continuous-time Markov model, the sojourn time in an arbitrary state r is exponentially distributed with a mean of $-\frac{1}{h_{rr}}$, where h_{rr} represents the negative of the total outgoing transition intensity from state r. The probability that an individual

in state r transitions directly to state s is given by $-\frac{h_{rs}}{h_{rr}}$.

The likelihood function for such a model is derived from the transition probability matrix $\mathbf{P}(t)$. For a time-homogeneous process, the element $p_{rs}(t)$ of $\mathbf{P}(t)$ represents the probability of being in state s at time $t+u$, given that the process was in state r at time u and doesn't depend on u itself. This probability does not provide specific information about when the transition occurred within the interval $[u, t+u]$ or whether other states were visited during this period. Given the closed form of KFE [1.2], transition probabilities can be computed using the matrix exponential of the scaled transition intensity matrix $\mathbf{Q}$ and applying eigensystem decomposition. For simple models, it is feasible to analytically compute each element of $\mathbf{P}$(t). This second approach, when achievable, is faster and avoids numerical instability.

For the illness-death model with no recovery, the transition intensity matrix $\mathbf{Q}$ is defined as:

$$\mathbf{Q} = \begin{bmatrix} -(h_{01}+h_{02}) & h_{01} & h_{02} \\ 0 & -h_{12} & h_{12} \\ 0 & 0 & 0 \end{bmatrix}$$

where h_{01}, h_{02}, and h_{12} denote the transition intensities between the respective states. Assuming $h_{01}+h_{02}=h_{12}$, the transition probabilities $p_{rs}(t)$ obtained by solving the Kolmogorov Forward Equations (KFE) are:

$$\begin{aligned}
p_{00}(t) &= e^{-(h_{01}+h_{02})t}, \\
p_{01}(t) &= h_{01}te^{-(h_{01}+h_{02})t}, \\
p_{02}(t) &= e^{-(h_{01}+h_{02})t}\big(-1+e^{(h_{01}+h_{02})t}-h_{01}t\big), \\
p_{10}(t) &= 0, \\
p_{11}(t) &= e^{-h_{12}t}, \\
p_{12}(t) &= 1-e^{-h_{12}t}, \\
p_{20}(t) &= 0, \\
p_{21}(t) &= 0, \\
p_{22}(t) &= 1
\end{aligned}$$

The likelihood function is conveniently expressed in terms of these transition probabilities. The specific likelihood formulation, varying depending on the observational scheme, has already been detailed within the mathematical framework [1.4]. The likelihood formulation properly accounts for the presence of unknown transition times and censored events.

Implementations details

In `msm` package data are organized as a series of observations grouped by patient, with each patient identified by a unique identification number. The dataset should include a data frame containing variables that specify the observation time and the observed state of the process . The model specification involves defining the generator matrix $\mathbf{Q}$, which determines the instantaneous transitions of the Markov process. Transition intensities can depend on covariates, allowing the generator matrix $\mathbf{Q} = \mathbf{Q}(\mathbf{Z})$ to vary with individual-specific or time-varying characteristics. In order to calculate transition probabilities $\mathbf{P}(t, \mathbf{Z}(t))$ on which the likelihood depends, time dependent covariates are assumed to be piecewise-constant.

Fitting the model entails finding estimates of the unknown transition intensities by maximizing the likelihood. The latter is maximized using numerical methods, which require an initial set of values to begin the optimization process. To assist users without prior knowledge on the process, the `msm` package provides a default function that generates reasonable starting values for the non-zero entries of the $\mathbf{Q}$ matrix. To ensure that the true maximum likelihood estimates are obtained, it is recommended to run the model multiple times, each time starting from different initial values. Internally, the optimization process leverages the R function `optim`.

The `msm` package also includes functions to extract key quantities from the model fit, such as the estimated transition intensity matrix and its confidence intervals, transition probabilities, mean sojourn times, total length of stay in transient states, hazard ratios, and other relevant metrics. Confidence intervals can be computed using bootstrap methods when asymptotic standard errors are underestimated.

When overly complex models are applied to insufficient data, the model parameters may become unidentifiable. As a result, the optimization algorithm may fail to locate the maximum of the log-likelihood or, in some cases, be unable to evaluate the likelihood altogether.

Relaxing time-homogeneity assumption by introducing a piecewise time-varying covariate

The idea behind this procedure stems from the fact that, as we have highlighted, the `msm` package is a functional, intuitive, well-documented tool, rich in utilities for extracting information from model fits. However, the fact that the baseline hazard is constrained to an exponential form can be limiting.

The Gompertz distribution allows for exponentially changing hazards, which means it can model both increasing and decreasing hazards over time. This property makes it particularly useful for modeling aging processes, where the rate of event occurrence is assumed to accelerate over time.

The Gompertz hazard function is given by:

$$h(t) = \lambda \cdot \exp(\gamma t) \tag{2.1}$$

where λ is the scale parameter, and γ represents the rate of change of the hazard with respect to time.

The Gompertz distribution is therefore highly suitable for our context, and we would like to model it within `msm`. However, this is not directly possible in `msm`, so we aim to approximate it. By introducing a continuous time-dependent covariate representing the time scale on which we are modeling the process, we obtain the following hazard function:

$$h(t) = \lambda \cdot \exp(\beta t) \tag{2.2}$$

Here, λ is the constant transition intensity predicted by `msm`, and $\exp(\beta t)$ represents the effect of the covariate on the transition.

This covariate is often represented for each patient by his age, as it is often used as a time scale in multi-state models. Additionally, the patient's age has a clear interpretative meaning and can provide useful insights into the model.

We can observe the similarity between the expressions (4.1) and (2.2). However, We should keep in mind that `msm` handles time-varying covariates in a piecewise manner, so the effect of that covariate can be estimated assuming that it is constant in between the times that it is observed, so that $\mathbf{P}(u; t+u) = \mathbf{P}(t)$. Hence, expression (2.2) would be an approximation of the Gompertz hazard function.

While models with piecewise constant intensities allow for temporal non-homogeneity, they are limited in several ways. The piecewise structure requires estimation of all transition intensities within each interval. When few transitions are observed, the estimates are highly uncertain. Further, these models require the assumption of homogeneity within intervals, which is quite limiting for a highly temporally volatile disease process.

2.5. Parametric time-inhomogeneous multi-state model for interval-censored data

In the analysis of multi-state models, the `msm` package has proven to be a robust and widely-used tool for fitting multi-state models with panel data. However, it is often necessary to model baseline hazard functions as time-dependent, particularly in scenarios where time-varying intensities better reflect the underlying biological processes.

To address these limitations, we opted to leverage the `nhm` package for fitting parametric time-inhomogeneous multi-state Markov models. This package extends the functionalities of `msm` by enabling the modeling of smoothly changing transition intensities while preserving compatibility with interval-censored data. Misclassification type hidden models with non-homogeneous transition intensities can also be fitted. The `nhm` package provides flexibility for specifying parametric intensity functions, including Gompertz, Weibull, and B-spline forms, and even allows for user-defined intensity functions to accommodate bespoke modeling needs.

One of the key challenges in fitting such models lies in solving the Kolmogorov Forward Equations (KFE), which are often intractable for most forms of generator matrices $\mathbf{Q}(t)$. This necessitates numerical solutions to nonlinear partial differential equations, which are implemented in `nhm` using advanced numerical solvers.

Theoretical details

Consider i indexing N individuals. The data for individual i consist of a series of observation times $(t_{i0}, \ldots, t_{in_i})$ and corresponding states $(S(t_{i0}), \ldots, S(t_{in_i}))$. The likelihood function for a single individual is computed as:

$$\mathcal{L}_i(\boldsymbol{\theta}) = \prod_{j=0}^{n_i-1} p_{S(t_j)S(t_{j+1})}(t_j, t_{j+1}; \boldsymbol{\theta})$$

The entries of the matrix $\mathbf{P}(t)$ are computed by leveraging the solutions to the KFE 1.1. For additional efficiency in optimization, Titman [27] proposed to solve the extended system of differential equations, incorporating the systems of equations defining the derivatives with respect to the parameter vector $\boldsymbol{\theta}$:

$$\frac{d\mathbf{P}'(t_0, t; \boldsymbol{\theta})}{dt} = \mathbf{P}'(t0, t, \boldsymbol{\theta}) \cdot \mathbf{Q}(t; \boldsymbol{\theta}) + \mathbf{P}(t0, t, \boldsymbol{\theta}) \cdot \mathbf{Q}'(t; \boldsymbol{\theta}) \quad \mathbf{P}'(t0, t0; \boldsymbol{\theta}) = 0 \tag{2.3}$$

$$\mathbf{P}'(t0, t; \boldsymbol{\theta}) = \frac{\partial \mathbf{P}(t0, t; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \quad \text{and} \quad \mathbf{Q}'(t; \boldsymbol{\theta}) = \frac{\partial \mathbf{Q}(t; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \tag{2.4}$$

The differential equations are solved by using the *LSODA* routine of the `deSolve` package, it adaptively selects step sizes and determines whether to use stiff ODE solvers, ensuring accurate and efficient computation.

A computationally convenient approach to compute the likelihood, implies first solving a single system of differential equation (initial value problem) with starting time t_{i0} for all observation times $(t_{i1}, \ldots, t_{ni})$. The required transition probabilities between two arbitrary times t_j and t $(t > t_j)$ are then derived using:

$$\mathbf{P}(t_j, t; \boldsymbol{\theta}) = \mathbf{P}(t_0, t_j; \boldsymbol{\theta})^{-1} \mathbf{P}(t_0, t; \boldsymbol{\theta}).$$

This approach is computationally efficient since $\mathbf{P}(t_0, t_j; \boldsymbol{\theta})^{-1}$ can typically be obtained quickly, for example for progressive models where $\mathbf{P}(t_0, t_j; \boldsymbol{\theta})$ is upper triangular. However, there are some cases in which inversion arises problems. The transition probability matrix may become singular if t_0, t_j are far apart and, for instance, the probability of remaining in a state over that time gets very close to zero. Singularities can be avoided, by supplying a vector of split times. Separate initial value problems for any intervals within new start times, determined by the split times, will be solved. Adding splits will tend to increase the computation time, but not substantially. This is realized in `nhm` by using the splits option within *nhm.control*.

In case $\mathbf{Q}(t; \mathbf{Z})$ is affected by a set of covariates $\mathbf{Z}$, the transition probabilities are computed by solving a single initial value problem for each unique covariate pattern. As a consequence, models with continuous covariates can be substantially more computationally demanding. It may be desirable for large datasets with continuous covariates to consider an approximation to the likelihood based upon coarsening the set of unique covariate values. A simple method based on using K-means clustering involves grouping similar values of the continuous covariates and assuming approximating transition probabilities within the same cluster by the transition probability that would arise from the mean covariate values within that cluster. While coarsening the covariates will tend to introduce some attenuation bias in the covariate effects, it may be useful either in the model building stage or in cases where the full model is computationally impractical.

Implementation details

Data are prepared as in the `msm` package 2.4. The model specification involves defining the generator matrix $\mathbf{Q}(t)$. There are different arguments fundamental to define its nature. The `type` argument specifies the type of non-homogeneous model for the intensity matrix of the Markov process, among those mentioned previously. The number of states in the model and the set of admissible transitions is specified using the `trans` argument, with the possibility of constraining some transitions to have the same baseline parameters. The `nonh` argument specifies which transition intensities are non-homogeneous with respect to time. Finally, `covm` argument specifies which transitions are affected by covariates.

Fitting the model requires finding estimates of the unknown transition intensities by maximizing the likelihood. The likelihood function is maximized using numerical methods, which requires an initial set of values to begin the optimization process. `nhm` provides a function to generate a guess, specifically it computes initial parameters for the baseline transition intensities by assuming a homogeneous Markov model with no covariates effect and that the observation times correspond to the exact transition times.

The likelihood function is maximized using numerical optimization techniques. The default method is the Berndt–Hall–Hall–Hausman (BHHH) algorithm. This method is preferred over the BFGS algorithm because it tends to be more robust in situations where the objective function is noisy or has a poor condition number. BHHH relies on the gradient of the log-likelihood that is directly available thanks to the extended system introduced in 2.3.

For models without censoring or misclassification, the Fisher scoring algorithm is offered as an option. Fisher scoring is advantageous when the likelihood is well-behaved, and it converges faster compared to methods like BHHH in these conditions.

The package provides the score test option, which is most useful as an exploratory step to determine which, if any, transition intensities exhibit time non-homogeneity, and in which direction. The advantage of the score test is that it is not necessary to fit a more complicated model in order to assess whether it may fit better than the basic model. This approach allows us to assess time-dependence in the transition intensities without the need to explicitly fit non-homogeneous models initially.

The `nhm` package offers several strengths, particularly its ability to fit non-homogeneous Markov models with smoothly changing transition intensities. However, it is particularly sensitive to the choice of starting values for the optimization procedure, which may affect the convergence of the model. Additionally, the numerical solution approach is compu-

tationally intensive, especially when dealing with large datasets or continuous covariates. To address this challenge, the package offers parallelization capabilities, which allow for the parallel solving of systems of ODEs. This is especially useful when there are multiple covariate patterns to consider, as it can significantly reduce computation time and improve efficiency when fitting complex models.

2.6. Summary Table of key attributes

This tables summarizes the most distinctive features of the discussed methods, highlighting their limitations and noteworthy characteristics. It also includes some cutting-edge methods that we thoroughly investigated but ultimately did not use in the simulation study for the reasons outlined in the introduction of the chapter.

	surival	flexsurv	msm	nhm	SmoothHazrd	smms
Markov assumption	Yes	Yes	Yes	Yes	Yes	No
Semi-Markov assumption	Yes	Yes	No	No	No	Yes
Models interval censoring	No	No	Yes	Yes	Yes	Yes
Time-homogeneity assumption	No	No	Yes	No	No	No
Parametric baseline hazards	No	Yes	Yes (Exponential)	Yes	Yes (only Weibull)	Yes
Handles covariates	Any kind	Any kind	Time-varying piecewise constant	Continuous clustered in groups	Any kind	Any kind
Strengths	Well-documented, intuitive	Independent MLE for each transition	Closed-form of KFE exists	Direct numerical solution of KFE	Comprehensive functionality	Parametric distributions of sojourn times
Limitations	Assumes independent transitions and exact observation times	Assumes exact observation times	Assumes time-homogeneity	Computationally intensive, sensitive to initialization	Limited to illness-death models	Does not account for delayed entry with age as time scale

Table 2.1: Key attributes of methods under analysis

3 | A novel method for multi-state models with panel data

This chapter will focus on introducing an entirely new strategy, referred to as **Multiple Imputation for Panel Data (MIPD)**. The concept behind this method arises from the need to preserve the intuitiveness and functionality of widely used approaches in the literature, such as those developed in the `survival` and `flexsurv` packages, while also introducing the ability to account for and model the uncertainty associated with intermittently observed data. This methodology draws inspiration from Yu's article [33] and the accompanying script, which have been adapted to the specific context of our study.

Specifically, the method was implemented for a progressive illness-death model, where the disease state is represented by the onset of dementia. For each patient $i = 1, ...N$, the exact age at study entry and at exit/death is known, while the transition time to the dementia state, if observed, is subject to interval censoring. Additionally, for those patients whose transition to dementia is not directly observed, it cannot be ruled out that dementia may have occurred between the last visit and the time of exit from the study.

Our approach is highly adaptable, enabling the incorporation of covariates into the regression to account for variability between patients. Additionally, with appropriate modifications, it can be implemented under either the Markovian assumption or the Semi-Markovian assumption, allowing for consideration of the time spent in the dementia state.

The implementation of this method is divided into two main phases:

- In the first phase, a **multiple imputation** method is used to impute the exact time of dementia onset for patients who developed the disease and to impute the disease status for patients whose onset was not directly observed. The choice of multiple imputation over standard imputation methods is based on several key advantages. It preserves the variability and uncertainty associated with missing data, generating M complete datasets each representing a plausible scenario based on the observed data. Moreover, it enhances the robustness and reduce the bias of estimates compared to

methods that impute a single value per observation.

- In the second phase, the M complete datasets generated during the imputation process are used to **fit M parametric multi-state models**, one for each dataset, incorporating the imputed clinical histories and transition times of the patients. This approach allows us to maintain a tractable estimation framework thanks to the assumption that the transition times occur exactly at the imputed time points [2.3]. The choice of parametric models over semi-parametric ones is justified by their numerous advantages, as discussed in earlier sections.

Given that multiple imputed datasets produce multiple sets of parameter estimates, a proper combination of results is required to obtain valid statistical inferences. This is achieved through *Rubin's rules* [18], which provide a framework for combining parameter estimates across multiple imputations while accounting for both within- and between-imputation variability.

Pooled parameter estimates

Let $\hat{\boldsymbol{\theta}}_m$ be the estimated parameter from the m-th imputed dataset, for $m = 1, \ldots, M$. The final pooled estimate is obtained as the arithmetic mean of the estimates across all imputations:

$$\bar{\boldsymbol{\theta}} = \frac{1}{M} \sum_{m=1}^{M} \hat{\boldsymbol{\theta}}_m \tag{3.1}$$

Variance estimation

The total variance of the pooled estimate consists of two components:

- The *within-imputation variance*, which measures the average variability of estimates within each imputed dataset:

$$W = \frac{1}{M} \sum_{m=1}^{M} \hat{V}_m \tag{3.2}$$

  where $\hat{V}_m$ is the estimated variance of $\hat{\boldsymbol{\theta}}_m$ in the m-th imputed dataset.

- The *between-imputation variance*, which captures the variability of the estimates across different imputations:

$$B = \frac{1}{M-1} \sum_{m=1}^{M} (\hat{\boldsymbol{\theta}}_m - \bar{\boldsymbol{\theta}})^2 \tag{3.3}$$

The total variance of the pooled estimate is given by:

$$T = W + \left(1 + \frac{1}{M}\right) B \tag{3.4}$$

Confidence intervals

The standard of $\bar{\boldsymbol{\theta}}$ is given by $\sqrt{T}$. A $(1-\alpha)$ confidence interval for $\bar{\boldsymbol{\theta}}$ is:

$$\bar{\boldsymbol{\theta}} \pm z_{1-\alpha/2}\sqrt{T}, \tag{3.5}$$

where $z_{1-\alpha/2}$ is the critical value from the standard normal distribution.

Although confidence intervals are often computed using the t-distribution to account for the additional uncertainty introduced by the imputation process, we instead use the z-distribution. This choice is motivated by two factors: first, we chose a sufficiently large number of imputations (M) so that the degrees of freedom are high enough, making the t-distribution approximation unnecessary; and second, to ensure consistency with the computation of confidence intervals in other methods used in the simulation study.

3.1. Multiple imputation under Markovian assumption

In order to impute the quantities of interest from the original dataset, we require certain measures related to the development of dementia. To extract these, we fit two Cox models based on the initial dataset, ignoring the phenomenon of interval censoring:

- **Onset Model:** the first model focuses on the development of dementia as event of interest. Before fitting this model, we used the midpoint imputation technique [2.2] for individuals who developed the disease.
- **Death Model:** the second model is used to shape the transition to death. In this model, the effect of dementia onset on the response is accounted for by a coefficient δ.

Both models are capable of considering the effect of arbitrary covariates $\mathbf{Z}$.

After fitting the models, for each patient i we computed:

- the survival function for living disease-free

$$S_1(t|\mathbf{Z}_i) = e^{-H_{01}(t|\mathbf{Z}_i)-H_{02}(t|\mathbf{Z}_i)}$$

- the survival function adjusted for the dementia status

$$S_2(t|\mathbf{Z}_i) = e^{-H_{02}(t|\mathbf{Z}_i)+\delta}$$

- two vectors of length M of uniformly distributed values in [0,1]

$$c_1(i) \sim \text{Uniform}(0,1) \quad c_2(i) \sim \text{Uniform}(0,1)$$

Step 1 : **Imputation of Transition Times**

For patients for whom dementia development was observed at time b_i, with the preceding visit at time a_i, the transition time was imputed within the interval $[a_i, b_i]$.

We defined the quantity $\bar{p}_i(t|\mathbf{Z}_i)$ as:

$$\bar{p}_i(t|\mathbf{Z}_i) = \frac{S_1(t|\mathbf{Z}_i) \cdot h_{01}(t|\mathbf{Z}_i)}{S_2(t|\mathbf{Z}_i)}$$

which represents the probability of developing dementia at time t, conditioned on survival until time t. The cumulative version of this quantity, denoted $\bar{P}(t|\mathbf{Z}_i)$, is then calculated. Additionally, we computed the normalized version of this cumulative quantity using a min-max normalization:

$$\bar{P}_{\text{norm}}(t|\mathbf{Z}_i) = \frac{\bar{p}_i(t|\mathbf{Z}_i) - \bar{p}_i(a_i|\mathbf{Z}_i)}{\bar{p}_i(b_i|\mathbf{Z}_i) - \bar{p}_i(a_i|\mathbf{Z}_i)}$$

For each dataset $m = 1, ...M$, we imputed the transition time $\bar{t}$ as follows:

- if $\bar{P}_{\text{norm}}(b_i|\mathbf{Z}_i) > \bar{P}_{\text{norm}}(a_i|\mathbf{Z}_i) \quad \bar{t} = \max\{t \mid \bar{P}_{\text{norm}}(t|\mathbf{Z}_i) \leq c_1(i)[m]\}$
- if $\bar{P}_{\text{norm}}(b_i|\mathbf{Z}_i) \leq \bar{P}_{\text{norm}}(a_i|\mathbf{Z}_i) \quad \bar{t} = \frac{(a_i+b_i)}{2}$

In the first case, the inverse transform sampling technique for a discrete distribution was used to impute the value of the transition time $\bar{t}$ [24].

Step 2 : **Imputation of Dementia Status**

For patients for whom dementia development was not observed within the interval $[a_i, b_i]$,

where a_i corresponds to the last observed visit and b_i to the time of death or dropout, we then proceeded with imputing the dementia status.

We defined the quantity :

$$\bar{P}_{\text{adj}}(b_i|\mathbf{Z}_i) = S_2(b_i|\mathbf{Z}_i) \cdot e^{\delta \cdot \text{death}_i + \boldsymbol{\beta}_{02} \cdot \mathbf{Z}_{i,02}} \cdot (\bar{P}(b_i|\mathbf{Z}_i) - \bar{P}(a_i|\mathbf{Z}_i)) \tag{3.6}$$

The idea behind this quantity is to adjust the probability of developing dementia in the considered interval, conditioned on survival up to time b_i, on the effect of the covariates affecting transition to death and on the occurrence of death by the coefficient δ. The reason is that a higher value of the survival function at death/dropout implies a lower probability that the patient dies without first transitioning in dementia state. Moreover, the occurrence of death should increase the probability of developing dementia, as the death of the patient might have prevented us from observing the onset of the disease.

We then calculated:

$$\hat{P}_{\text{adj}}(b_i|\mathbf{Z}_i) = \frac{\bar{P}_{\text{adj}}(b_i|\mathbf{Z}_i)}{S_1(b_i|\mathbf{Z}_i) + \bar{P}_{\text{adj}}(b_i|\mathbf{Z}_i)} \tag{3.7}$$

which represents the ratio of developing dementia during the interval of interest with respect to the total probability of surviving until the end of the interval.
For each dataset $m = 1, ...M$, we imputed the dementia status as follows:

- if $\hat{P}_{\text{adj}}(b_i|\mathbf{Z}_i) \geq c_2(i)[m] \rightarrow$ dementia onset $= 1$
- if $\hat{P}_{\text{adj}}(b_i|\mathbf{Z}_i) < c_2(i)[m] \rightarrow$ dementia onset $= 0$

If the dementia status was imputed as positive, the transition time within the interval $[a_i, b_i]$ was imputed as described in Step 1.

3.2. Multiple imputation under Semi-Markovian assumption

In this section, we discuss the multiple imputation approach applied to panel data under the Semi-Markov assumption, incorporating modifications to account for the time spent in the dementia state. Before fitting any model, we used the midpoint imputation technique on the original dataset as explained in 2.2.

We fit a single Cox proportional hazards model stratified by transition type. Specifically:

- The transition from the dementia-free state to either dementia or death was modeled on the age scale, starting from the study entry age for each patient.
- The transition from the dementia state to death was modeled using the *clock-reset* principle. Here, the time scale resets to the time elapsed since the diagnosis of dementia. For this transition, we introduced the effect of a covariate through a coefficient δ, representing the patient's age at diagnosis. This adjustment allows us to capture the influence of age on the risk of death, which is realistic since age is expected to affect the mortality risk.

Finally, to enhance flexibility, the model accounts for an arbitrary set of covariates $\mathbf{Z}$ without prior constraints, allowing these covariates to influence the examined transitions differently.

For each patient $i = 1, ...N$, the following survival functions were computed:

- Survival function for living disease-free:

$$S_1(t \mid \mathbf{Z}_i) = e^{-H_{01}(t|\mathbf{Z}_i) - H_{02}(t|\mathbf{Z}_i)}$$

- Survival function for living:

$$S_2(t \mid \mathbf{Z}_i) = e^{-H_{02}(t|\mathbf{Z}_i)}$$

- Survival function for living with the disease:

$$S_{2,\text{dem}}(t \mid \mathbf{Z}_i) = e^{-H_{12}(t|\mathbf{Z}_i)}$$

Additionally, two vectors of length M of uniformly distributed random values in $[0, 1]$ were generated for imputation purposes:

$$c_1(i) \sim \text{Uniform}(0, 1), \quad c_2(i) \sim \text{Uniform}(0, 1)$$

Step 1 : Imputation of Transition Times

The imputation of transition times was performed analogously to the approach described in the previous section 3.1. This choice is justified by the fact the Semi-Markov assumption does not impact the transition to dementia and death without dementia, preserving the quantities and procedures outlined under the Markov assumption.

Step 2 : Imputation of Dementia Status

For patients whose dementia onset was not observed, the dementia status was imputed within the interval $[a_i, b_i]$, where a_i is the last observed visit and b_i is the time of death or dropout. We defined the adjusted probability:

$$\begin{aligned}\bar{P}_{\text{adj}}(b_i \mid \mathbf{Z}_i) = \big[\bar{P}(b_i \mid \mathbf{Z}_i) - \bar{P}(a_i \mid \mathbf{Z}_i)\big] \cdot \big[S_2(b_i \mid \mathbf{Z}_i) e^{\boldsymbol{\beta}_{02} \cdot \mathbf{Z}_{i,02}} \\ + \text{death}_i \big(1 - S_{2,\text{dem}}(b_i - a_i \mid \mathbf{Z}_i)\big) e^{\boldsymbol{\beta}_{12} \cdot \mathbf{Z}_{i,12} + \delta a_i}\big].\end{aligned} \tag{3.8}$$

This quantity adjusts the conditional probability of developing dementia in the interval of interest, based on the following factors:

- Survival up to time b_i adjust by the covariates affecting transition to death without disease: a higher value of the survival function at b_i indicates a lower probability of natural death, thereby increasing the likelihood of dementia development.
- Effect of observing death for patient i, accounted for by the following contributes:
  - The complementary quantity of a survival function for living with the disease over a time interval corresponding to $(b_i - a_i)$, which spans from the last visit without the disease to the time of death. We consider this contribution because we aim for the probability to increase as the survival function decreases. This reflects the idea that the longer the time spent with dementia, the more likely it is that death occurred specifically due to this unobserved condition.
  - The effect of the patient's age (a_i) at diagnosis by the coefficient δ and covariates affecting the transition from dementia to death, which contribute to increasing the adjusted probability.

The ratio of developing dementia during the interval of interest with respect to the total probability of surviving until the end of the interval $\hat{P}_{\text{adj}}(b_i \mid \mathbf{Z}_i)$ is computed as in 3.7.

For each dataset $m = 1, \ldots, M$, the dementia status was imputed as follows:

- If $\hat{P}_{\text{adj}}(b_i \mid \mathbf{Z}_i) \geq c_2(i)[m]$, then dementia onset = 1.
- If $\hat{P}_{\text{adj}}(b_i \mid \mathbf{Z}_i) < c_2(i)[m]$, then dementia onset = 0.

If dementia onset was imputed as positive, the transition time within $[a_i, b_i]$ was determined using the method outlined in paragraph 3.1.

4 | Simulation Study

Simulation studies are computer-based experiments that involve generating data through pseudo-random sampling. They serve as a powerful tool in statistical research, allowing us to evaluate and compare different methodological approaches under controlled conditions. A key advantage of simulation studies lies in their ability to provide insights into the behavior of statistical methods, as the true underlying values of parameters—referred to as the "ground truth"—are known by design. This unique feature enables an in-depth assessment of various properties of the methods under investigation.

In our study, we leverage the strengths of simulation to evaluate different methodologies for modeling **progressive illness-death processes** under an intermittent observation scheme. The process $X = (X(t), t \geq 0)$ takes values in $\{0, 1, 2\}$ [1.4], subjects are initially disease-free ($X(0) = 0$) and may transition to a state representing dementia ($0 \to 1$) and die ($1 \to 2$), or directly die without disease ($0 \to 2$). Recovery from dementia is not considered, making the process irreversible.

To ensure a structured and comprehensive approach to our simulation study, we follow the ADEMP framework [21], which consists of the following five key components:

- **Aims:** Defining the primary objectives and research questions of the study.
- **Data-generating mechanisms and Study Design:** Specifying the processes and assumptions used to create simulated datasets and the scenarios in which we want to evaluate them.
- **Estimands:** Identifying the key quantities of interest that will be estimated.
- **Methods:** Describing the statistical methods under evaluation, including model specifications and estimation techniques.
- **Performance measures:** Establishing criteria to assess the accuracy and reliability of the methods.

By structuring our study according to the ADEMP framework, we aim to ensure clarity, reproducibility, and complete evaluation of the statistical methods under investigation.

4.1. Aims

The primary aim of this simulation study is to test and analyze different methods for multi-state models applied to panel data across various scenarios. Specifically, we seek to address critical methodological questions that researchers commonly encounter when employing multi-state models with this kind of data. These challenges are explored through the following key aspects:

- **Determining the role of the sample size in obtaining consistent estimation**

  A fundamental requirement in statistical modeling is that estimators should be consistent, meaning that as the sample size increases, the estimated parameters converge to their true values. However, determining the exact minimum sample size required to ensure consistency is challenging, as it depends on various factors such as model complexity and data characteristics. While we do not aim to establish this precise threshold, we will investigate how the consistency of our estimators varies with sample size, providing insights into the robustness of our results and indicating the range where estimates become more reliable.

- **Impact of different observation schemes**

  A key aspect of this study is to investigate how different observation schedules influence the analysis and the accuracy of estimated parameters in multi-state models. Specifically, we examine the following factors:

  - Regular vs. irregular observation grids: in a regular observation scheme, data are collected at fixed, evenly spaced time intervals, ensuring a structured and predictable dataset. In contrast, irregular observation schedules involve data collection at unevenly spaced intervals, introducing additional complexity in the estimation process. Since not all events are observed following the same pattern, the estimation of transition-specific quantities may be affected. For instance, if certain events, such as the onset of dementia, are underrepresented and recorded at highly irregular intervals, this irregularity may introduce challenges in accurately estimating the transition probabilities and associated measures.

  - The impact of varying the length of observation intervals: the frequency of observations can significantly influence the quality of the data. Shorter intervals allow for more precise tracking of state transitions, reducing the consequences of interval censoring, whereas longer intervals may lead to increased uncer-

tainty about the exact timing of transitions. However, in population-based studies, conducting frequent visits requires substantial financial and logistical resources, increasing the overall cost and effort of data collection. Balancing study feasibility with the accuracy of parameter estimation is therefore crucial, as more frequent observations may not always be practical in large-scale studies.

- **Effects of incorrectly specifying the functional form of transition intensities**

  The functional form of the transition intensities plays a pivotal role in multi-state models, as it determines how the instantaneous risk of transitioning between states is modeled over time. In most practical applications, however, the true parametric form of these intensities is unknown. Consequently, during the study design phase, researchers can at best rely on prior knowledge to make an informed assumption about the form of the intensities, choosing one functional specification over another. Moreover, overly simplistic functional forms are often selected to facilitate estimation and reduce computational burden. While such choices can make model implementation more feasible, they risk oversimplifying the transition dynamics.

  Hence, it is of fundamental importance to assess how potential misspecifications in the assumed functional form can impact the reliability of the estimated model parameters. Incorrect assumptions may introduce bias, affect the precision of estimates, and ultimately bring to misleading conclusions about the underlying process.

- **Implications of neglecting interval censoring**

  In panel data, transitions between states are typically observed only at discrete time points, meaning that the exact transition times are unknown. A common simplifying assumption is to treat observation times as if they represent exact transition times, potentially introducing bias into the analysis. This study investigates the implications of neglecting interval censoring and assesses how this assumption impacts parameter estimates, providing insights into the importance of properly accounting for interval censoring in multi-state models.

4.2. Data-generating Mechanism

In our simulation study, data were generated by drawing parametric samples from a known model. Specifically, the parameter values were informed by population data from The Swedish National Study on Aging and Care in Kungsholmen, described in 1.3.

The choice of the parametric distribution for the transitions was guided by the need to model the instantaneous risk of transitioning as an increasing function of time, reflecting the reasonable assumption that the subjects' age positively influences this risk. Among the potential options, the Gompertz distribution was selected over the Weibull distribution for modeling transitions to dementia and death. This decision was made because the Gompertz distribution explicitly accounts for an exponential increase in hazard over time, which aligns better with the biological understanding of aging-related risks. The Gompertz hazard function for $t \geq 0$ is defined as:

$$h(t) = \lambda \cdot \exp(\gamma t) \tag{4.1}$$

where λ is the scale parameter, and γ determines the rate at which the hazard changes over time. This formulation provides an intuitive way to understand the impact of age, expressed by the time scale t, on the hazard function. Specifically, when $\gamma > 0$, the hazard increases exponentially with age, indicating a positive relationship between age and risk.

Computation of parameters' value

The dataset obtained from The Swedish National Study on Aging and Care in Kungsholmen comprises information on 2150 subjects and includes the following variables for each subject:

- Identifier code;
- Age at baseline and at each follow-up visit;
- Dates of each visit;
- Dementia status recorded at each visit;
- Mortality status, along with the recorded date of death (if applicable);
- A range of covariates of different kind.

The observation schedule varied based on patient age: subjects younger than 78 years were assessed at six-year intervals, whereas those aged 78 years or older were followed

up every three years. Each individual was observed a maximum of six times, baseline included.

The dataset was rearranged to include the following key variables for each patient:

- **Onset indicator:** true if dementia was observed in at least one wave, false otherwise
- **Onset age:** the age at which the onset indicator became true for the first time.
- **Death indicator:** true if the subject died during the study, false if it was lost to follow-up
- **Death age:** the age at which death occurred, if the death indicator was true, or the censoring time otherwise.

This restructuring allowed us to format the data as required for multi-state modeling, clearly defining the time intervals during which each patient was at risk for each possible transition and whether the transition occurred.

Two parametric multi-state models were fitted using the Gompertz distribution:

1. A **Markov multi-state model**, where each of the three transitions was modeled independently, using the age of the patient as the time scale.
2. A **Semi-Markov multi-state model**, where:
    - transitions from dementia-free status to dementia and to death were modeled on the time scale of age.
    - the time scale for transitions from dementia to death was reset to the time elapsed since the diagnosis of dementia.

We extracted the baseline parameter vectors $\boldsymbol{\lambda}$ and $\boldsymbol{\gamma}$ from the Markov and Semi-Markov multi-state models and subsequently we used them to generate data for the simulation study.

Data simulation

The data were simulated using the `hesim` package in R [7]. Like all statistical packages capable of Monte Carlo simulation, `hesim` utilizes a pseudo-random number generator, where each random number is a deterministic function of the current state of the generator. The state can be controlled by setting a seed, which ensures reproducibility of the results.

To assess the robustness of our estimations and ensure that the results are not driven by a specific dataset, **we conducted simulations using 100 different datasets** ($n_{sim} =$

100). This approach allows us to evaluate the stability and consistency of our findings across multiple replications, reducing the risk that observed results arise from random peculiarities in a single dataset. By averaging results over multiple datasets, we obtain a more reliable assessment of the estimands under various conditions.

For each seed, **data were simulated under both Markov and Semi-Markov assumptions**, using the Gompertz parametric distribution with the baseline parameters estimated from the real-world data. We set the number of individuals' trajectories to draw from the simulation process as $n_{obs} = 100.000$.

The hazard function for each transition is defined as:

$$h_k(t) = \lambda_k \cdot \exp(\gamma_k t) \cdot \exp\left(\boldsymbol{\beta}_k^T \mathbf{Z}_i\right), \quad k \in \{01, 02, 12\} \tag{4.2}$$

where:

- λ_k is the scale parameter for transition k,
- γ_k represents the rate of change of the hazard for transition k,
- $\boldsymbol{\beta}_k$ is the vector of covariate effects specific to transition k,
- $\mathbf{Z}_i$ represents the vector of covariates for an arbitrary individual i.

We introduced three different covariates into the simulation: one continuous covariate and two binary covariates, with prevalence rates of 15% and 30%, respectively.

To ensure the general applicability of our study, we did not specify the exact nature of these covariates. However, for illustrative purposes, the continuous covariate could represent physiological measures such as a patient's body mass index (BMI) or blood pressure, while the binary covariates could reflect conditions such as smoking status, the presence of diabetes, or other factors like educational level.

In our simulation, we considered different covariate effects, allowing them to influence the hazard of the admissible transitions in distinct ways. Their impact is represented in terms of hazard ratios and is summarized in Table 4.1.

We briefly define the the hazard ratio (HR) associated to a covariate. For a binary covariate Z (e.g., presence/absence of a condition), the hazard ratio is modeled as:

$$\text{HR} = \frac{h_0(t)e^{\beta \cdot 1}}{h_0(t)e^{\beta \cdot 0}} = e^{\beta}$$

For a continuous covariate Z, the hazard ratio for a unit increase in Z is:

$$\text{HR} = \frac{h_0(t)e^{\beta \cdot (Z+1)}}{h_0(t)e^{\beta \cdot Z}} = e^{\beta}$$

Transition	$\exp(\beta_1)$	$\exp(\beta_2)$	$\exp(\beta_3)$
Dementia-free → Dementia	1.0	2.0	1.5
Dementia-free → Death	1.5	2.0	1.5
Dementia → Death	1.5	1.0	1.5

Table 4.1: Hazard ratios for different covariates across transitions.

The interpretation of hazard ratios is based on the proportional hazards assumption, implying that a value greater than 1 suggests an increased risk of transition associated with the respective covariate, while a value equal to 1 indicates no effect.

From Table 4.1, it can be observed that:

- The first covariate has no effect on the transition from dementia-free state to dementia, meaning its corresponding regression coefficient is zero.
- The second covariate does not influence the transition from dementia to death.
- The third covariate consistently affects all transitions, with a constant hazard ratio.

In our simulation process, we also accounted for the presence of left truncation, meaning that we only generated trajectories for individuals who entered the study in the dementia-free state.

To further enhance the realism of our simulated cohort and reflect the heterogeneity of an aging population, we modeled the age at study entry as a random variable following a uniform distribution between 60 an 96 years. This approach allows us to capture a diverse range of entry ages, simulating a representative sample of older adults, and ensures that the study population includes individuals with varying baseline risks and exposure durations.

Aspect	Details
Software	`hesim` package in R
Number of simulations	100 datasets for each modeling assumption
Transitions	Dementia-free → Dementia, Dementia-free → Death , Dementia → Death
Sample size	$n_{obs} = 100,000$
Modeling assumptions	Markov and Semi-Markov
Baseline hazard distribution	Gompertz ($\boldsymbol{\lambda}, \boldsymbol{\gamma}$ assessed from real data)
Hazard function	$h_k(t) = \lambda_k \exp(\gamma_k t) \exp(\boldsymbol{\beta}_k^T \mathbf{Z}_i) \quad k \in \{01, 02, 12\}$
Covariates	1 continuous, 2 binary (prevalence: 15% and 30%)
Covariate effects	See Table 4.1
Left truncation	Individuals enter in dementia-free state
Age at study entry	Uniform distribution between 60 and 96 years

Table 4.2: Summary of simulation process

4.3. Study Design

For each of the $n_{\text{sim}} = 1, \ldots, 100$ datasets generated under both the Markov and Semi-Markov frameworks, we considered different sample sizes n_{obs} to evaluate the effect of study size on models' performance. Specifically, we investigated the following scenarios:

- **Small population-based study:** $n_{\text{obs}} = 500$
- **Medium population-based studies:** $n_{\text{obs}} = 2000$ and $n_{\text{obs}} = 5000$
- **Large population-based study:** $n_{\text{obs}} = 10.000$

We sampled individuals according to the predefined sample sizes, ensuring that the original proportion of diseased individuals was preserved within each dataset.

Each dataset was subject to *right and interval censoring*, assuming that the censoring distribution is independent of the event of interest and non-informative. In our study design, we assumed a total follow-up period of 20 years.

For each patient i, the right censoring time was defined by generating a variable Y_i uniformly distributed in the interval $[0, 20]$, representing the number of years the individual would remain under observation. The censoring time was then calculated as:

$$C_i = \text{age}_{i,0} + Y_i$$

where $\text{age}_{i,0}$ represents the patient's age at baseline. Considering T_i as the transition time into the absorbing state we defined the following quantities:

$$\tilde{T}_i = \min(T_i, C_i)$$

$$\delta_{2i} = \mathbf{1}(T_i \leq C_i)$$

where δ_{2i} indicates whether the death was observed ($\delta_{2i} = 1$) or censored ($\delta_{2i} = 0$).

Observation Schemes

To assess the impact of different observation strategies, we considered four distinct observation schemes, characterized by different visit frequencies and patterns:

- **Scheme A:** fixed annual visits (1-year observation intervals), representing frequent and regular observations, commonly used in small population studies.
- **Scheme B:** fixed triennial visits (3-year observation intervals), representing regular but less frequent follow-ups.
- **Scheme C:** age-dependent visit scheduling, with intervals of 3 years for subjects aged < 78 years and 6 years for those aged ≥ 78 years, reflecting real-world age-based follow-up strategies such as in SNAC-K.
- **Scheme D:** irregular observation times, simulating scenarios where follow-up occurs due to self-reporting or electronic health record linkage, generating intervals at random. Naturally, if this were the actual study design, we would no longer be operating under a non-informative censoring assumption. However, here, we merely aim to capture the essence of such studies while ensuring that the generated censoring mechanism remains consistent with the underlying assumptions.

The observation schemes A, B, and C were generated based on a predefined expected frequency of patient follow-ups. Specifically, for each patient i, the expected number of follow-up visits was determined using the following formula:

$$n_{\text{visits},i} = \left\lceil \frac{20}{\text{W}} \right\rceil \tag{4.3}$$

where W represents the average time expected between two consecutive visits. For example, under Scheme C, the expected number of visits was computed as $n_{\text{visits},i} = \lceil \frac{20}{\frac{6+3}{2}} = 5 \rceil$.

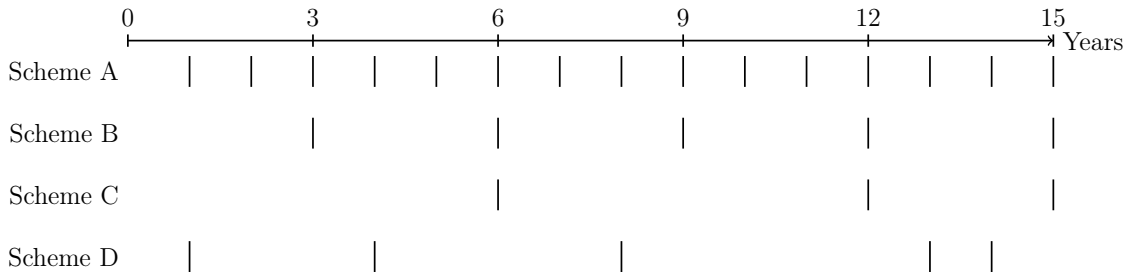


Figure 4.1: Observation schedules under different schemes. Years are calculated since entry into the study. For illustration purposes, the follow-up ends at 15 years and there is no variability in the time between consecutive visits.

To introduce a degree of realism and variability in the observation times, the actual visit times were simulated using a uniform distribution centered around the expected time between two consecutive visits.

$$\begin{aligned} t_i[1] &= \text{age}_{i,0}, \\ t_i[j] &= \text{age}_{i,0} + \text{Uniform}\left(\text{W} - 0.5,\ \text{W} + 0.5\right), \quad j = 2, \ldots, n_{\text{visits},i}. \end{aligned} \tag{4.4}$$

This approach accounts for natural variations in scheduling, acknowledging that patients are unlikely to be observed at perfectly fixed intervals. The variability was assumed to be within a range of six months.

For Scheme D, representing irregular observations, both the number of visits and the visit times were randomly generated. Specifically:

$$n_{\text{visits},i} \sim \text{Uniform}(1, 20) \tag{4.5}$$

$$t_i[j] \sim \text{age}_{i,0} + \text{Uniform}(1, 8) \quad j = 1, \ldots, n_{\text{visits},i} \tag{4.6}$$

Below, we present an overview of the study design strategy [4.3, 4.2]. By considering all possible sample sizes and observation schemes, we obtain a total of **16 distinct scenarios** to analyze across each of the $n_{\text{sim}} = 1, \ldots, 100$ datasets generated under both the Markov and Semi-Markov frameworks.

For clarity, we will refer to all analyses conducted on datasets generated under the Markovian assumption as *Simulation Study Part 1*, while those based on datasets generated under the Semi-Markovian assumption will be referred to as *Simulation Study Part 2*.

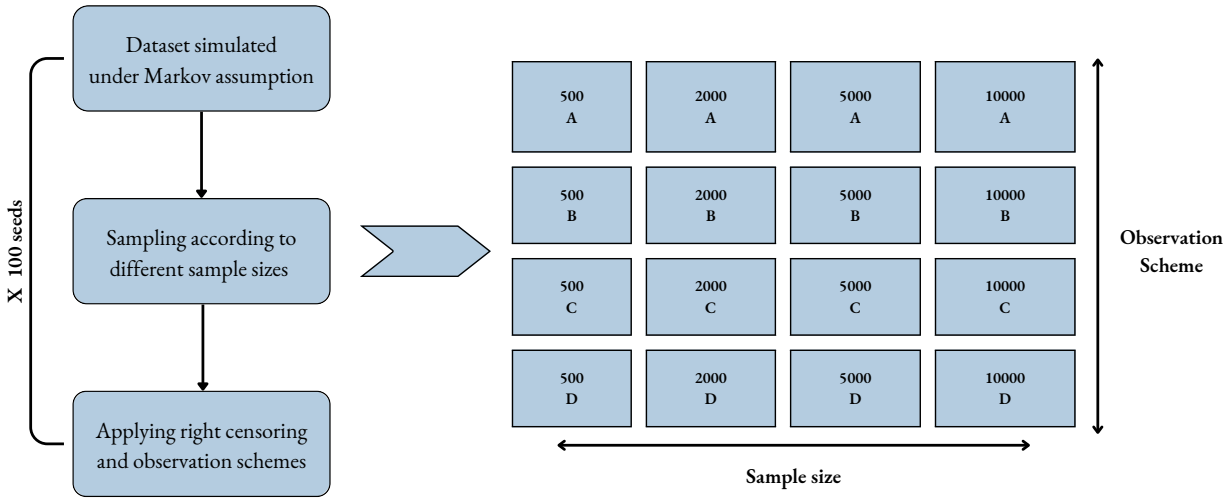


Figure 4.2: Study design for simulation study part 1

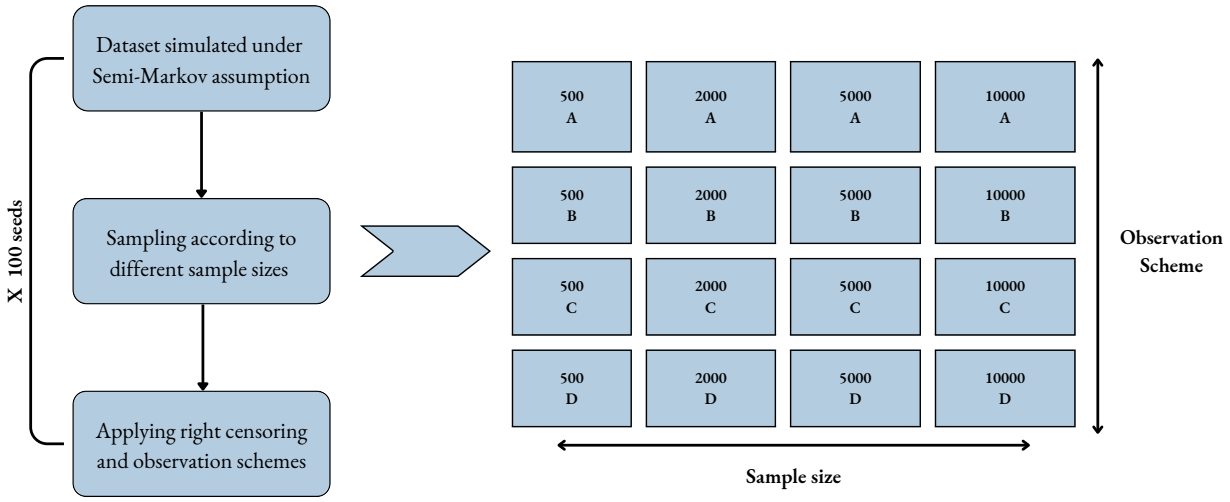


Figure 4.3: Study design for simulation study part 2

Final Data Structure

Finally, the defined observation schemes were applied to each simulated dataset.

In cases where dementia was observed in the original dataset, the indicator variable δ_{1i} was set to 1 if the simulated onset time of dementia occurred before the time of the last observed visit; otherwise, it was set to 0, indicating no observed onset of dementia during the follow-up period.

For individuals where $\delta_{1i} = 1$, meaning dementia onset was detected, the right endpoint R_i was set to the visit time immediately after the dementia onset, and the left endpoint L_i was set to the visit time preceding the onset.

Through this process, we obtained, for each patient, the vector $(L_{0i}, L_i, R_i, \delta_{1i}, \delta_{2i}, \tilde{T}_i)$, as described in 1.4, which contains the patient's baseline age, left and right endpoints of the dementia onset interval, disease indicator, censoring indicator, and time of death or lost to followup.

4.4. Methods

The methods selected for this study aim to address the research questions outlined in the objectives of our simulation study. Our chosen strategy involves progressively increasing the complexity of the applied methodologies to thoroughly investigate the impact of different modeling approaches.

For each simulated dataset generated under the Markov assumption, and for each of the considered 16 scenarios, we will apply the following methods:

a. *Semi-parametric multi-state model assuming exact transition times* [2.2]

b. *Parametric multi-state model assuming exact transition times* [2.3]

c. *Time-homogeneous multi-state model for interval-censored data* [2.4]

d. *Time-homogeneous multi-state model for interval-censored data with piecewise time-varying covariates (age of the individuals)* [2.4]

e. *Parametric time-inhomogeneous multi-state model for interval-censored data* [2.5]

f. *Multiple imputation for panel data strategy* [3.1]

A visual representation of the process that led us to the selection of methodologies *a-f* is provided in figure 4.4.

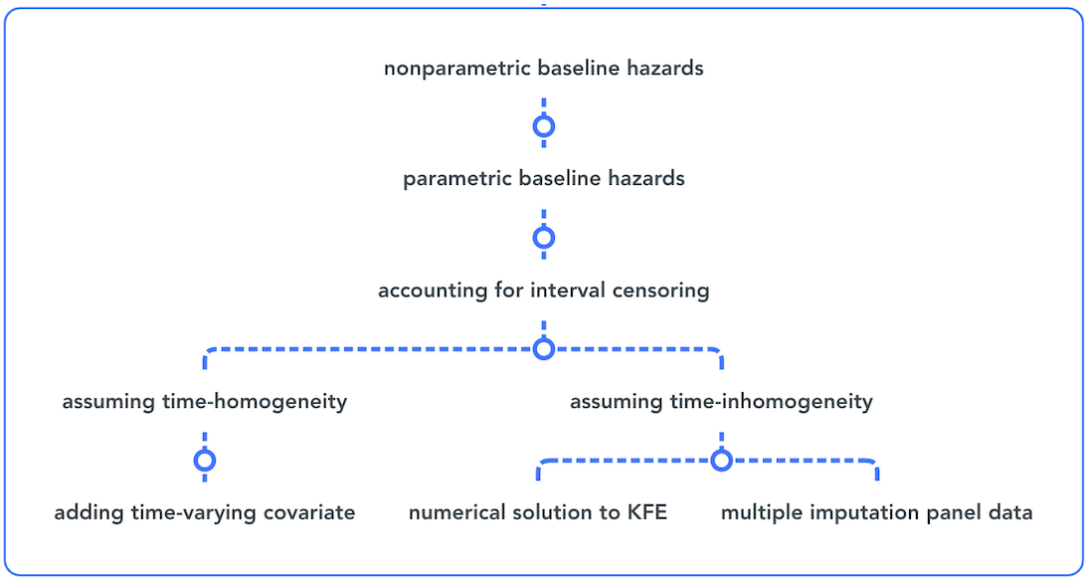


Figure 4.4: Starting from the simplest semi-parametric multi-state model a with the assumption of parametric hazard functions, we defined model b. We then incorporated the interval censoring mechanism, initially within a time-homogeneous framework, which led to models c and d. Subsequently, by assuming a time-inhomogeneous model, we explored two entirely different strategies, resulting in models e and f.

A detailed description of methods a to e, along with their implementation, is provided in Chapter 2, while the multiple imputation strategy for panel data (method f) is elaborated in Chapter 3.

Since the data were simulated from a Gompertz parametric distribution, methods b, e, and f will incorporate the Gompertz distribution for modeling the hazard functions.

For each simulated dataset under the Semi-Markov assumption, where the time spent in the dementia state is explicitly considered, the following methods will be applied for each of the 16 scenarios:

a. *Semi-parametric multi-state model assuming exact transition times* [2.2]

b. *Parametric multi-state model assuming exact transition times* [2.3]

c. *Multiple imputation for panel data strategy* [3.2]

In addition to the above methods, a benchmark model will be fitted both under Markov and Semi-Markov assumption. The *benchmark model* (indexed by 0) consists in a parametric multi-state model assuming a Gompertz hazard distribution, fitted to the original datasets for each sample size, without applying any interval censoring mechanism. Consequently, in this case, the transition times are correctly assumed to be exactly observed.

The benchmark model serves as a reference to compare the performance of alternative strategies that either ignore interval censoring or incorporate it using more complex approaches. It is important to note that, in practice, such a model could never be fitted in real-world studies, as chronic diseases are never observed continuously over time. However, this benchmark allows us to objectively evaluate the performance of our proposed methodologies under ideal conditions.

4.5. Estimands

The quantities of interest that we decided to investigate are described in the following Table 4.3. The estimands will be extracted from each fitted model under all 16 possible scenarios for all datasets $j = 1, ...n_{sim}$.

The baseline hazard functions can provide important insights into the underlying transition process. However, they cannot be estimated under semi-parametric models, and they do not provide meaningful information when the assumed parametric distribution does not align with the true underlying process. So they would be computed only for model b, d, e and f.

The effect of covariates on different transitions is crucial to identify exposure factors influencing the progression of dementia and its risk factors. Moreover, it is also interesting to investigate whether the factors that contribute to a faster death in individuals without dementia are the same as those that accelerate the progression from dementia to death.

Regarding the average total time spent in different states, we have assumed 60 years as the starting age for computation, as it represents a reasonable age both for study enrollment and for the onset of dementia risk. These quantities are of particular interest in medical research, as they provide valuable insights into how long, on average, an individual is expected to remain healthy and how long he is likely to live with dementia once diagnosed.

In our analysis, we calculated these measures assuming that the continuous covariate $\mathbf{Z}[1]$ is centered at its mean value, while the binary covariates $\mathbf{Z}[2]$ and $\mathbf{Z}[3]$ are set at their respective prevalence levels of 15% and 30%. Clearly, these quantities can also be used for patient-specific predictions by setting the covariates to the values corresponding to the patient of interest.

Estimand	Description	Formula
Baseline hazards (when applicable)	For each transition k, we are interested in the scale parameter and the rate of change of the hazard. These parameters provide insights into the transition dynamics between states.	$(\hat{\lambda}_k^{(j)}, \hat{\gamma}_k^{(j)})$
Vector of covariates effect	The impact of covariates associated with each transition k is of primary interest in determining exposure factors. The estimated coefficients allow for the derivation of hazard ratios (HRs), facilitating interpretation.	$\hat{\boldsymbol{\beta}}_k^{(j)}$
Total length of stay in dementia-free state	This quantity is calculated as the integral of the transition probability of remaining in the dementia-free state. The integration limits assume study entry at age $t_0 = 60$ and an upper bound t_{end}, ensuring that all individuals have experienced either death or study dropout.	$T_{00}^{(j)} = \int_{t_0}^{t_{\text{end}}} p_{00}^{(j)}(t)\, dt$
Total length of stay in dementia state	This quantity is calculated as the integral of the transition probability of going to state 1 at $t_1 \geq t_0$ from state 0, assuming the individual enters in the study at $t_0 = 60$. The integration limits assume lower bound t_1 and upper bound t_{end}, ensuring that all individuals have experienced either death or study dropout.	$T_{11}^{(j)} = \int_{t_1}^{t_{\text{end}}} p_{01}^{(j)}(t)\, dt$

Table 4.3: Definitions and formulas of key estimands in the simulation study.

4.6. Performance Measures

The performance measures we have chosen to evaluate the quality of the estimators are summarized in Table 4.4. To ensure robust results, these measures have been computed for each fitted model across all possible scenarios by averaging over the 100 datasets generated using different random seeds. By averaging across multiple datasets, we aim to reduce variability and obtain reliable estimates of the estimator's properties.

Metric	Description	Formula
Absolute Bias	Measures the systematic absolute difference between the expected value of an estimator and the true parameter value. It ignores whether the estimator tends to overestimate or underestimate the parameter of interest.	$\frac{1}{n_{\text{sim}}} \sum_{j=1}^{n_{\text{sim}}} \lvert\hat{\theta}^{(j)} - \theta\rvert$
Relative Bias	Bias expressed as a proportion of the true parameter value, providing a scale-invariant measure. It's applicable only when the true parameter is not null.	$\frac{1}{n_{\text{sim}}} \sum_{j=1}^{n_{\text{sim}}} \frac{\hat{\theta}^{(j)} - \theta}{\theta}$
Coverage	Proportion of times the estimated confidence interval contains the true parameter value, reflecting the accuracy of uncertainty quantification.	$\frac{1}{n_{\text{sim}}} \sum_{j=1}^{n_{\text{sim}}} \mathbf{1}\{\theta \in [L_j, U_j]\}$
Type I Error	The probability of incorrectly rejecting the null hypothesis when it is true (e.g., assessing a covariate effect as significant when it's not).	$\frac{1}{n_{\text{sim}}} \sum_{j=1}^{n_{\text{sim}}} \mathbf{1}\{\text{Reject } H_0 \mid H_0 \text{ is true}\}$
Power	The probability of correctly rejecting the null hypothesis when the alternative hypothesis is true (e.g., assessing a covariate effect as significant when it's true).	$\frac{1}{n_{\text{sim}}} \sum_{j=1}^{n_{\text{sim}}} \mathbf{1}\{\text{Reject } H_0 \mid H_1 \text{ is true}\}$

Table 4.4: Definitions and formulas of key performance measures in the simulation study.

It is important to note that leveraging these performance measures is feasible in our case because we are conducting a simulation study, where the true parameter values are known in advance. This allows us to objectively assess the performance of the estimators by directly comparing their estimates to the ground truth, an advantage that would not be available in real-world applications where the true values remain unknown.

In addition to the previously described performance measures, we have introduced further metrics to assess the quality and reliability of the considered strategies. Specifically, we have analyzed:

- **Average computational time**: the average computational time of each strategy has been evaluated across different sample sizes and observation schemes. This measure is crucial since a methodology cannot be considered practical if it is associated with an unfeasible computational time. Researchers often have limited computational resources, and we aim to compare only strategies that are truly applicable in practice. Furthermore, analyzing computational time can provide valuable insights into whether it increases significantly as the sample size increases or whether for some methods it is particularly sensitive to the choice of observation schemes.
- **Proportion of convergent models**: the proportion of models achieving convergence has been calculated across different sample sizes and observation schemes. We have defined *lack of convergence* as the situation where a given strategy, applied to a specific dataset, fails to meet the convergence criteria imposed by the optimization algorithm used—criteria that are specific to the software package employed for the maximization process. Additionally, we have defined *non-optimal convergence* as the condition where a strategy meets the required convergence criteria but fails to compute the covariance matrix of the estimated parameters. This issue arises when the Hessian (or its equivalent in the specific algorithm) is not positive definite and therefore not invertible. Consequently, in the following analysis, we will consider as *convergent models* only those that do not fall into either of the two aforementioned cases, ensuring an accurate evaluation of the stability of the adopted estimation strategies.

4.7. Results

In this section, we present the results of the simulation study, supported by both graphical and tabular representations. The performance measures were compared for each model across all possible sample sizes and observation schemes. However, for clarity and conciseness, we report only the most relevant plots that highlight the key findings and differences across the various scenarios. The remaining material will be provided in the appendix. Specifically, supplementary results of *Simulation Study Part 1* will be presented in appendix A, while results from *Simulation Study Part 2* will be included in appendix B.

Throughout this section, we will refer to the transitions as follows:

- **Transition 1:** from dementia-free to dementia,
- **Transition 2:** from dementia-free to death,
- **Transition 3:** from dementia to death.

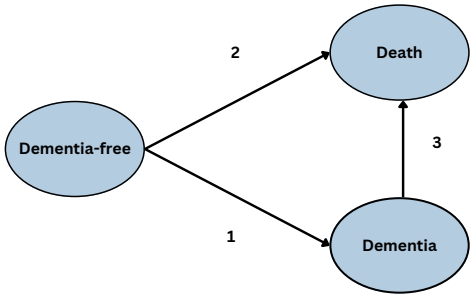


Figure 4.5: Feasible transitions in the progressive illness-death model

Furthermore, we remind the reader that the observation schemes labeled as A, B, C, and D correspond to the following:

- **A:** fixed annual visits,
- **B:** fixed triennial visits,
- **C:** Age-dependent visit scheduling (every 3-6 years),
- **D:** irregular observation times.

All plots presented in this section include the different strategies described earlier, along with the benchmark model introduced previously. The benchmark serves as a reference to compare the best possible performance achievable in each scenario.

4.7.1. Results from simulation study part 1

Rate of convergence and computational time

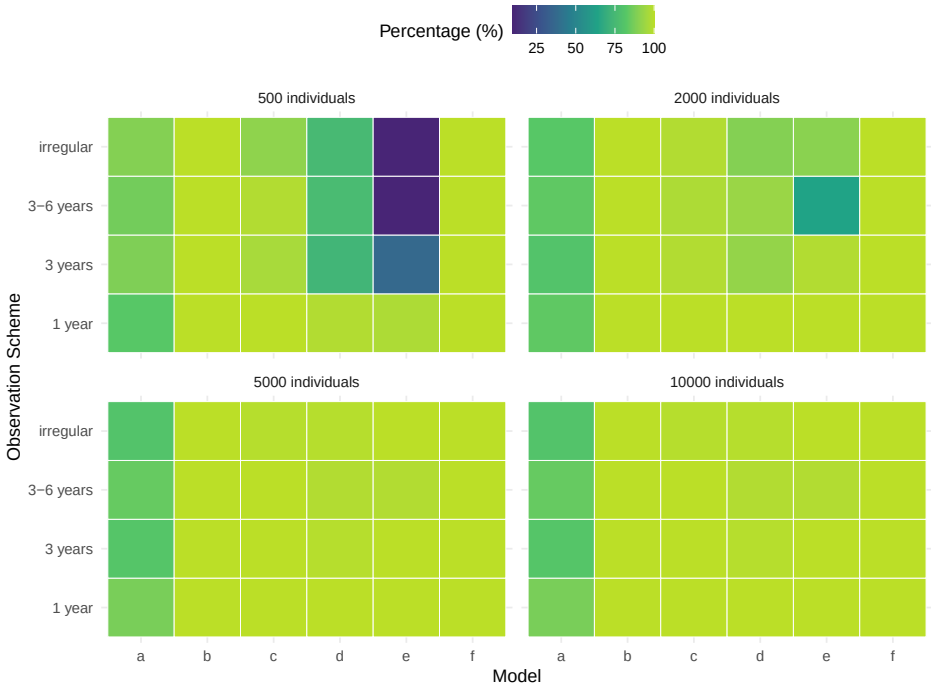
The first step in our analysis involved assessing the **rate of convergence** associated with different models across various observation schemes. Consistent with the criteria defined in Section 4.6, a model was considered to have converged only if the optimization algorithm satisfied the required convergence criteria and correctly computed the covariance matrix of the estimated parameters. The results of this assessment are illustrated in figure 4.6a.

We observe that the models accounting for interval censoring, specifically the time-homogeneous model with extra covariate *age* (d) and the time-inhomogeneous model (e), implemented using the `msm` and `nhm` packages, are the most affected by convergence issues. Several factors contribute to these issues, which we discuss below.

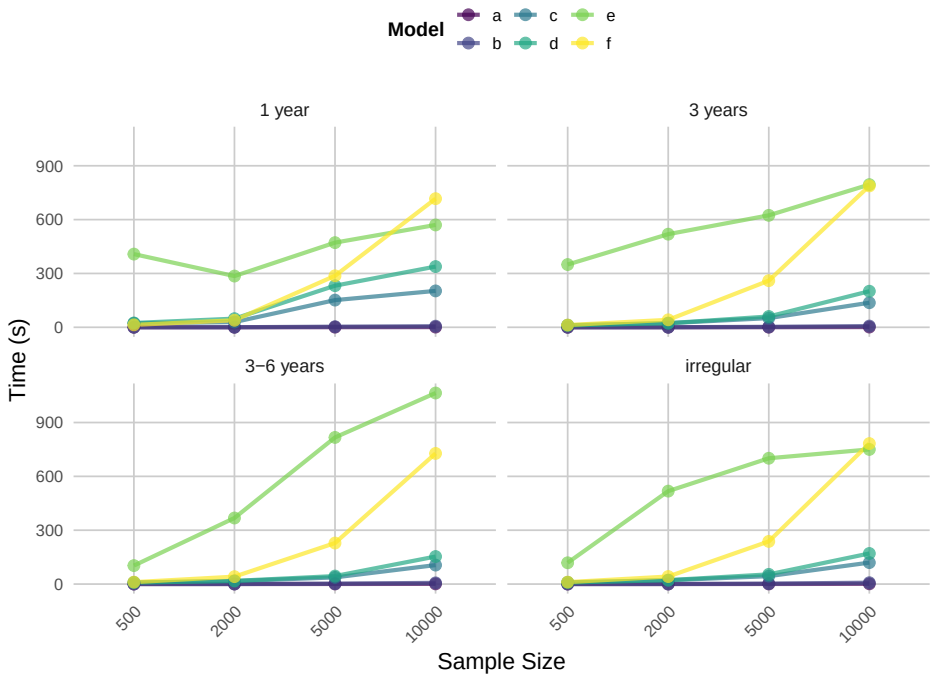
One of the primary reasons for convergence difficulties lies in the complexity of the likelihood function induced by the presence of interval censoring. This additional uncertainty makes the likelihood function highly non-linear and prone to multiple local optima, which complicates the task of numerical optimization algorithms in locating the global maximum. Poor initialization strategies can further exacerbate the issue resulting in slow convergence or premature termination before reaching a valid solution.

Another critical factor affecting convergence is overparameterization. This issue arises due to the inclusion of covariate effects, as seen in model d compared to model c, and the complexity of baseline hazard functions, which are more intricate in model e than in c. When the sample size is small, the available data may be insufficient to accurately estimate the increased number of parameters, leading to convergence failures and unreliable estimates.

It is also important to note that **wider observation intervals** between consecutive visits, which introduce greater uncertainty, tend to exacerbate **convergence issues**. However, our results indicate that these convergence issues in moderately complex multi-state models can be largely **resolved by increasing the sample size**, which provides more data to inform the estimation process.



(a) Rate of convergence



(b) Average computational time

Figure 4.6: **Rate of models convergence and average computational time across different observation schemes and sample sizes.**
0) Benchmark model. a) Semi-parametric multi-state model assuming exact transition times. b) Parametric multi-state model assuming exact transition times (Gompertz baseline). c) Time-homogeneous multi-state model for interval-censored data d) Time-homogeneous multi-state model or interval-censored data with 'age' covariate. e) Parametric time-inhomogeneous multi-state model for interval-censored data (Gompertz baseline). f) Multiple imputation for panel data strategy (Gompertz baseline)

On the other hand, multi-state models fitted using the Cox regression approach consistently met the convergence criteria across all scenarios. However, we observed a small proportion of models that converged with non-optimal solutions, primarily due to ill-conditioned Hessian matrices.

From Fig. 4.6b we note that the **average computational time** is generally an increasing function of the sample size, as expected, but is not excessively affected by the choice of the observation scheme.

It is worth noting that the computational time associated with models that ignore interval censoring is negligible under any scenario. Similarly, the computational time for time-homogeneous models accounting for interval censoring remains almost constant across different observation schemes, staying under 5 minutes even for the largest sample sizes.

A particular case is represented by the time-inhomogeneous model fitted by numerically solving the KFE, which generally requires the longest execution time. In the worst-case scenario, it takes around 18 minutes per dataset, despite the parallelization and the coarsening of continuous covariates employed to make the computation feasible. In some cases, the latter requires excessively long execution times even when the sample size is small, which can be explained by the convergence issues discussed earlier. These issues cause slow iterations and the search for the optimum in incorrect parametric spaces, leading to inefficient computation. As the number of individuals in the study increases, the computational time naturally grows, but there is also a reduction associated with more facilitated convergence.

Finally, the strategy implemented with multiple imputation for panel data must account for both the time to perform the imputation, which involves generating m datasets with complete disease history, and the time to fit a parametric model independently to each of the m imputed datasets. In our analysis, $m = 30$ and it is fixed to this value; however, it must be pointed out that as the sample size increases, the approach becomes unfeasible. For $m = 1$, we expect the computational time of strategy f to be similar to that of strategy b, since we used the same parametric implementation. Therefore, the strategy is promising if a parallelization technique is implemented.

Estimated effect of covariates for different modeling strategies

The effect of covariates on the different modeled transitions, using various estimation strategies, was first assessed through the **absolute bias** of the estimated covariate coefficients.

We first compare the effect of different observation schemes on a population study with 500 individuals in Fig. 4.7a. Subsequently, we explore how the absolute bias changes with increasing sample size under a triennial observation scheme in Fig.4.7b. To enhance readability, we have used distinct color palettes for these two comparisons and maintained this consistency across all related results.

It is important to note that the choice of different y-axis scales is intentional. This decision aims to emphasize the varying trends of different strategies, which would be less noticeable if a fixed scale was applied across different coefficients.

Firstly, we observe that for each transition the coefficient associated to the continuous covariate ($\hat{\beta}_1$) exhibits a higher bias compared to the ones associated to the binary covariates ($\hat{\beta}_2, \hat{\beta}_3$)

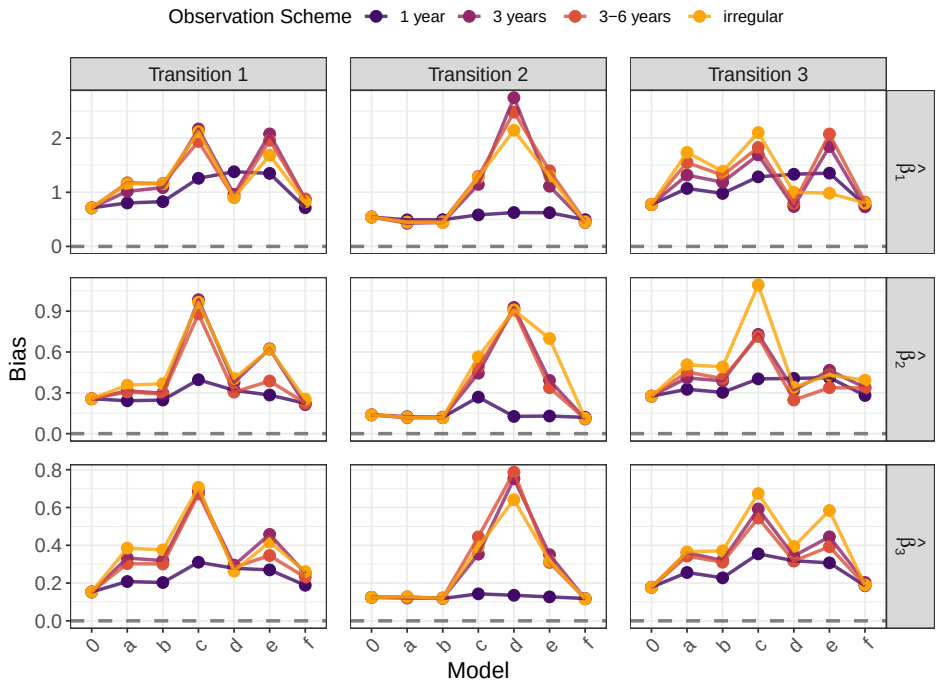
Under the fixed annual visit scheme (A), the bias associated with significantly different strategies appears to be similar. This suggests that whether or not interval censoring is modeled does not have a substantial impact on estimation performance under this observation scheme.

Under observation schemes B, C, and D, the bias trends across different strategies are similar. Moreover:

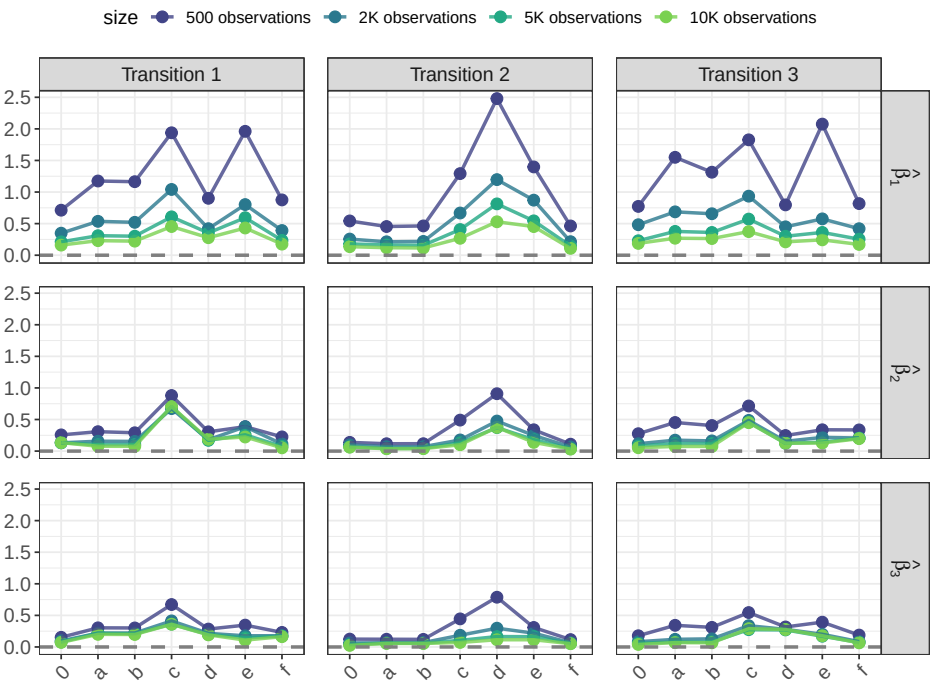
- For **Transition 2**, methodologies that ignore interval censoring tend to perform better. This is because the entry into the study and the event of death are known precisely and not subject to uncertainty.
- For **Transitions 1 and 3**, it is beneficial to use methods that account for interval censoring. Specifically, models d and f exhibit lower bias compared to the semi-parametric and parametric models that do not consider interval censoring.

As the **sample size increases**, the following trends are observed:

- The overall bias decreases across all scenarios.
- The time-homogeneous model (c) confirms its inadequacy under any scenario due to its constrained exponential baseline assumption.
- The bias associated with the time-inhomogeneous model e decreases and, in some cases, becomes preferable to model d, in which time-inhomogeneity is modeled in a piecewise fashion.



(a) Absolute bias with 500 observations



(b) Absolute bias with varying sample sizes under triennial observation scheme

Figure 4.7: **Comparison of absolute bias of covariate coefficient estimates for different models across different observation schemes and sample sizes**

0) Benchmark model. a) Semi-parametric multi-state model assuming exact transition times. b) Parametric multi-state model assuming exact transition times (Gompertz baseline). c) Time-homogeneous multi-state model for interval-censored data d) Time-homogeneous multi-state model or interval-censored data with 'age' covariate. e) Parametric time-inhomogeneous multi-state model for interval-censored data (Gompertz baseline). f) Multiple imputation for panel data strategy (Gompertz baseline)

The results highlight the importance of considering interval censoring for transitions where the exact timing of events is unknown. Additionally, the choice of an appropriate baseline hazard function plays a crucial role in achieving accurate estimates.

Further details and additional results can be found in the appendix.

We now investigate another quantity associated with the effect of covariates: the proportion of times the confidence interval of the estimated coefficients, $\hat{\boldsymbol{\beta}}$, contains the true value of the coefficients. If the model appropriately represents the data and the 95% confidence intervals are correctly constructed, we expect the **coverage** to be approximately 95%.

To highlight the obtained results, we first compare the effect of different observation schemes on a population study with 5000 individuals in Fig. 4.8. Subsequently, we explore how the coverage changes with increasing sample size under an annual observation scheme in Fig.4.9.

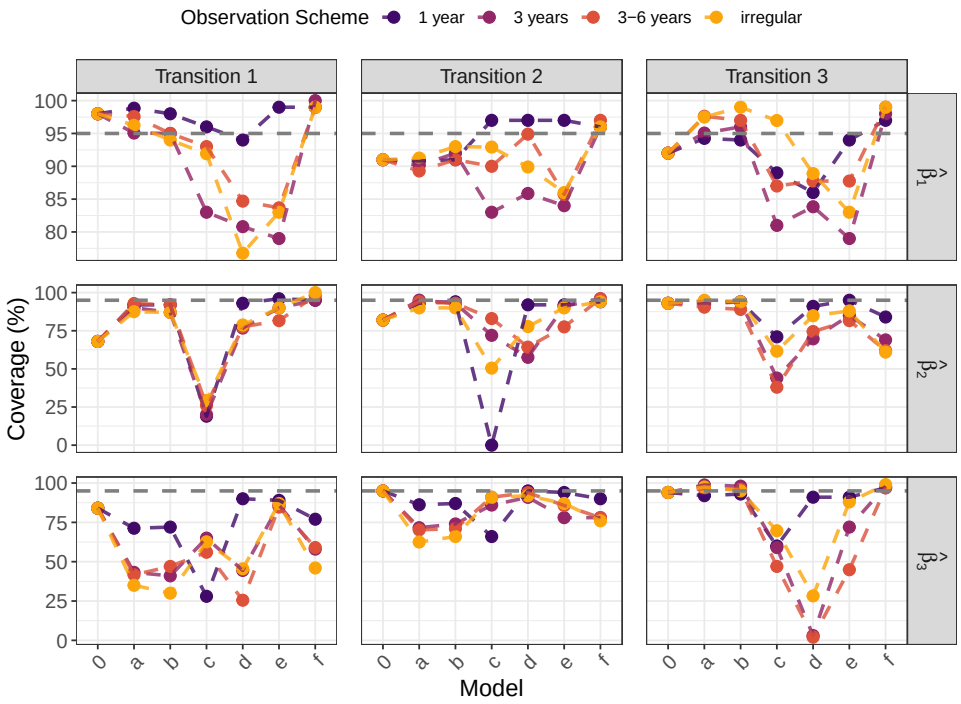


Figure 4.8: **Coverage of covariate coefficient estimates for different models across different observation schemes over 5000 individuals**
0) Benchmark model. a) Semi-parametric multi-state model assuming exact transition times. b) Parametric multi-state model assuming exact transition times (Gompertz baseline). c) Time-homogeneous multi-state model for interval-censored data d) Time-homogeneous multi-state model or interval-censored data with 'age' covariate. e) Parametric time-inhomogeneous multi-state model for interval-censored data (Gompertz baseline). f) Multiple imputation for panel data strategy (Gompertz baseline)

First, we observe that models which ignore interval censoring (*e.g.*, models *a* and *b*) show

a coverage value that remains nearly constant across different scenarios, with a slight decrease in performance as the time intervals between consecutive visits increase.

The time-homogeneous model (c) proves to be completely inadequate in describing the data across all cases.

On the other hand, the model with a time-varying baseline hazard (e), achieves coverage close to the expected value even in the presence of longer observation intervals. However, it encounters issues in estimating the confidence intervals for the coefficient $\hat{\beta}_1$ associated with continuous covariates, due to challenges in handling such covariates within `nhm` package.

In general, confidence intervals were computed based on asymptotic standard errors derived from the Hessian matrix or transformations of these using the delta method. This has two important implications in our analysis:

- If the model presents a significant **bias**, it will result in systematically shifted confidence intervals, leading to reduced coverage.
- If the Hessian matrix is **ill-conditioned**, the asymptotic standard errors may be inaccurate. Since models implemented using the `msm` and `nhm` packages are more susceptible to these issues due to computational challenges, the obtained results should not be surprising.

Thus, the inaccuracy in confidence intervals is not primarily due to the modeling of interval censoring itself but rather the computational difficulties associated with it. This conclusion is supported by the results obtained from model f, which uses the same confidence interval computation methods as model b but achieves better performance by mitigating the uncertainty associated with intermittent observation schemes through a multiple imputation technique for panel data.

We now analyze the effect of **increasing sample size**:

- A deterioration in coverage is observed, particularly for transition 1 (dementia-free state to dementia), which is the least represented transition in our datasets, as only a small percentage of patients develop dementia.
- A severe deterioration is observed in models with higher bias (e.g., model c). As the sample size increases, the bias does not decrease, whereas the variance does. This leads to narrower confidence intervals that are centered around an incorrect value, thereby reducing the actual coverage.

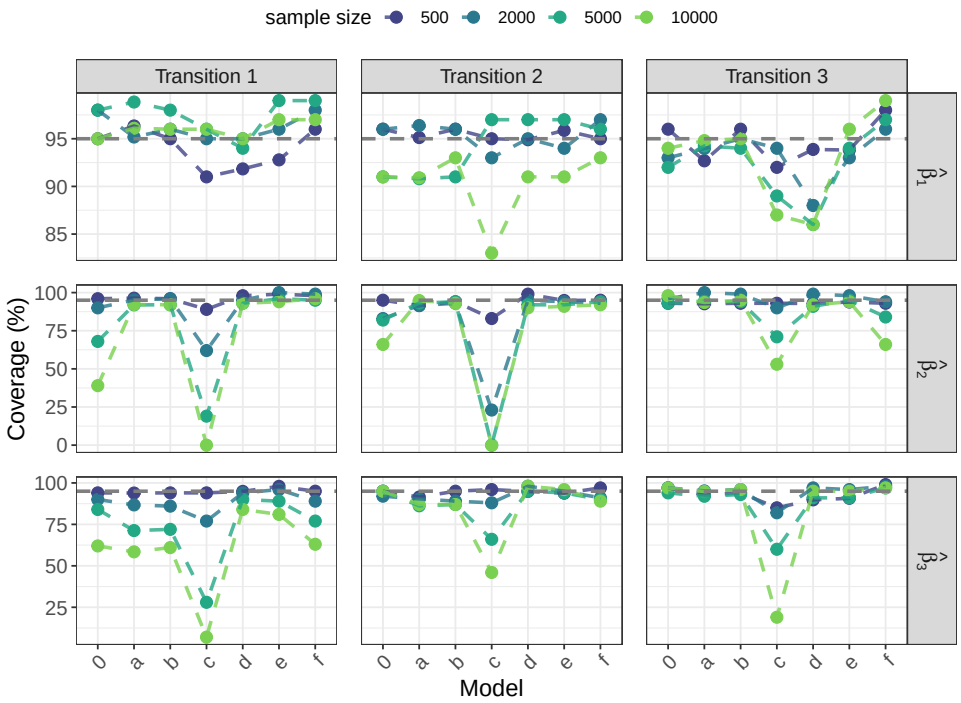


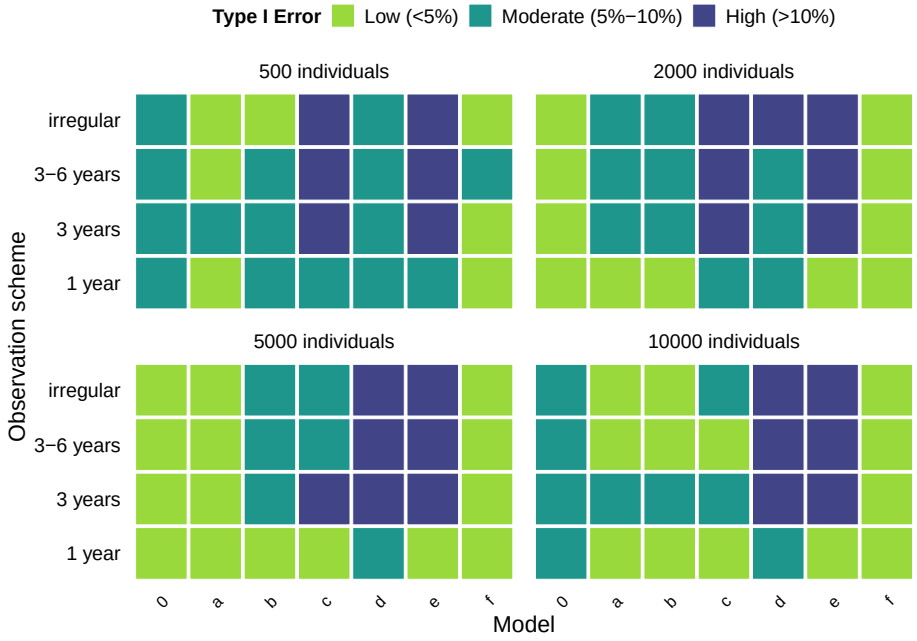
Figure 4.9: **Coverage of covariate coefficient estimates for different models across different sample sizes for annual observation scheme**
0) Benchmark model. a) Semi-parametric multi-state model assuming exact transition times. b) Parametric multi-state model assuming exact transition times (Gompertz baseline). c) Time-homogeneous multi-state model for interval-censored data d) Time-homogeneous multi-state model or interval-censored data with 'age' covariate. e) Parametric time-inhomogeneous multi-state model for interval-censored data (Gompertz baseline). f) Multiple imputation for panel data strategy (Gompertz baseline)

Now, we are going to analyze the **Type I Error** rates associated with different models across various transitions. Specifically, the covariates considered non-influential in the data simulation process were:

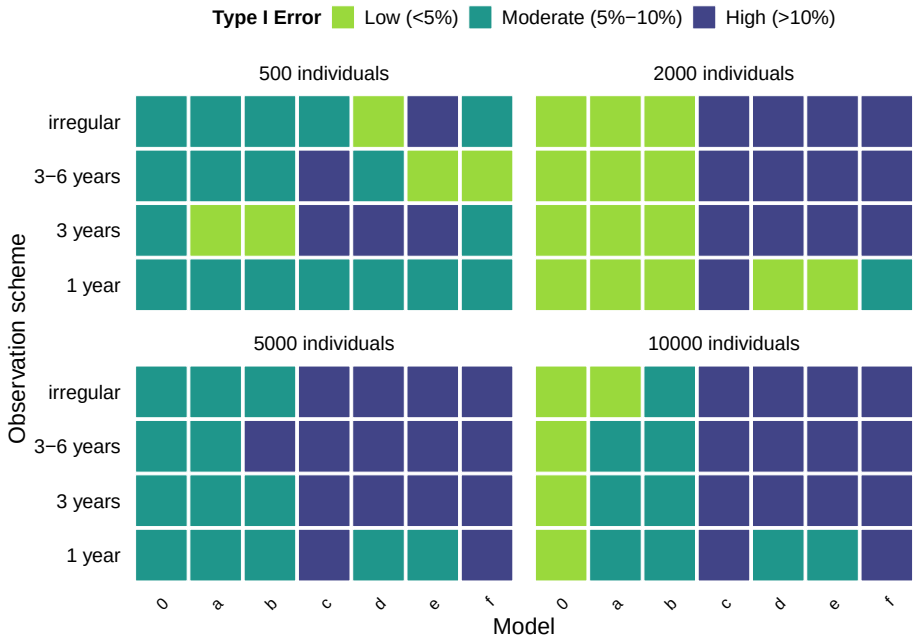
- Covariate 1 (continuous) in the transition to dementia.
- Covariate 2 (binary with lower prevalence) in the transition from dementia to death.

For each analytical strategy, we examine the proportion of instances in which these covariates were deemed significant under different observational scenarios.

It is generally expected that the Type I Error should decrease as the sample size increases if the model is well-specified. Since the datasets were randomly drawn from the same initial population while preserving the proportion of diseased individuals, each dataset may capture slightly different structures of the population. Thus, the benchmark model plays a crucial role in our analysis, providing a reference to determine whether any unexpected behavior arises due to the model itself or the inherent randomness in the sampling process.



(a) Categorized Type I Error associated to $\hat{\beta}_1$ in transition to Dementia



(b) Categorized Type I Error associated to $\hat{\beta}_2$ in transition from Dementia to Death

Figure 4.10: **Categorized Type I Error across different observation schemes and sample sizes.**
0) Benchmark model. a) Semi-parametric multi-state model assuming exact transition times. b) Parametric multi-state model assuming exact transition times (Gompertz baseline). c) Time-homogeneous multi-state model for interval-censored data d) Time-homogeneous multi-state model or interval-censored data with 'age' covariate. e) Parametric time-inhomogeneous multi-state model for interval-censored data (Gompertz baseline). f) Multiple imputation for panel data strategy (Gompertz baseline)

The results indicate that the Type I Error rate is more pronounced in scenarios characterized by long intervals between consecutive visits and irregular observation schemes.

- For what concerns **transition 1** [4.10a] models $c, d,$ and e exhibit the highest inaccuracies, with Type I Error rates exceeding 5% in observation schemes B, C, and D. Conversely, the multi-state models that ignore interval censoring a and b demonstrate relatively good performance, with rates staying within the expected range. Notably, model f consistently achieves Type I Error rates below 5% across all observation schemes and sample sizes, highlighting its robustness in this transition.
- Similarly, in **transition 3** [4.10b] models that ignore interval censoring (a, b) tend to perform better, maintaining Type I Error rates below 10%. While models c, d, and e exhibit a high Type I Error even in this transition, we also observe elevated Type I Error rates for model f. This discrepancy can be attributed to the distribution of $\hat{\beta}_2$ in model f, which appears to be systematically biased compared to the true null value. This suggests that the imputation of the disease state may have altered the distribution of influential covariates for this transition, leading to a higher Type I Error rate. This hypothesis is further supported by the similar behavior observed for $\hat{\beta}_2$ in the transition to dementia (see figure A.9 and A.10 in Appendix A for further analysis).

The **power** of a model in determining the significance of a covariate is defined as the proportion of times it correctly identifies the covariate as significant when it truly is. As expected, power tends to **increase with increasing sample size**, a trend that is generally observed in our analysis.

An interesting exception arises in the case of the continuous covariate in the transition from dementia to death [A.8]. Here, we observe that none of the models are able to appropriately detect the significance of this covariate. This suggests that even with a sample size of $n_{obs} = 10.000$, the data might still be insufficient to achieve the required statistical power for detecting the covariate's effect.

Focusing specifically on **transition 1** [4.11], we observe that the homogeneous time model (c) exhibits the lowest power, often falling below 50% across most observation schemes. Models that ignore interval censoring tend to have lower power in observation schemes C and D, whereas models d and e demonstrate superior performance in the majority of cases.

Model f emerges as the best-performing strategy, achieving power levels greater than 80% even with moderate sample sizes for $\hat{\beta}_2$. Moreover, it is the only model that matches the

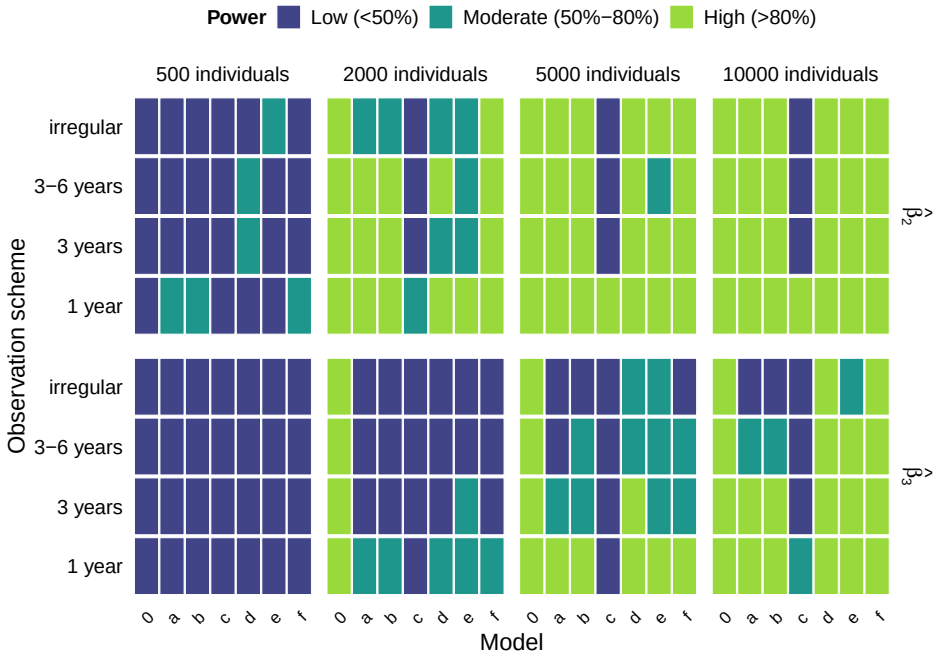


Figure 4.11: **Categorized Power of models across different observation schemes and sample sizes for transition to Dementia**
0) Benchmark model. a) Semi-parametric multi-state model assuming exact transition times. b) Parametric multi-state model assuming exact transition times (Gompertz baseline). c) Time-homogeneous multi-state model for interval-censored data d) Time-homogeneous multi-state model or interval-censored data with 'age' covariate. e) Parametric time-inhomogeneous multi-state model for interval-censored data (Gompertz baseline). f) Multiple imputation for panel data strategy (Gompertz baseline)

benchmark model's performance in detecting the significance of $\hat{\beta}_3$ for larger sample sizes.

Further analysis for transition 2 and 3 are presented in figure A.7 and A.8.

Estimated baseline parameters for different modeling strategies

In this section, we leverage the **relative bias** as a measure of performance to compare the estimates of the baseline hazard parameters for each transition [4.1]. This approach is chosen because in most implementations of the methodologies used, the scale parameter λ is expressed on a logarithmic scale. Since the scales of λ and γ are therefore very different, we make them comparable by using the relative bias.

Furthermore, we present results only for the parametric methods assuming a Gompertz distribution. Specifically, for models 0, b, e and f the baseline hazard parameters were easily extracted from the model. For the time-homogeneous multi-state model including a time-varying covariate (d), the parameters were estimated as follows:

1. **Computing the hazard functions:**
   The hazard function for each transition is calculated using the fitted model. While the model is assumed to be time-homogeneous, incorporating a time-varying covariate (age) requires us to compute the hazard at each discretized time point. This is done by adjusting the hazard values at each time step to reflect the covariate's effect, ensuring that the hazard function updates dynamically over time.

2. **Modeling the hazard function with a Gompertz distribution:**
   Assuming that the hazard values follow a Gompertz distribution, we model the relationship between the hazard function and *age*, since they are expressed on the same time scale imposed by the process. The hazard function at time t, where parameters are in their natural scale, is expressed as:

$$h(t) = \lambda \cdot \exp(\gamma \cdot age(t)) \tag{4.7}$$

   By taking the natural logarithm of both sides, we obtain:

$$\log h(t) = \log \lambda + \gamma \cdot age(t) \tag{4.8}$$

   This transformation converts the problem into a linear regression framework, where the logarithm of the hazard function is linearly dependent on *age* at t.

3. **Extracting the parameters:**
   For each transition, the logarithm of the computed hazard values is regressed against the corresponding *age* values. This regression yields two essential parameters:

   - λ (in logarithmic scale), which corresponds to the intercept of the regression model.
   - γ, which is given by the slope of the regression.

From the following tables, which report the relative bias values associated with λ and γ for each of the mentioned models and observational schemes on a population of 500 individuals, we observe that the **parametric multi-state model that does not account for interval censoring (*b*) generally performs best**. The two strategies incorporating time-varying hazards, *e* and *f*, follow. Specifically, the first approach (*e*) achieves better predictions for the parameters in the transition to dementia [4.5], while the second approach (*f*) provides more accurate estimates in the transition from dementia to death [4.7].

Moreover, we notice that for all analyzed methods, **the relative bias increases in**

observational schemes B, C, and D for transitions 1 and 3. This trend arises because, as previously noted, the uncertainty associated with transition times is higher when the intervals between observation visits are wider. While this is mitigate in transition 2 since exact transition times are known.

We then evaluated the percentage decrease in the relative bias of the parameters observed when the population size increases to 5000 (columns 4 and 5 of the tables). **As the sample size grows**, the performance of **model** e **improves significantly**, making it a reasonable choice when the objective is to accurately estimate the baseline parameters while incorporating the modeling of interval censoring.

Model	$\frac{(\hat{\lambda}-\lambda)}{\lambda}$	$\frac{(\hat{\gamma}-\gamma)}{\gamma}$	**Relative Bias Variation (%)**		**Scheme**
0	-0.0611	0.0672	-35.171	-63.110	benchmark
b	**-0.089**	0.115	-0.741	-1.767	A
d	-0.109	0.121	-19.639	-46.706	
e	-0.096	**0.110**	**-37.729**	**-50.068**	
f	-0.136	0.157	-33.854	-37.552	
b	**-0.117**	0.170	-8.638	0.199	B
d	-0.250	**0.163**	-1.101	-10.631	
e	-0.157	0.191	**-71.900**	**-61.899**	
f	-0.173	0.229	-11.330	-11.945	
b	**-0.098**	**0.133**	**-39.609**	-13.613	C
d	-0.328	0.249	2.221	1.725	
e	-0.221	0.224	-28.438	**-55.390**	
f	-0.178	0.238	-9.711	-9.267	
b	**-0.117**	**0.168**	-34.010	-11.307	D
d	-0.321	0.245	1.607	0.898	
e	-0.167	0.172	**-68.515**	**-55.534**	
f	-0.210	0.290	-3.329	-3.745	

Table 4.5: Relative Bias of baseline parameters in transition 1, for a population of 500 individuals, with the percentage decrease in relative bias observed when the population size increases to 5000.

Model	$\frac{(\hat{\lambda}-\lambda)}{\lambda}$	$\frac{(\hat{\gamma}-\gamma)}{\gamma}$	Relative Bias Variation (%)		Scheme
0	-0.059	0.068	-69.667	-73.469	benchmark
b	-0.060	**0.066**	-53.993	-64.156	A
d	-0.073	0.071	-4.200	-35.738	
e	**-0.057**	**0.066**	**-70.785**	**-64.575**	
f	-0.066	0.071	-53.120	-62.451	
b	**-0.061**	**0.067**	-48.413	**-55.882**	B
d	-0.333	0.368	**-51.506**	-50.372	
e	-0.118	0.154	-25.651	-28.422	
f	-0.070	0.078	-37.505	-40.984	
b	**-0.058**	**0.065**	-52.414	-59.951	C
d	-0.393	0.439	-57.931	-60.253	
e	-0.242	0.278	**-73.064**	**-70.852**	
f	-0.065	0.074	-39.271	-43.475	
b	**-0.057**	**0.064**	-48.449	**-55.704**	D
d	-0.384	0.448	**-49.394**	-48.333	
e	-0.109	0.137	6.366	-1.194	
f	-0.068	0.077	-34.772	-35.907	

Table 4.6: Relative Bias of baseline parameters in transition 2, for a population of 500 individuals, with the percentage decrease in relative bias observed when the population size increases to 5000.

Model	$\frac{(\hat{\lambda}-\lambda)}{\lambda}$	$\frac{(\hat{\gamma}-\gamma)}{\gamma}$	Relative Bias Variation (%)		Scheme
0	-0.169	0.191	-72.330	-72.193	benchmark
b	-0.180	0.204	**-68.274**	**-67.739**	A
d	-0.315	0.318	-24.136	-30.991	
e	-0.232	0.254	-65.088	-66.252	
f	**-0.143**	**0.164**	-64.615	-64.876	
b	-0.233	0.270	**-69.933**	**-70.195**	B
d	-0.933	0.971	-6.038	-7.320	
e	-0.424	0.477	-21.742	-36.935	
f	**-0.174**	**0.181**	-49.834	-62.574	
b	-0.261	0.308	**-70.567**	**-71.941**	C
d	-0.757	0.749	-3.398	-3.947	
e	-0.407	0.461	-3.363	-25.112	
f	**-0.225**	**0.231**	-29.371	-41.515	
b	-0.290	0.339	**-74.058**	**-74.678**	D
d	-0.880	0.914	-3.710	-3.826	
e	-0.322	0.364	-23.064	-35.962	
f	**-0.272**	**0.267**	-35.829	-47.902	

Table 4.7: Relative Bias of baseline parameters in transition 3, for a population of 500 individuals, with the percentage decrease in relative bias observed when the population size increases to 5000.

Average life expectancy

We now compare the average time spent without developing dementia (T_{00}) since the entry in the study and the average time spent in the dementia state (T_{11}) as predicted by different models under various scenarios.

To obtain these predictions, we first generated trajectories for 100.000 patients, assuming that they followed different underlying processes of interest. Subsequently, we computed the time spent in state i as the difference between the time of transition into state $j \neq i$ and the time of entry into state i. We then averaged these durations across all patients. This averaging process is justified by the law of large numbers, ensuring that for a sufficiently large n, the sample mean provides an accurate estimate of the expected value.

Model	T_{00}	T_{11}	Scheme
0	**20.6**	**4.56**	**benchmark**
a	19.3	1.03	A
b	20.9	**4.72**	
c	7.35	6.72	
d	18.6	3.92	
e	19.5	4.38	
f	**20.6**	5.19	
a	19.6	0.743	B
b	21.3	**4.86**	
c	6.49	13.8	
d	15.9	15.2	
e	18.0	2.57	
f	**20.6**	6.39	
a	19.9	0.590	C
b	21.4	**4.65**	
c	6.29	13.3	
d	14.7	3.26	
e	17.6	2.46	
f	**20.6**	6.76	
a	19.9	0.534	D
b	21.5	**4.73**	
c	6.14	14.5	
d	14.5	11.5	
e	17.5	3.08	
f	**20.5**	7.46	

Table 4.8: Average total length of time spent in Dementia-free state and in Dementia State computed with different models across different observational schemes for a population of 2000 individuals.

Table 4.8 reports the estimated average total time spent in the dementia-free state and in the dementia state for a population of 2.000 individuals. We chose this sample size because

we assessed that T_{00} and T_{11} were not sensitive to changes in sample size. Reliable estimates were already obtained with as few as 500 individuals.

Furthermore, we observed that the time spent in the disease state was not realistically represented in the dataset with 5.000 individuals. In fact, these estimates deviated from those obtained with all other sample sizes, suggesting that they might be biased due to dataset-specific characteristics. We provide further evidence for these considerations through the plots in the appendix A (see figure A.11 and A.12).

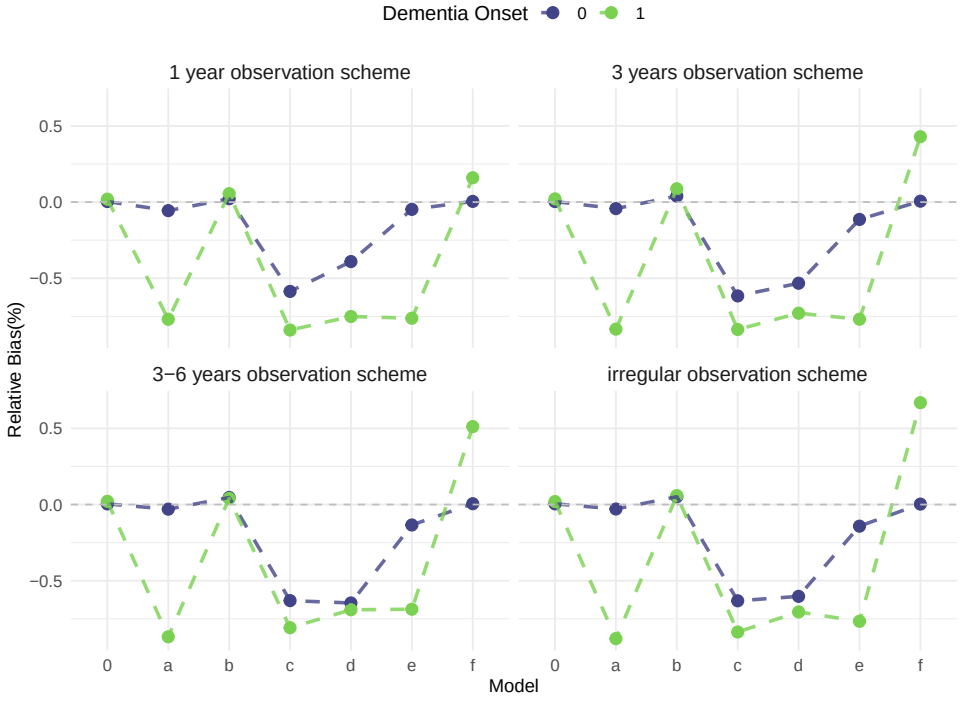


Figure 4.12: **Relative Bias of the total length of time spent in Dementia-free state and in Dementia state computed with different models across different observational schemes for a population of 2000 individuals**

0) Benchmark model. a) Semi-parametric multi-state model assuming exact transition times. b) Parametric multi-state model assuming exact transition times (Gompertz baseline). c) Time-homogeneous multi-state model for interval-censored data d) Time-homogeneous multi-state model or interval-censored data with 'age' covariate. e) Parametric time-inhomogeneous multi-state model for interval-censored data (Gompertz baseline). f) Multiple imputation for panel data strategy (Gompertz baseline)

Additionally, Fig. 4.12 presents a comparison of the relative bias associated with these estimates for a population of 2000 individuals. The relative bias was used to make the comparison between time spent in dementia and dementia-free states more meaningful, as the latter is generally larger and thus subject to greater absolute errors.

When estimating T_{00}, we find that all strategies perform well, except for the time-homogeneous models (c and d), which assumes constant hazards and consequently underestimates this quantity. Among the methods considered, ***f* exhibits the lowest relative**

bias, followed by a and b.

In the estimation of T_{11} , the semi-parametric multi-state model proves inadequate, significantly underestimating the time spent in the dementia state. Methods c, d, and e exhibit similar underestimation, as previously observed for the estimation of T_{00}. Although incorporating *age* as a covariate reduces the bias of model c, this adjustment is not sufficient to fully correct for the underestimation.

In contrast, method f slightly overestimates the duration spent in the dementia state. **For predicting $\boldsymbol{T_{11}}$**, the best-performing model is the Gompertz multi-state model that ignores interval censoring (***b***). Finally, we observe that the **relative bias did not vary significantly across different observational schemes**, but gets slightly worst when considering extensive or irregular schemes.

4.7.2. Results from simulation study part 2

For this second part of the results, we will follow the same structure as the previously presented findings. However, one exception is that we will not further investigate the convergence rate and average computational time. This is because the models applied to the datasets generated under the Semi-Markovian assumption do not experience convergence issues and consistently satisfy the convergence criteria imposed by the maximization algorithm used in their implementation.

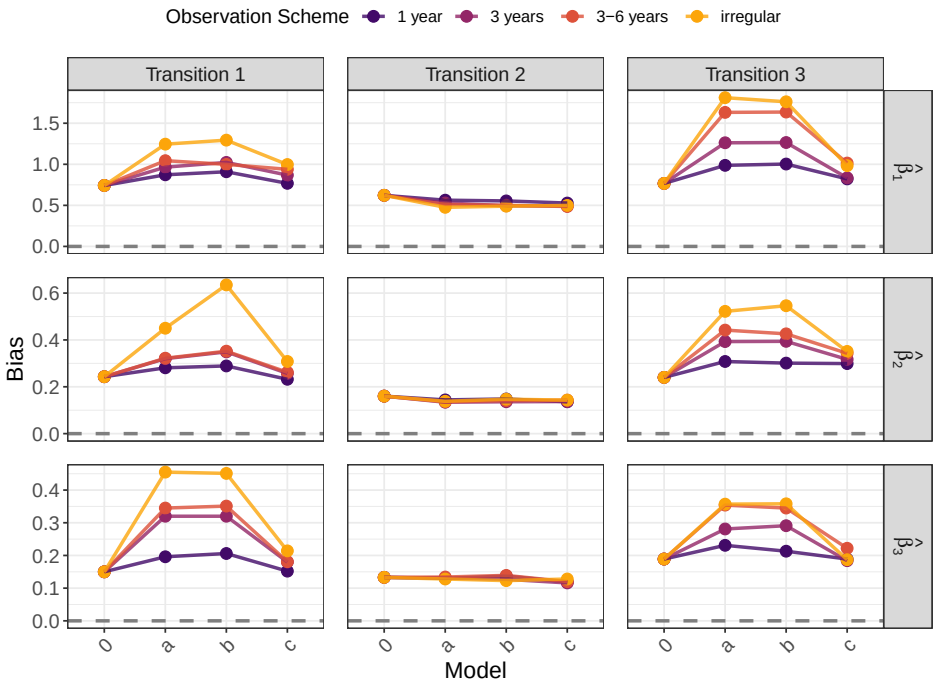
Regarding the analysis of computational time, the conclusions for strategies a, b, and c are identical to those already discussed for models a, b, and f in the previous section.

Finally, the comparison of baseline hazard parameters will also be omitted, as the results align with those already observed. Additionally, comparing only two parametric models would provide limited insights and is therefore not particularly informative.

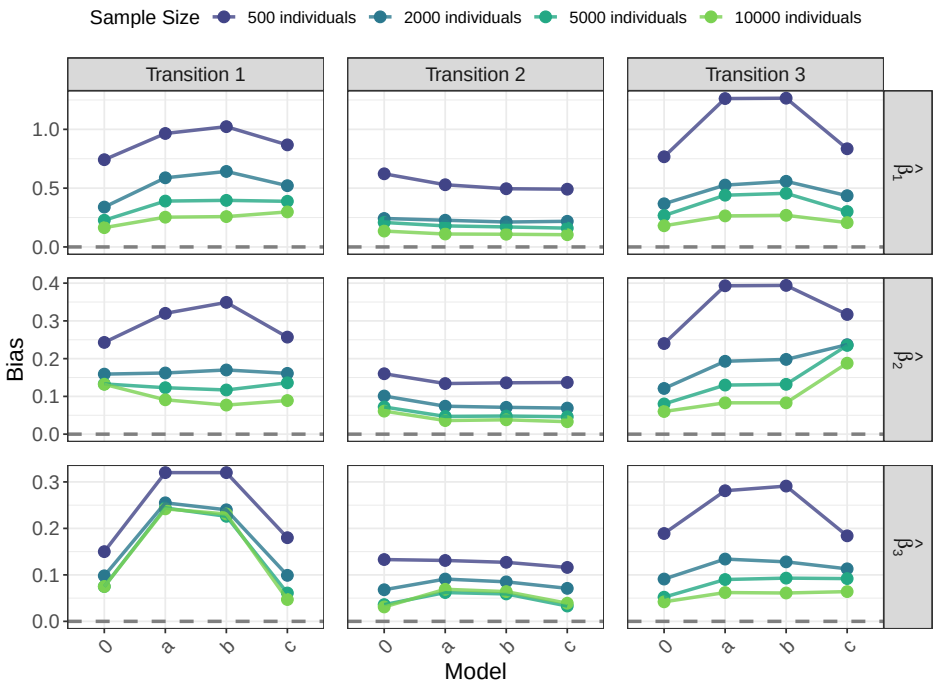
Estimated effect of covariates for different modeling strategies

The impact of covariates on the various modeled transitions, using different estimation strategies, was first assessed through the **absolute bias** of the estimated covariate coefficients.

We begin by analyzing the effect of different observation schemes on a population study comprising 500 individuals, as shown in Fig. 4.13a. Subsequently, we examine how absolute bias changes with increasing sample size under a triennial observation scheme in Fig. 4.13b.



(a) Absolute bias with 500 observations.



(b) Absolute bias with varying sample sizes under triennial observation scheme.

Figure 4.13: **Comparison of absolute bias of covariate coefficient estimates for different models across different observation schemes and sample sizes.**
0) Benchmark model. a) Semi-parametric multi-state model assuming exact transition times b) Parametric multi-state model assuming exact transition times (Gompertz baseline). c) Multiple imputation for panel data strategy (Gompertz baseline)

The analysis reveals that for transition 2, where entry into the study and either death or dropout times are precisely known, there are no significant differences across modeling strategies and observation schemes. The uncertainty introduced by the interval censoring mechanism primarily affects the transition to dementia and the transition from dementia to death. In these cases, wider and irregular observation intervals increase the bias of strategies that ignore interval censoring (a, b). The **multiple imputation for panel data strategy mitigates this effect under observation schemes B, C, and D**, showing a general trend of outperforming other strategies.

As the sample size increases, the bias associated with the covariate coefficients decreases, and the overall pattern remains consistent with previous observations. However, some exception are observed for the multiple imputation for panel data strategy: in certain cases (e.g, $\hat{\beta}_1$ $\hat{\beta}_2$ in transition 1 and $\hat{\beta}_2$ in transition 2), its bias reduction is less pronounced compared to methods a and b, making it less favorable under those specific cases.

Next, we investigate the **95% coverage** of the confidence intervals for the estimated coefficients.

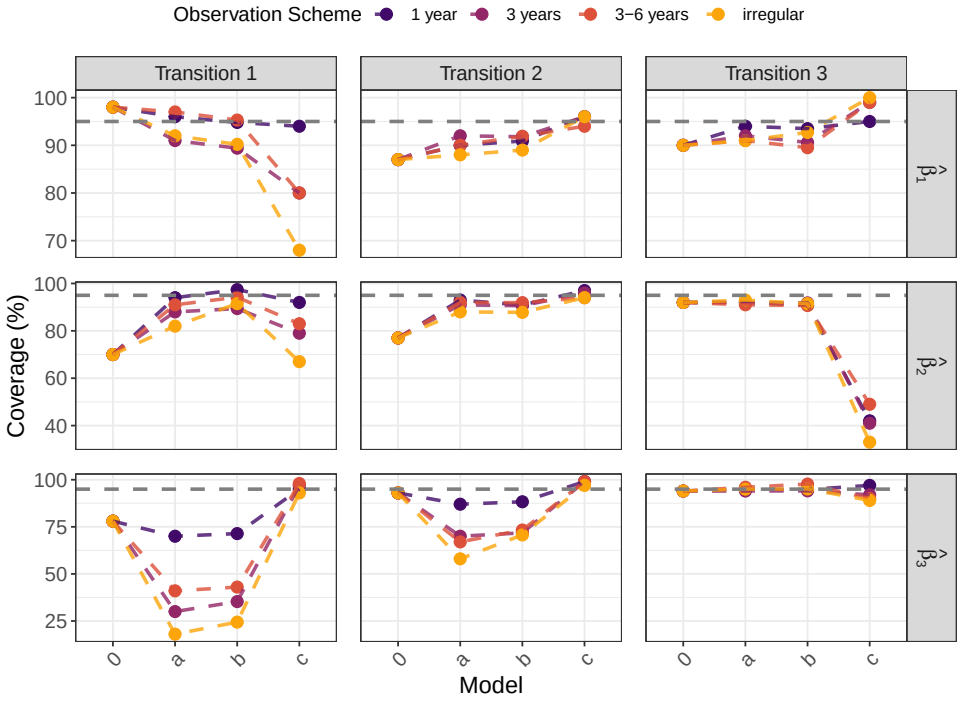


Figure 4.14: **Coverage of covariate coefficient estimates for different models across different observation schemes over 5000 individuals**

0) Benchmark model. a) Semi-parametric multi-state model assuming exact observation times. b) Parametric multi-state model assuming exact observation times (Gompertz baseline). c) Multiple imputation for panel data strategy (Gompertz baseline)

We first compare the effect of different observation schemes on a population of 5000 individuals in Fig. 4.14. Then, we examine how coverage evolves with increasing sample size under an annual observation scheme in Fig. 4.15.

Models that ignore interval censoring (e.g., a and b) maintain nearly constant coverage across different scenarios, though performance declines slightly as observation intervals increase.

The observed **coverage trend across different estimation strategies aligns with the bias results**. Specifically, in cases where the measured bias was substantial, the confidence intervals are notably shifted, leading to lower coverage. For instance, in model c, the coverage associated with the second covariate—characterized by very low prevalence—is particularly low in the first and third transitions, as well as for the continuous covariate in the first transition.

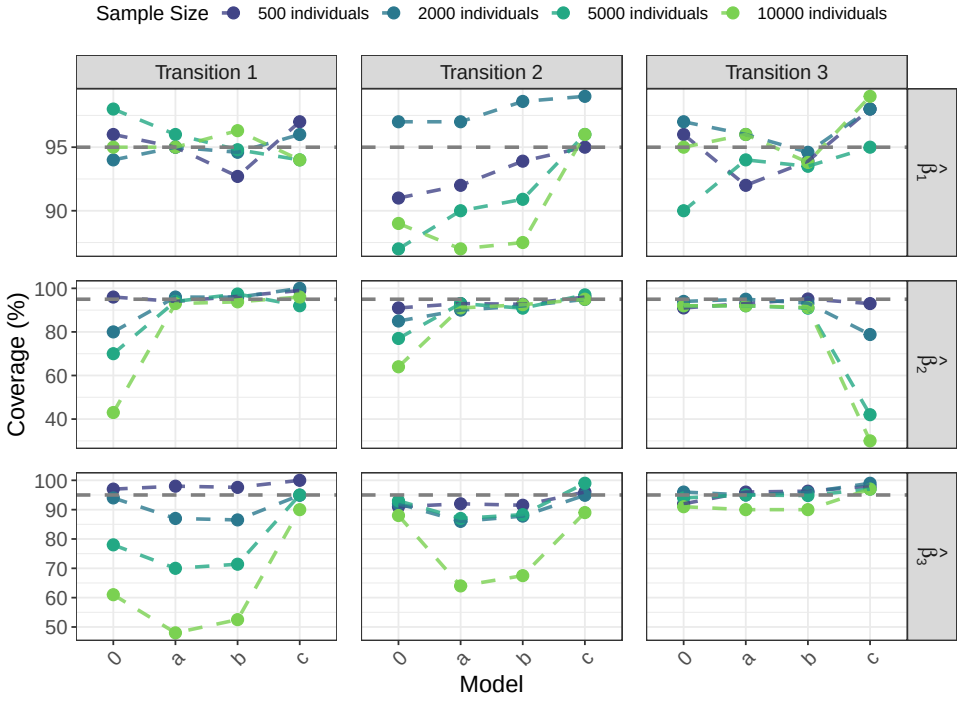
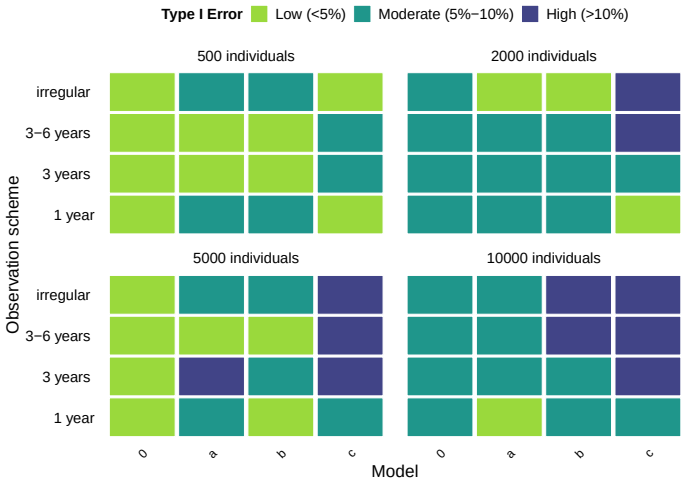


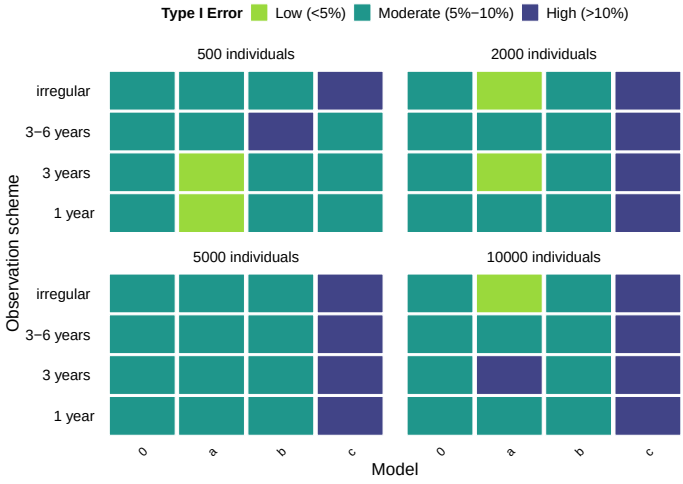
Figure 4.15: **Coverage of covariate coefficient estimates for different models across different sample sizes for annual observation scheme**
0) Benchmark model. a) Semi-parametric multi-state model assuming exact observation times. b) Parametric multi-state model assuming exact observation times (Gompertz baseline). c) Multiple imputation for panel data strategy (Gompertz baseline)

Overall, no clear pattern emerges regarding the effect of observation schemes on coverage. However, as the sample size increases, coverage deteriorates for methods that already exhibited high bias, consistent with previous findings part 1.

Type I Error was assessed by examining covariates known to be non-influential in the data simulation process. The results indicate that the Type I Error rate remains largely unaffected by variations in observational schemes.



(a) Categorized Type I Error associated to $\hat{\beta}_1$ in transition to Dementia



(b) Categorized Type I Error associated to $\hat{\beta}_2$ in transition from Dementia to Death

Figure 4.16: **Categorized Error Type I Error across different observation schemes and sample sizes.**
0) Benchmark model. a) Semi-parametric multi-state model assuming exact observation times. b) Parametric multi-state model assuming exact observation times (Gompertz baseline). c) Multiple imputation for panel data strategy (Gompertz baseline)

The best overall performance is achieved using the semi-parametric model that ignores interval censoring (a). Parametric models appear to face greater challenges in estimating the effect of covariates when these are not significant. Moreover, **compared to results obtained under the data-generating scenario based on the Markov assumption, the performance of the multiple imputation for panel data strategy deteriorates considerably**. This decline may be attributed to the different imputation approach used for disease status. Specifically, incorporating the effect of covariates in the transition from dementia to death as a reinforcing element in the probability estimation of developing dementia may have further distorted the distribution of coefficients associated to those covariates.

We then analyzed the **power** of different modeling strategies in determining the significance of a covariate.

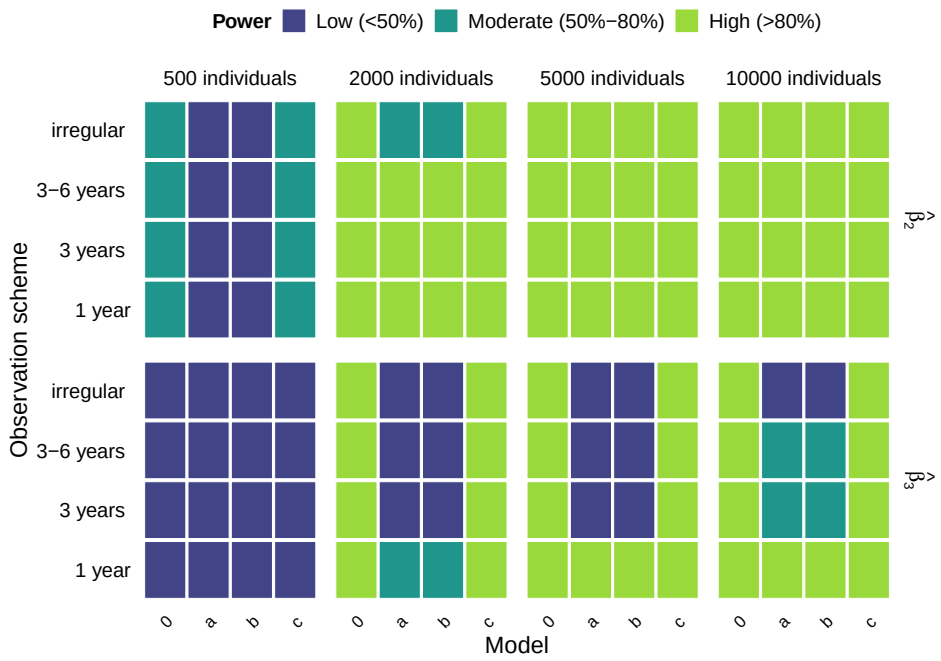


Figure 4.17: **Categorized Power of models across different observation schemes and sample sizes for transition to Dementia**
0) Benchmark model. a) Semi-parametric multi-state model assuming exact observation times. b) Parametric multi-state model assuming exact observation times (Gompertz baseline). c) Multiple imputation for panel data strategy (Gompertz baseline)

As expected, **statistical power increases with larger sample sizes**, a trend consistently observed in our analysis.

Focusing on the transition to dementia (Fig.4.17), we observe that models that ignore interval censoring tend to exhibit lower power under observation schemes B, C, and D.

Among all strategies, **model c demonstrates the best performance**, achieving power levels exceeding 80% for both coefficients even with moderate sample sizes. Moreover, it is the only model that matches the benchmark model's performance in detecting the significance of $\hat{\beta}_3$.

Similar to the findings under the data-generating scenario based on the Markov assumption, none of the models successfully detect the significance of the continuous covariate in the transition from dementia to death, even when the sample size is large.

Further power analyses for transitions 2 and 3 are presented in Figures B.7 and B.8.

Average life expectancy

We now compare the average time spent without developing dementia (T_{00}) since entry into the study and the average time spent in the dementia state (T_{11}) as predicted by different models.

These predictions are obtained through the simulation of individuals' trajectories, following the same approach used in part 1. The key difference in this case is that we explicitly account for the time spent in the dementia state in the simulation process.

Model	T_{00}	T_{11}	Scheme
0	**20.6**	**2.24**	**benchmark**
a	**20.6**	4.60	
b	21.0	**2.13**	A
c	20.8	2.63	
a	**20.6**	2.79	
b	21.3	**1.99**	B
c	21.1	2.97	
a	**20.7**	**2.09**	
b	21.4	2.00	C
c	21.1	3.03	
a	**20.7**	1.90	
b	21.4	**1.93**	D
c	21.1	3.15	

Table 4.9: Average total length of time spent in the Dementia-free state and in the Dementia state computed with different models across different observational schemes for a population of 2000 individuals.

Table 4.9 presents the estimated average total time spent in the dementia-free state and in the dementia state for a population of 2000 individuals. This sample size was chosen since **T_{00} and T_{11} are not highly sensitive to changes in sample size**. In fact,

reliable estimates were already obtained with as few as 500 individuals. Further evidence supporting this consideration is provided in Appendix B, where Figures B.9 and B.10 illustrate the stability of these estimates across different sample sizes.

Additionally, Figure 4.18 presents a comparison of the relative bias associated with these estimates for a population of 2000 individuals.

When estimating T_{00}, we find that all strategies perform well, exhibiting approximately the same relative bias. However, for the estimation of T_{11}, the semi-parametric multi-state model proves inadequate, showing considerable fluctuations between underestimation and overestimation depending on the observation scheme.

In contrast, method c tends to slightly overestimate the duration spent in the dementia state. Among the considered approaches, the best-performing model for predicting T_{11} is the Gompertz multi-state model that ignores interval censoring (b).

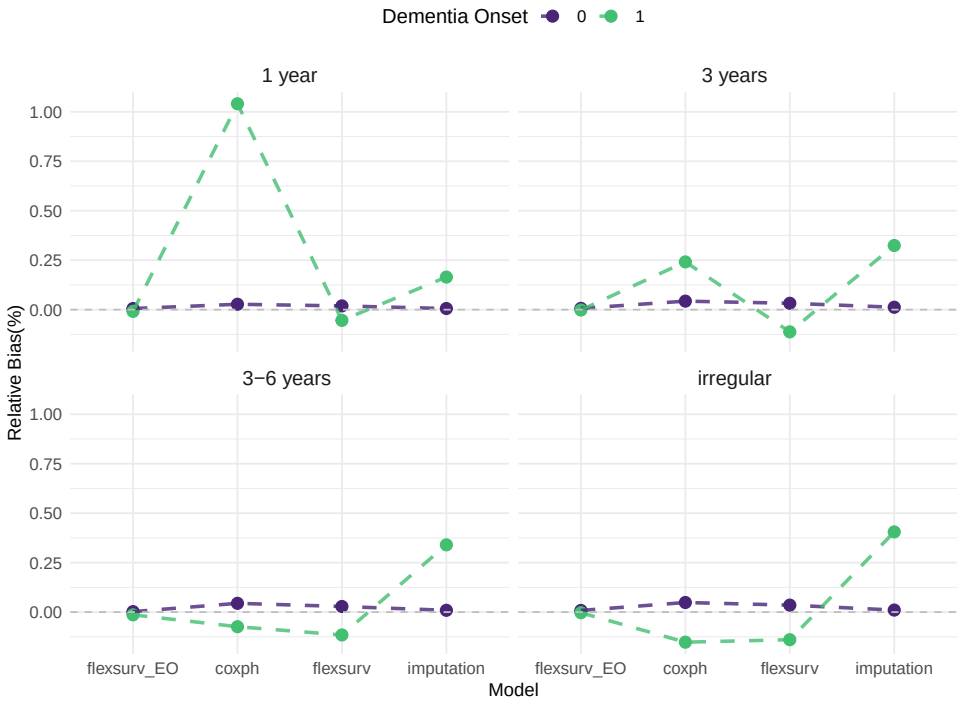


Figure 4.18: **Relative Bias of the total length of time spent in Dementia-free state and in Dementia state computed with different models across different observational schemes for a population of 2000 individuals**

0) Benchmark model. a) Semi-parametric multi-state model assuming exact transition times. b) Parametric multi-state model assuming exact transition times (Gompertz baseline). c) Multiple imputation for panel data strategy (Gompertz baseline)

4.7.3. Considerations on MIPD

After examining the performance of the various methods presented in this thesis, it is crucial to understand why the Multiple Imputation for Panel Data (MIPD) method provided good estimates for nearly all the quantities of interest. As previously discussed, this technique allows us to impute the exact transition times, as well as determine the disease status for each individual. Specifically, it helps us to record whether an individual has developed dementia where this transition was not observed, because the individual was not subjected to a follow-up visit that would have allowed for the diagnosis either due to death or withdrawal from the study before the event could be recorded.

By imputing these missing transition times and disease statuses, we are able to recover valuable information that would otherwise be lost. Therefore, it is essential to quantify the extent to which MIPD helps recover "missing" cases of dementia that were not observed in the original dataset.

To evaluate this, we will examine a sample of 2000 individuals and present the proportion of observed dementia cases for each observational scheme, averaged across 100 simulated datasets. We will compare this observed proportion with the same metric calculated on the imputed datasets. The imputed datasets provide an estimate of how many additional cases of dementia are identified when using MIPD. This analysis will be conducted both for data generated under Markovian assumption and for data generated under Semi-markovian assumption.

Scheme	Demented	Demented MIPD	Recovered (%)
A	0.158	0.188	+19.05
B	0.107	0.149	+39.25
C	0.099	0.143	+44.44
D	0.083	0.130	+56.63

(a) Data generated under Markovian assumption

Scheme	Demented	Demented MIPD	Recovered (%)
A	0.155	0.188	+21.29
B	0.097	0.157	+61.86
C	0.088	0.156	+77.27
D	0.072	0.160	+122.22

(b) Data generated under Semi-Markovian assumption

Table 4.10: Comparison of the proportion of recorded diseased before and after the application of MIPD for each observational scheme under Markovian and Semi-Markovian assumptions.

We can observe that in both cases, the proportion of observed dementia cases decreases as the length of observation intervals increases, compared to the true proportion of approximately 22% of actual dementia cases. With the multiple imputation technique for panel data, we are able to "recover" the diagnosis of many patients. **The percentage of recovered cases is particularly significant in schemes B, C, and D, and it becomes even more relevant when considering the time spent in the dementia state.**

4.8. Discussion

In this section, we conclude our simulation study by contextualizing our findings within the existing literature. To add further validity to our results, we aim to compare them with previous works. However, given that such a comprehensive overview of multi-state models for interval-censored data has not been conducted before, we can only rely on pairwise comparisons to either validate or challenge our findings.

A relevant reference in this context is the work by Leffondré et al. [15], who compared the accuracy of effect estimates obtained from a **semi-parametric illness-death Markov model for interval-censored data against those from a standard Cox model**. A key strength of this comparison is that it is based on the PAQUID study, a French population-based cohort on dementia development, making it particularly relevant to our study.

The main conclusion of Leffondré's work is that when follow-up intervals are wide and the exposure affects the risk of death, the illness-death model for interval-censored data should be preferred over the standard Cox regression analysis. In particular, the evaluation of model performance is based on two key metrics: the relative bias and the 95% coverage associated with the coefficient of the covariate influencing the transition from a dementia-free state to dementia.

A crucial finding from this study is that the relative bias of this coefficient is significantly lower when using a modeling strategy that accounts for the uncertainty introduced by interval and event censoring. Specifically, this means incorporating the possibility that an individual may develop dementia between their last visit and death or loss to follow-up. This effect is particularly pronounced when covariates influencing mortality risk are included in the model.

In Leffondré's work, the illness-death model for interval-censored data is estimated using a semi-parametric approach, where the baseline hazard is modeled with penalized splines.

Consequently, there is no direct comparison with the specific strategies explored in our simulation study. Therefore, we consider all strategies designed for interval-censored data (models c, d, e, f) as the second term of comparison, while we have a direct comparison with the standard Cox model (model a).

To assess the impact of interval-censoring modeling strategies, we focus on the same estimands chosen in Leffondrè's work, specifically the covariates effects influencing the transition from a dementia-free state to dementia ($\hat{\boldsymbol{\beta}}_{\mathbf{01}}$). We draw the following conclusions:

- **The absolute bias is reduced when using models that account for interval-censored data** compared to methods that ignore this uncertainty, particularly in settings with wide or irregular observation intervals. However, it is essential to balance the advantages of modeling interval censoring against the potential drawbacks of baseline hazard misspecification. As a result, in our analysis, models d, e, f are preferred over model a, but model c does not show the same improvement.
- **The 95% coverage increases when using models that account for interval-censored data under observations schemes with wide intervals**, provided that the selected strategy is not affected by an ill-conditioned Hessian matrix, which could lead to incorrect confidence interval estimation. Consequently, model f is preferred over model a, whereas models c, d, e do not exhibit a clear advantage in this respect.

These findings emphasize the importance of explicitly modeling interval censoring in multi-state models, especially in the presence of irregular observation schemes and covariates affecting both disease progression and mortality, provided that these strategies do not lead to misspecification of the baseline hazard shape.

Another relevant study that we considered is **the work conducted by Kendall et al.** [12]. In their paper, the authors provide insights and illustrations on the potential **biases in parameter estimation that may arise when the assumption of piecewise-homogeneous transition rates is violated**. They advocate for a computation of the likelihood function leveraging a fully time-inhomogeneous approach.

The key conclusion from this study is that assuming piecewise-constant transition rates—where changes occur at discrete time points—can introduce non-negligible biases in estimation when the underlying process follows a continuous-time Markov model. Numerical solutions to the Kolmogorov Forward Equations (KFE) lead to more consistent estimations, which can only be matched by piecewise approximations when an extremely fine

resolution grid is chosen. However, such a fine discretization significantly increases computational time compared to solving the KFE directly.

The comparison in Kendall et al.'s work is based on performance measures evaluating the accuracy of baseline hazard parameter estimation. In our analysis, a similar comparison can be drawn between model d, which assumes piecewise time-homogeneous transition rates, and model e, which numerically solves the KFE.

In our specific case, the assumption of piecewise-constant transition rates appears reasonable since the transition hazard changes with an individual's age, which naturally defines the time intervals. While this assumption is pragmatic, it limits our ability to assess performance across different levels of resolution and discretization. As a result, **we find that baseline parameter estimates from model *d* are consistently worse than those from model *e*, likely due to the coarser resolution used**. However, this coarser resolution makes model d significantly more computationally efficient, though not necessarily the best choice for achieving highly accurate estimates.

In this simulation study, we have systematically analyzed the effects of different observation schemes, the implications of misspecifying the functional form of the baseline hazards, and the consequences of ignoring interval censoring, comparing our findings with existing literature where possible.

One potential limitation of our study is that we did not assess the impact of ignoring the time spent in the dementia state when the underlying process actually depends on it. In other words, we did not investigate the estimation error that arises when Markov models are applied to data generated under a Semi-Markovian assumption. This aspect is relevant because methods for estimating Semi-Markov multi-state models in the presence of panel data are limited. As a result, researchers might be forced to adopt simpler assumptions that disregard the time spent in the states. The reason why we did not pursue this analysis is that, when evaluating Markov multi-state models for interval-censored data, we had already identified multiple factors contributing to model inaccuracy. Introducing an additional layer of complexity—specifically, assessing the impact of misspecifying the Markovian assumption—would have made it difficult to isolate its specific effect on model performance.

A possible approach would have been to focus on the best-performing Markovian strategy, namely multiple imputation for panel data, and test its robustness when applied to Semi-Markovian data. However, since this method is implemented under Semi-Markov assumption as well, such a comparison would not have been meaningful.

A crucial aspect of any simulation study is the quantification of the Monte Carlo Standard Error (MCSE), which measures the uncertainty introduced by using a finite number of simulated datasets n_{sim}. The MCSE provides an estimate of the standard error associated with the estimated performance measures, reflecting how much they would vary if we repeatedly ran the simulation study with different random seeds and it can be computed without any knowledge on the true value of the parameters.

In designing our simulation study, we aimed to balance computational feasibility with the need for a small MCSE . Ideally, increasing n_{sim} reduces Monte Carlo Standard Error, as it converges at a rate of $1/\sqrt{n_{\text{sim}}}$. However, given the complexity of our models and the computational cost associated with each run, we opted for $n_{\text{sim}} = 100$, which provides a reasonable trade-off between computational efficiency and precision.

Further studies with enhanced computational resources could refine these estimates by increasing the number of repetitions, thereby improving the robustness of our findings.

5 | Recommendations for researchers

In this chapter, we outline key principles from our study which may offer practical recommendations for researchers using multi-state models in panel data analysis. These recommendations address critical issues discussed in the aims of the simulation study, including the influence of sample size on estimation consistency, the impact of various modeling choices and the implications of ignoring interval censoring.

Sample size considerations

- The sample size should be sufficiently large to ensure that a **meaningful number of events of interest** (e.g., dementia onset) **are observed**, providing enough information to fit the model reliably. Naturally, this required quantity depends on the incidence of the disease in the studied population. By leveraging the benchmark model, where all occurring events are assumed to be observed, we estimate that approximately 500 observed events are necessary to ensure that the effects of underrepresented covariates—such as those with low prevalence or continuous variables spanning a wide range—are accurately estimated, leading to satisfactory statistical power.
- Models based on the **numerical solution of Kolmogorov Forward Equations (KFE)** with time-varying coefficients tend to perform poorly in small population studies. These models should be avoided when the number of observed events of interest is fewer than 150, as they frequently encounter convergence issues and produce unreliable results.

Modeling choices

- **Semi-Parametric vs. Parametric models:**
  - Semi-parametric models, such as transition-specific Cox models, provide flex-

ibility and are a suitable choice when there is no prior knowledge about the transition rates.

  - Parametric models are preferable when baseline hazards follow known distributions. They enable more precise estimation with fewer data requirements and facilitate the extrapolation of estimands.

- **Time-homogeneous vs. Time-inhomogeneous models:**

  - If prior knowledge suggests that transition hazards change smoothly over time, assuming time-homogeneity can have severe consequences for the estimation of any quantity of interest. Such an assumption could indicate the model's inadequacy in capturing the complexity of the process, leading to potential biases in the results.

  - When the risk of transitioning is nearly constant over time, time-homogeneous models are the preferred choice, as they allow for closed-form solutions to KFE and require less computational time.

  - Piecewise constant transition rates can be useful, but only when the discretization grid is coarse. If smooth hazard modeling is required, a time-inhomogeneous model is preferable over a finer discretization grid, both in terms of performance and computational time.

- **Markov vs. Semi-Markov assumption:**

  - If the time spent in an intermediate state affects future transitions, the Semi-Markov assumption must be adopted, and the transition from the intermediate state should be modeled accordingly. Notably, incorporating this dependency significantly impacts the estimation of the average time spent in a given state. As a result, it becomes crucial when modeling quantities such as life expectancy following disease onset.

  - If the Markov property holds, Markov models are significantly more convenient, as there is a broad range of methodologies for handling multi-state models under this assumption, and their computational complexity is lower.

  - If Semi-Markov methods are required and there is a need to model uncertainty in panel data, Multiple Imputation for Panel Data (MIPD) serves as a valuable tool.

Handling Interval Censoring in Panel Data

- With **frequent observations** (e.g., annual follow-ups), it is reasonable to disregard interval censoring uncertainty and **assume transition times are exactly known**.
- If the observation scheme involves **infrequent or irregular follow-ups**, methods that **explicitly model interval censoring** should be preferred. They help to correct for the biases that arise in the measures of performance when transition times are not observed exactly. Among those methods, any technique that allows for smooth hazard variation should be preferred to minimize the bias. However, **MIPD is recommended over direct numerical solutions of KFE**, as it provides higher coverage of confidence intervals and generally leads to more reliable estimates in the presence of interval censoring.
- Ignoring interval censoring can result in **low Type I Error** but may lead to inaccurate estimates of transition rates and covariate effects. Conversely, explicitly modeling interval censoring when it's needed generally improves **statistical power**, enhancing the ability to detect true effects but at the risk of a higher Type I Error. This **trade-off** should be carefully considered based on the study's objectives.

Computational Efficiency

Strategies that explicitly model interval censoring are generally less favorable due to the additional computational burden. However, among these, **parallelized multiple imputation for panel data** is preferable. Parallelization helps reduce the computational time, making it feasible to handle larger datasets and more complex models.

Key Takeaways:

1. **No universally optimal strategy exists for multi-state modeling with panel data**: no single model outperforms all others across all scenarios. Instead, the best approach depends on the specific context (e.g. study design, observation scheme, computational constraints) and research goals.
2. When **interval censoring must be explicitly modeled**, **MIPD** is generally the preferred approach among those analyzed, as it balances bias reduction, coverage accuracy and computational feasibility.
3. When **interval censoring can be ignored**, the best strategy is to use a **parametric multi-state model assuming exact transition times**. This method is

computationally efficient and yields reliable estimates, provided that the assumed distribution accurately represents the underlying process.

6 | Conclusions and future developments

This thesis contributes to the growing field of multi-state modeling for panel data by synthesizing existing methods, identifying their strengths and weaknesses, and proposing a novel approach to address the limitations introduced by interval-censored data. Ultimately, we hope that this work will help to bridge the gap between theoretical advancements and practical applications, ensuring that multi-state models with intermittent observation schemes continue to evolve as valuable tools for understanding disease progression and informing healthcare policies.

Through the analyses carried out in this thesis, we have learned that observing continuous processes at discrete time intervals introduces two types of uncertainty: *interval censoring* and *event censoring*. The former arises from the inability to observe the exact transition time of individuals, leading to bias in the estimation of the main quantities. Event censoring, on the other hand, results in an underestimation of the number of observed events of interest, as they may have occurred before death or loss to follow-up, but were not recorded.

These types of phenomena are unavoidable since logistical and economic constraints prevent frequent observations of individuals in large-scale population studies. Hence, it is crucial to incorporate them appropriately into statistical models, adapting the modeling approach based on the degree of uncertainty.

Our study suggests that modeling uncertainty is not necessarily the best solution to obtain more robust estimates, since the computational challenges introduced often outweigh the benefits. Consequently, we believe that a promising approach is the Multiple Imputation for Panel Data (MIPD) method. This method retains all the advantages of multi-state models that assume exact transition times while being specifically designed to accommodate different parametric forms of hazard functions.

A key future objective is undoubtedly to focus on improving the imputation phase to enhance performance. More precise imputations would allow us to obtain estimates com-

parable to those of benchmark models. Furthermore, this method could be extended to accommodate non-parametric baseline hazard functions where needed, allowing for the development of transition-specific Cox models on datasets with imputed disease histories. Finally, as already highlighted in the computational results, parallelization strategies are necessary to make this method competitive in all aspects.

Several avenues for future research and methodological improvements emerge from this work:

- **Extending to multi-state models including backward transitions:** the current study focuses on a progressive illness-death model with three states. Expanding this framework to include additional intermediate states, such as mild cognitive impairment (MCI) before dementia, would provide a more granular representation of disease progression. This extension would allow us to explore the performance of different modeling strategies in the presence of an additional state that allows reversible transitions and to evaluate whether our conclusions hold under this scenario.
- **Refining software implementations for multi-state models with panel data:** currently, many advanced multi-state modeling approaches lack widely available and user-friendly software implementations. This limitation is particularly pronounced for methods based on Semi-Markovian assumption, given the increased complexity of the likelihood function. Future research should focus on improving such packages, making them more accessible to researchers without advanced programming expertise.
- **Applying these models to real-world longitudinal datasets:** while our simulation study provides a controlled evaluation of different modeling strategies, validating these approaches on real-world datasets will be crucial for assessing their robustness and practical utility. Incorporating data from large-scale epidemiological studies could provide valuable insights into the performance of different methods in real settings.

This thesis has laid the foundation for further research in multi-state modeling for panel data, providing a comprehensive comparison of methods and highlighting promising new approaches.

Bibliography

[1] M. E. Aastveit, C. Cunen, and N. Hjort. A new framework for semi-markovian parametric multi-state models with interval censoring. *Statistical Methods in Medical Research*, 32(6):1100– 1123, 2023. doi: 10.1177/09622802231160550.

[2] H. Aralis and R. Brookmeyer. A stochastic estimation procedure for intermittently-observed semi-markov multistate models with back transitions. *Statistical Methods in Medical Research*, 28(3):770– 787, 2019. doi: 10.1177/0962280217736342.

[3] C. Birkenbihl, Y. Salimi, and H. Fröhlich. Unraveling the heterogeneity in alzheimer's disease progression across multiple cohorts and the implications for data-driven disease modeling. *Alzheimer's Disease and Dementia*, 18(2):251–261, 2021. doi: https://doi.org/10.1002/alz.12387.

[4] Y. Chen, P. Bandosz, G. Stoye, and Y. L. et al. Dementia incidence trend in england and wales, 2002–19, and projection for dementia burden to 2040: analysis of data from the english longitudinal study of ageing. *The Lancet Public Health*, 8(11):859–867, 2023. doi: https://doi.org/10.1016/S2468-2667(23)00214-1.

[5] Elmståhl, Hagberg, Holst, Rennemark, Sjölund, Thorslund, Wiberg, Winblad, and W. A. A longitudinal study integrating population, care and social services data. the swedish national study on aging and care (snac). *Aging clinical and experimental research*, 16(2):158–168, 2004. doi: 10.1007/BF03324546.

[6] R. Hubbard, L. Inoue, and J. Fann. Modeling nonhomogeneous markov processes via time transformation. *Biometrics*, 64(3):843–850, 2008. doi: https://doi.org/10.1111/j.1541-0420.2007.00932.x.

[7] D. Incerti and J. P. Jansen. hesim: Health economic simulation modeling and decision analysis. *Statistics in Medicine*, 2021. doi: https://doi.org/10.48550/arXiv.2102.09437.

[8] C. Jackson. Multi-state models for panel data: The msm package for r. *Journal of Statistical Software*, 38(8):1–28, 2011. doi: https://doi.org/10.18637/jss.v038.i08.

[9] P. Joly, D. Commenges, C. Helmer, and L. Letenneur. A penalized likelihood approach for an illness–death model with interval-censored data: application to age-specific incidence of dementia. *Biostatistics*, 3(3):843–850, 2002. doi: https://doi.org/10.1093/biostatistics/3.3.433.

[10] J. Kalbfleisch and J. Lawless. The analysis of panel data under a markov assumption. *Journal of the American Statistical Association*, 80(392):863–871, 1985. doi: https://doi.org/10.1080/01621459.1985.10478195.

[11] M. Kang and S. Lagakos. Statistical methods for panel data from a semi-markov process, with application to hpv. *Biostatistics*, 2(8):252–264, 2007. doi: https://doi.org/10.1093/biostatistics/kxl006.

[12] E. B. Kendall, J. P. Williams, G. H. Hermansen, F. Bois, and V. H. Thanh. Beyond time-homogeneity for continuous-time multistate markov models. *International Journal of Epidemiology*, 2024. doi: https://doi.org/10.48550/arXiv.2211.03214.

[13] C. Law and R. Brookmeyer. Effects of mid-point imputation on the analysis of doubly censored data. *Statistics in Medicine*, 11(12):1569–1578, 1992. doi: 10.1002/sim.4780111204.

[14] J. Le-Rademacher, R. Peterson, T. Therneau, and et al. Application of multi-state models in cancer clinical trials. *Clinical Trials*, 5(15):489–498, 2018. doi: https://doi.org/10.1177/1740774518789098.

[15] K. Leffondré, C. Touraine, C. Helmer, and P. Joly. Interval-censored time-to-event and competing risk with death: is the illness-death model more accurate than the cox model? *International Journal of Epidemiology*, 42(4):1177–1186., 2013. doi: 10.1093/ije/dyt126.

[16] R. J. Machado and A. van den Hout. Flexible multistate models for interval-censored data: Specification, estimation, and an application to ageing research. *Statistics in Medicine*, 37(10):1587–1766, 2018. doi: https://doi.org/10.1002/sim.7604.

[17] R. J. Machado, A. van den Hout, and G. Marra. Penalised maximum likelihood estimation in multi-state models for interval-censored data. *Computational Statistics Data Analysis*, 153, 2021. doi: https://doi.org/10.1016/j.csda.2020.107057.

[18] A. Marshall, D. G. Altman, R. Holder, and P. Royston. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. *BMC Medical Research Methodology*, 9(57), 2009. doi: 10.1002/sim.2712.

[19] B. Mast and B. Yochim. *Alzheimer's Disease and Dementia*, volume 38 of *1*. Hogrefe Publishing, 1 edition, 1 2018. ISBN 9781616765033.

[20] T. Monfared, S. Fu, N. Hummel, L. Qi, A. Chandak, R. Zhang, and Q. Zhang. Estimating transition probabilities across the alzheimer's disease continuum using a nationally representative real-world database in the united states. *Neurology and Therapy*, (12):1235–1255, 2023. doi: https://doi.org/10.1007/s40120-023-00498-1.

[21] T. P. Morris, I. R. White, and M. J. Crowther. Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11):2074–2102, 2019. doi: https://doi.org/10.1002/sim.8086.

[22] H. Putter, M. Fiocco, and R. Geskus. Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine*, 26(11):2389–2430, 2007. doi: 10.1002/sim.2712.

[23] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023. URL `https://www.R-project.org/`.

[24] S. Ross. Simulation - chapter 4 - generating discrete random variables. *Academic Press*, pages 47–70, 2013. doi: //doi.org/10.1016/B978-0-12-415825-2.00004-8.

[25] B. Sun, S. Li, Y. Wang, and W. X. et al. Sarcopenia transitions and influencing factors among chinese older adults with multistate markov model. *Innovation in Aging*, 8(7), 2023. doi: https://doi.org/10.1093/geroni/igad105.

[26] Terry M. Therneau and Patricia M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer, New York, 2000. ISBN 0-387-98784-3.

[27] A. C. Titman. Flexible nonhomogeneous markov models for panel observed data. *Biometrics*, 67(3):780–787, 2011. doi: https://doi.org/10.1111/j.1541-0420.2010.01550.x.

[28] A. C. Titman. *Non-Homogeneous Markov and Hidden Markov Multistate Models*, 2023. current version 0.1.1.

[29] C. Touraine, T. A. Gerds, and P. Joly. Smoothhazard: An r package for fitting regression models to interval-censored observations of illness-death models. *Journal of Statistical Software*, 79(7):1–22, 2017. doi: https://doi.org/10.18637/jss.v079.i07.

[30] S. Wei and R. Kryscio. Semi-markov models for interval censored transient cognitive

states with back transitions and a competing risk. *Statistical Methods in Medical Research*, 25(6):2909– 2924, 2016. doi: https://doi.org/10.1177/0962280214534412.

[31] A. Wimo, M. Guerchet, G. Ali, Y. Wu, A. Prina, B. Winblad, L. Jönsson, Z. Liu, and M. Prince. The worldwide costs of dementia 2015 and comparisons with 2010. *Alzheimer's Dementia*, 13(1):1–7, 2017. doi: 10.1016/j.jalz.2016.07.150.

[32] *World report on ageing and health*. World Health Organization, 1 edition, 9 2015.

[33] B. Yu, J. Saczynski, and L. Launer. Multiple imputation for estimating the risk of developing dementia and its impact on survival. *Biometrical Journal*, 25(5):616–627, 2010. doi: 10.1002/bimj.200900266.

[34] M. Yuan, C. Xu, and Y. Fang. Multi-state models for bone marrow transplantation studies. *Statistical Methods in Medical Research*, 2(11), 2002. doi: https://doi.org/10.1191/0962280202sm277ra.

[35] M. Yuan, C. Xu, and Y. Fang. The transitions and predictors of cognitive frailty with multi-state markov model: a cohort study. *BMC Geriatr*, 550(22), 2022. doi: https://doi.org/10.1186/s12877-022-03220-2.

A | Appendix A

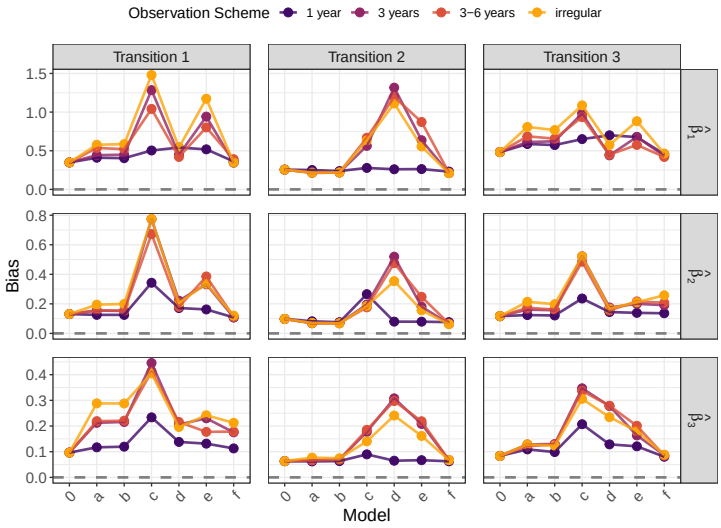


Figure A.1: Absolute bias of covariate coefficient estimates for different models across different observation schemes over 2000 individuals

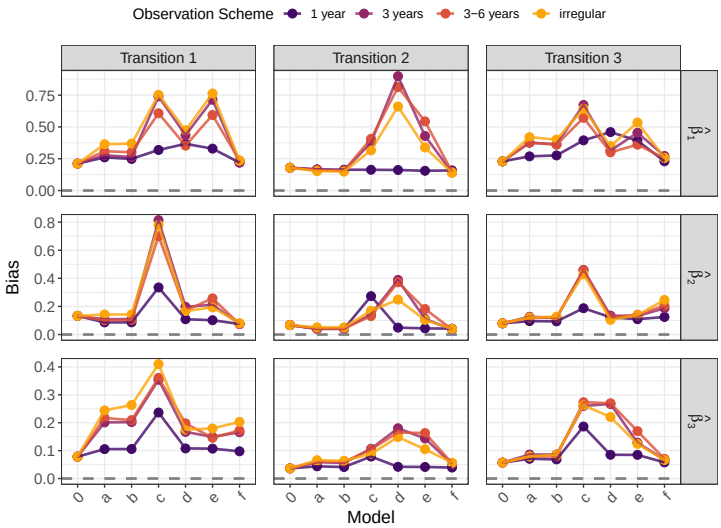


Figure A.2: Absolute bias of covariate coefficient estimates for different models across different observation schemes over 5000 individuals

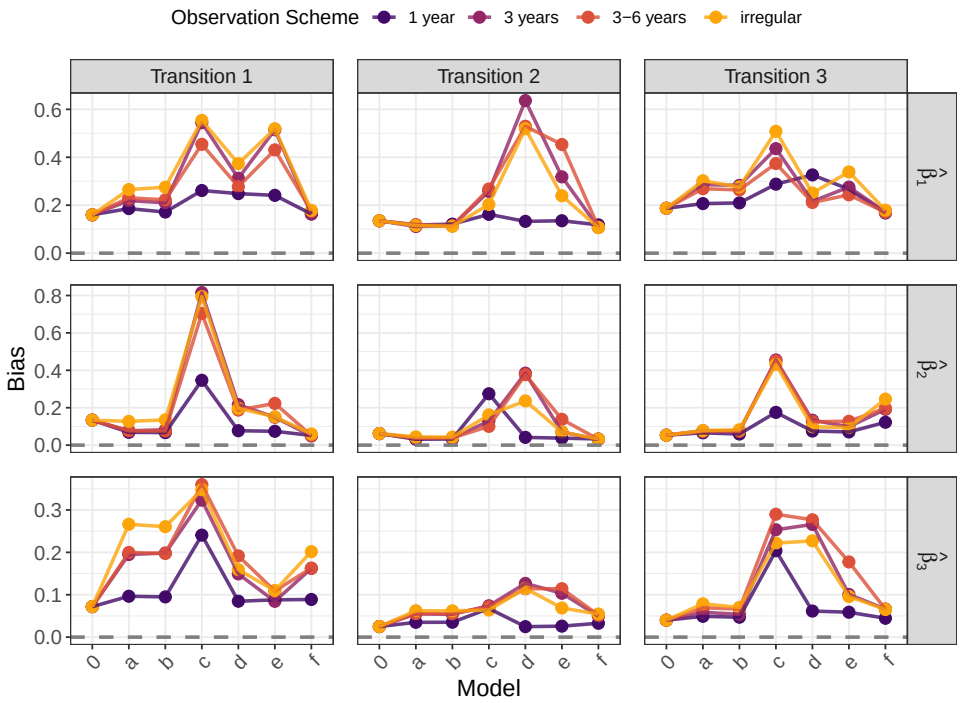


Figure A.3: Absolute bias of covariate coefficient estimates for different models across different observation schemes over 10.000 individuals

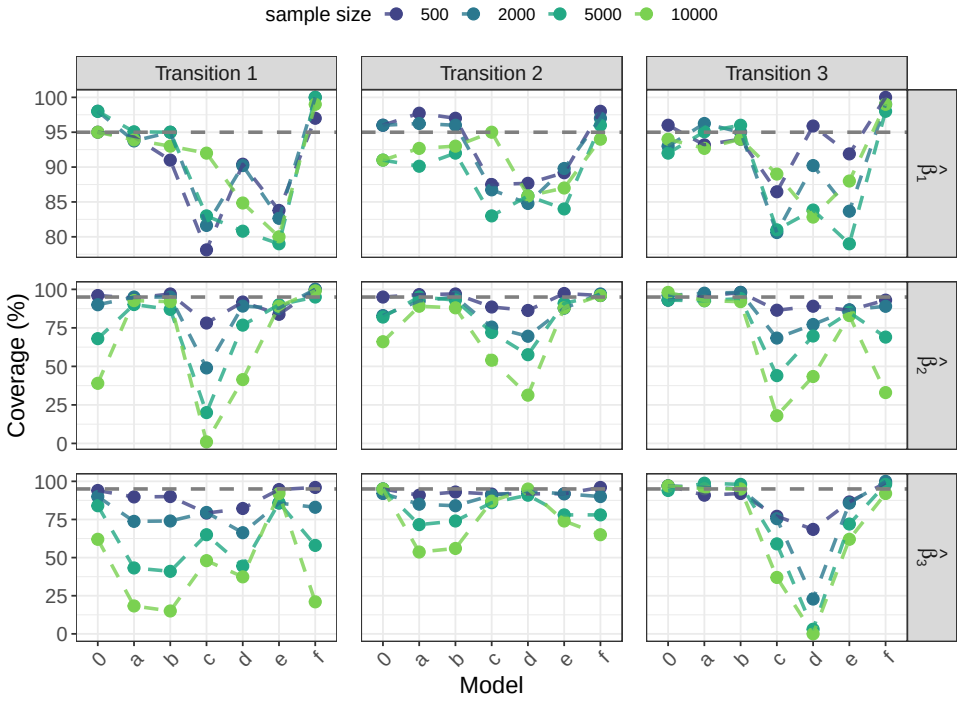


Figure A.4: Coverage of covariate coefficient estimates for different models across different sample sizes for scheme B

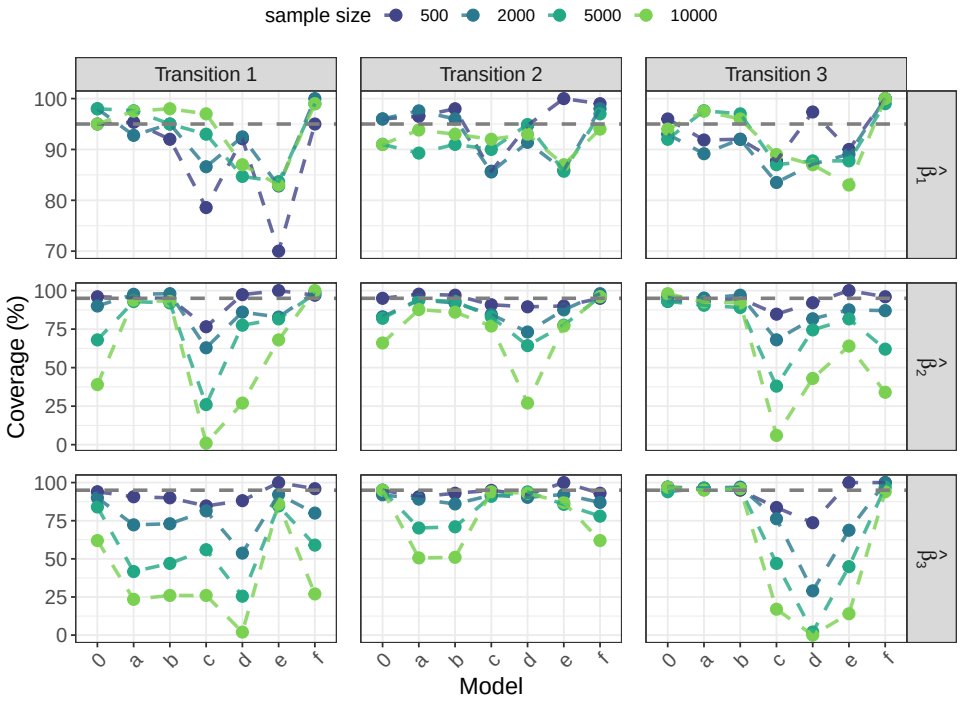


Figure A.5: Coverage of covariate coefficient estimates for different models across different sample sizes for scheme C

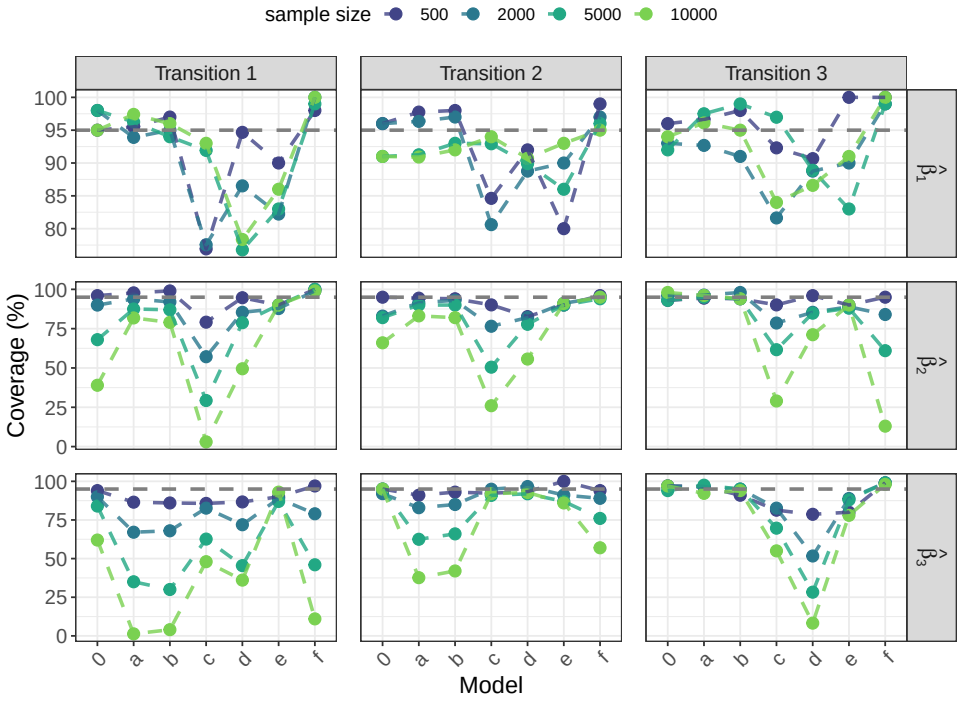


Figure A.6: Coverage of covariate coefficient estimates for different models across different sample sizes for scheme D

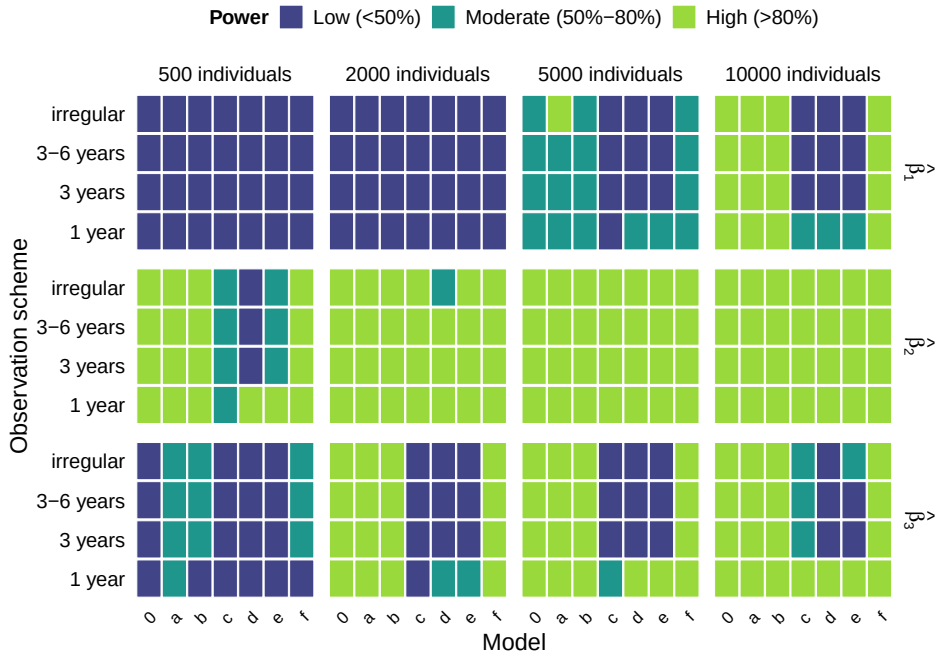


Figure A.7: Categorized Power of models across different observation schemes and sample sizes for transition from Dementia-free state to death

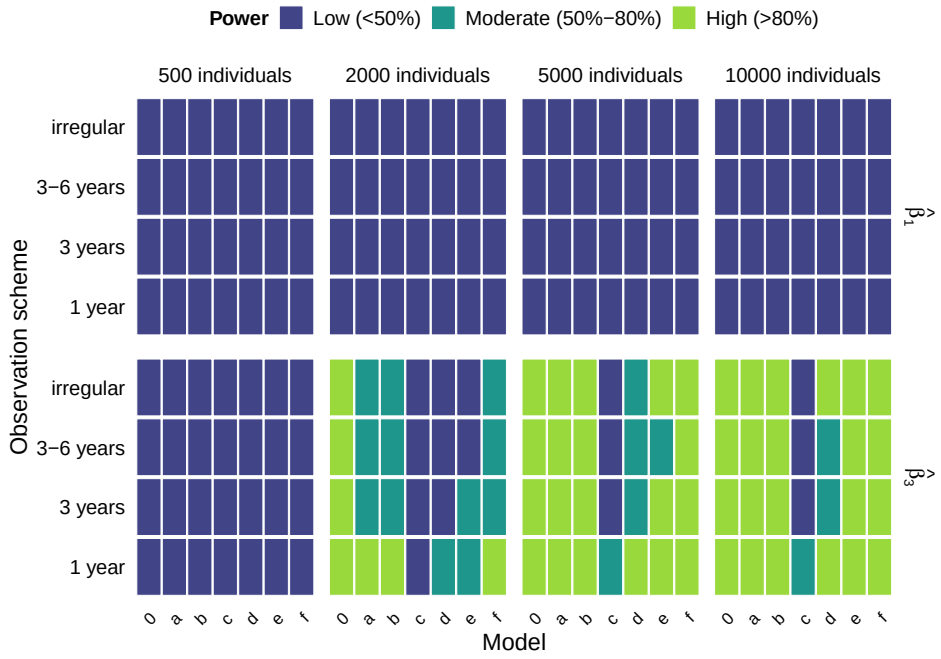


Figure A.8: Categorized Power of models across different observation schemes and sample sizes for transition from Dementia state to death

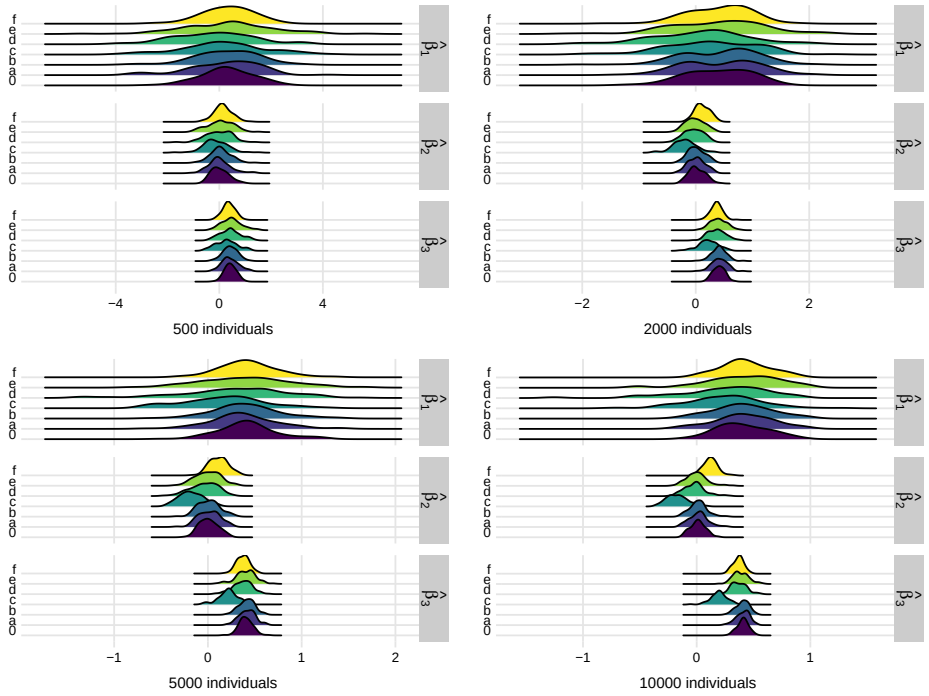


Figure A.9: Estimated coefficients' distribution for transition to Dementia in observational scheme C

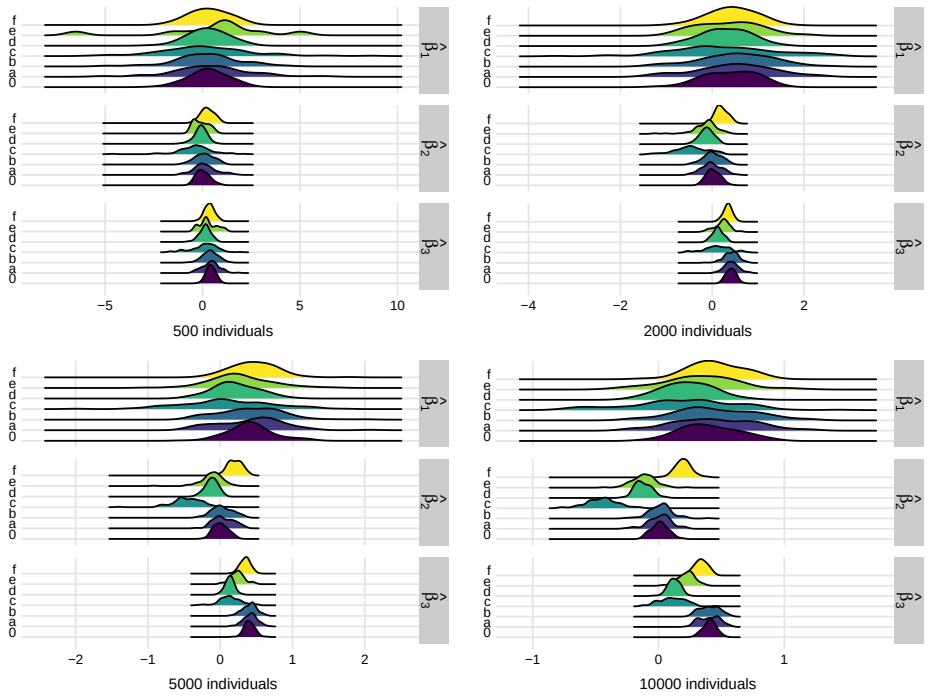


Figure A.10: Estimated coefficients' distribution for transition from Dementia to Death in observational scheme C

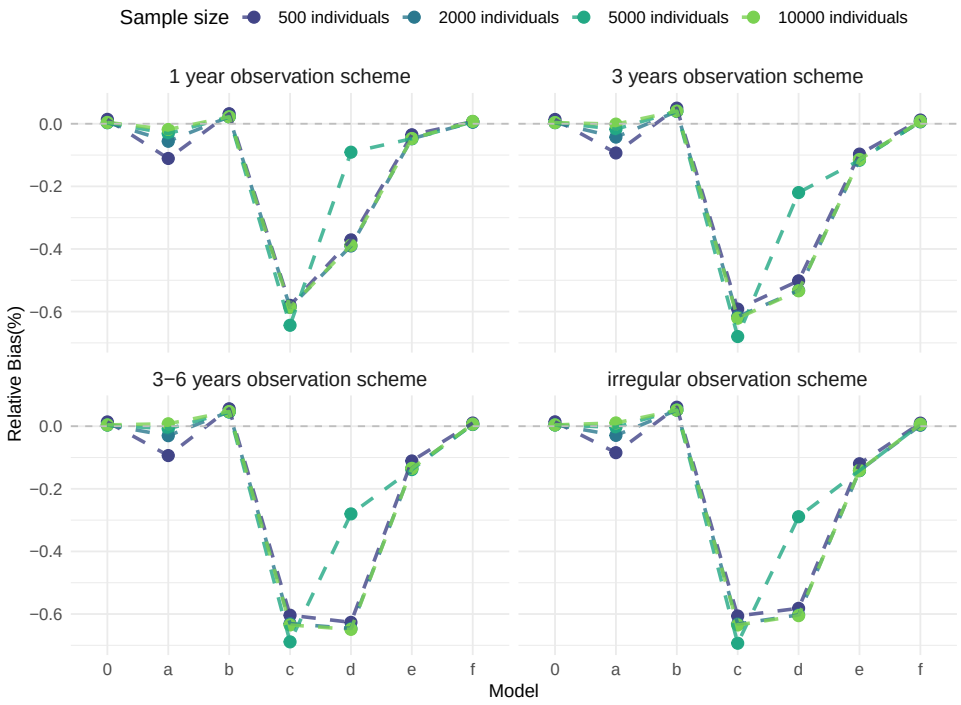


Figure A.11: Relative Bias of the total length of time spent in Dementia-free state computed with different models across different observational schemes and sample sizes

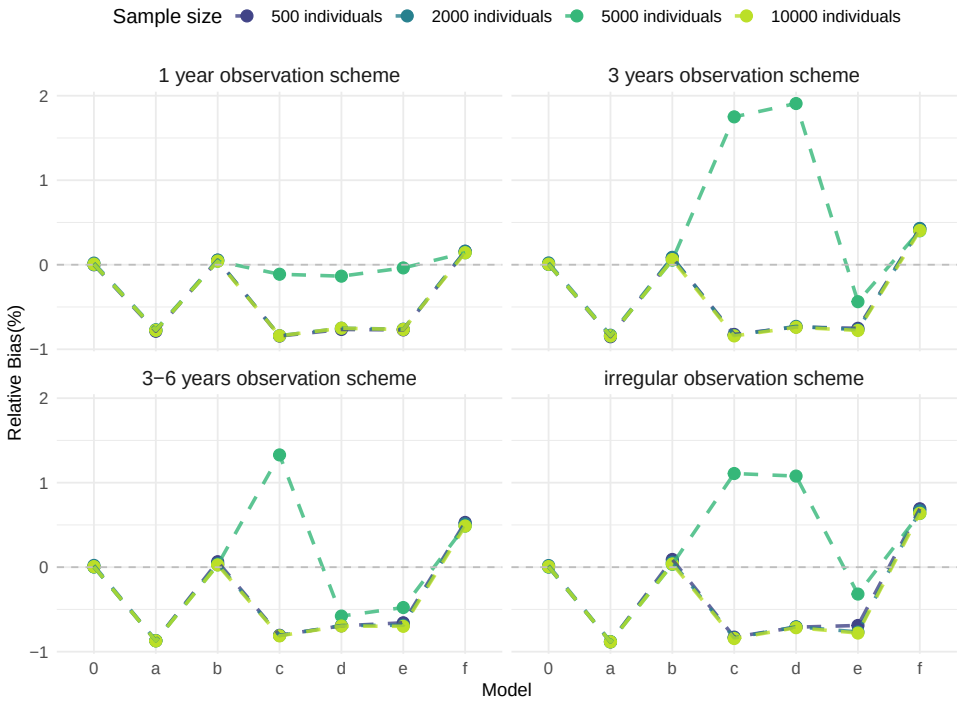


Figure A.12: Relative Bias of the total length of time spent in Dementia state computed with different models across different observational schemes and sample sizes

B | Appendix B

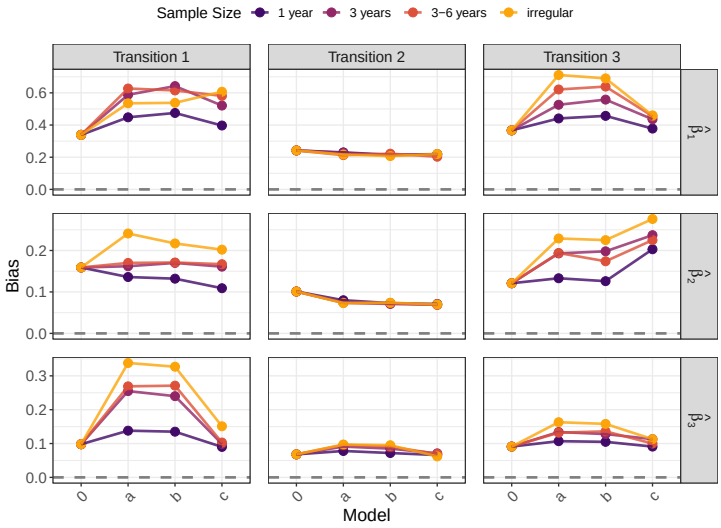


Figure B.1: Absolute bias of covariate coefficient estimates for different models across different observation schemes over 2000 individuals

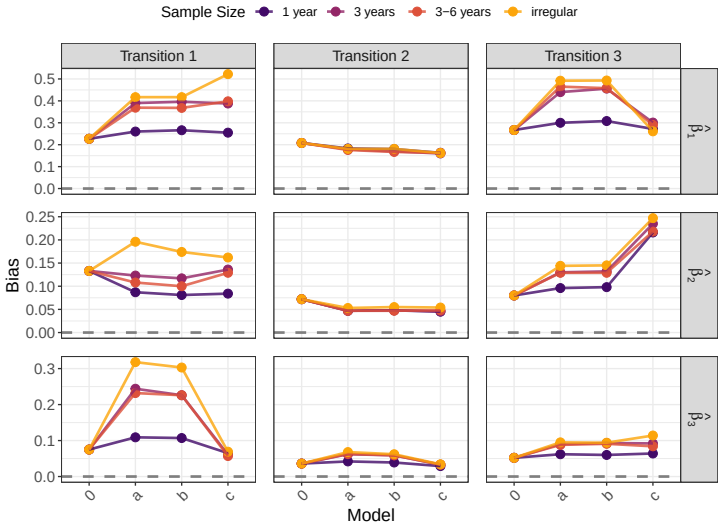


Figure B.2: Absolute bias of covariate coefficient estimates for different models across different observation schemes over 5000 individuals

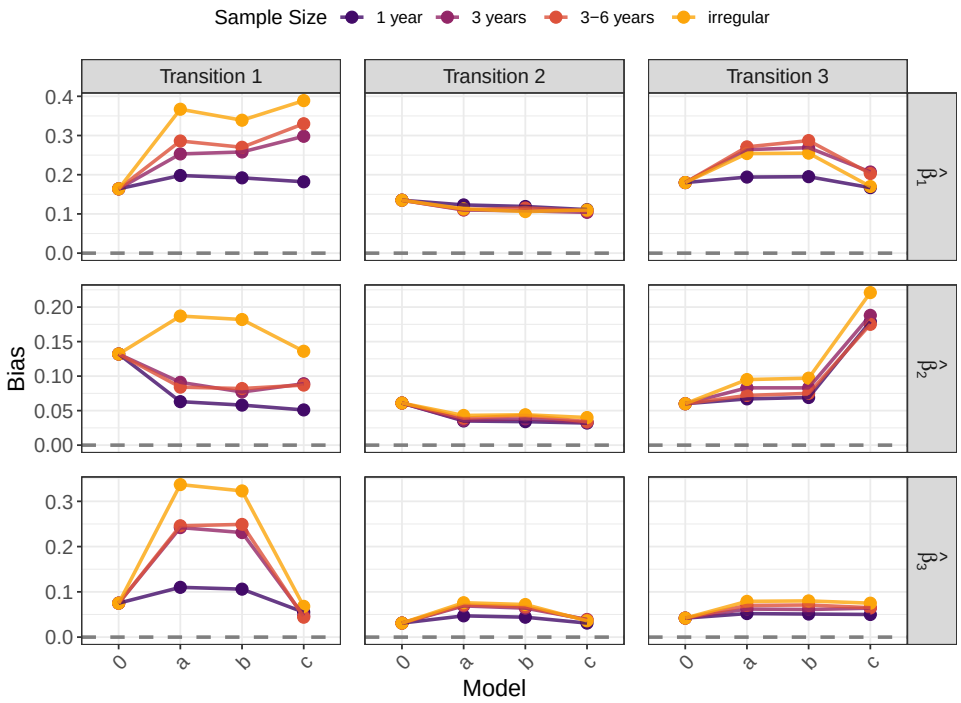


Figure B.3: Absolute bias of covariate coefficient estimates for different models across different observation schemes over 10.000 individuals

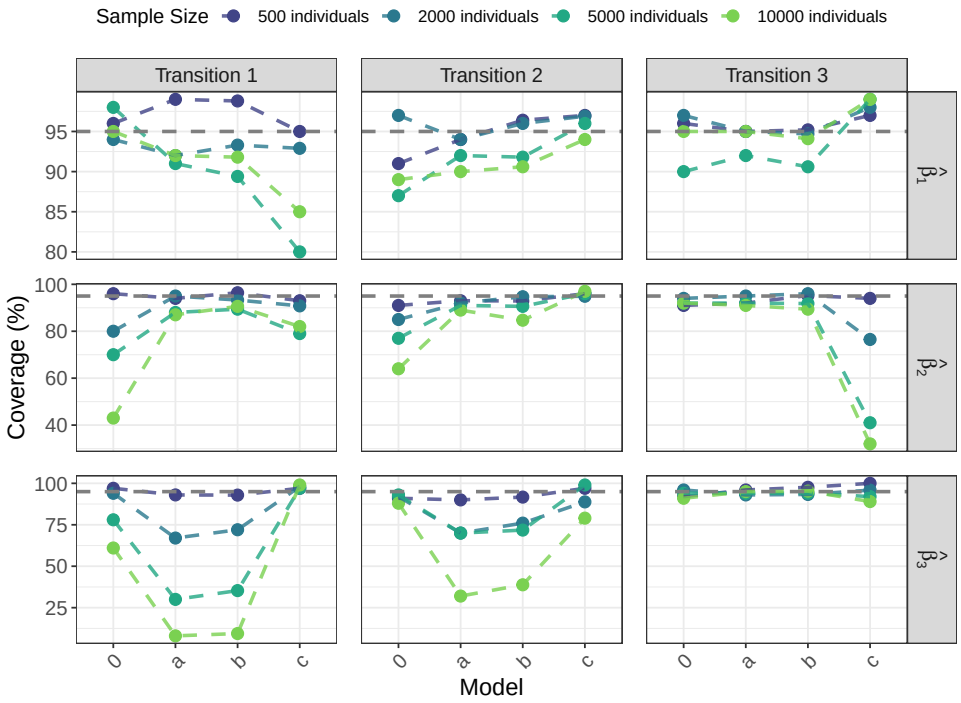


Figure B.4: Coverage of covariate coefficient estimates for different models across different sample sizes for scheme B

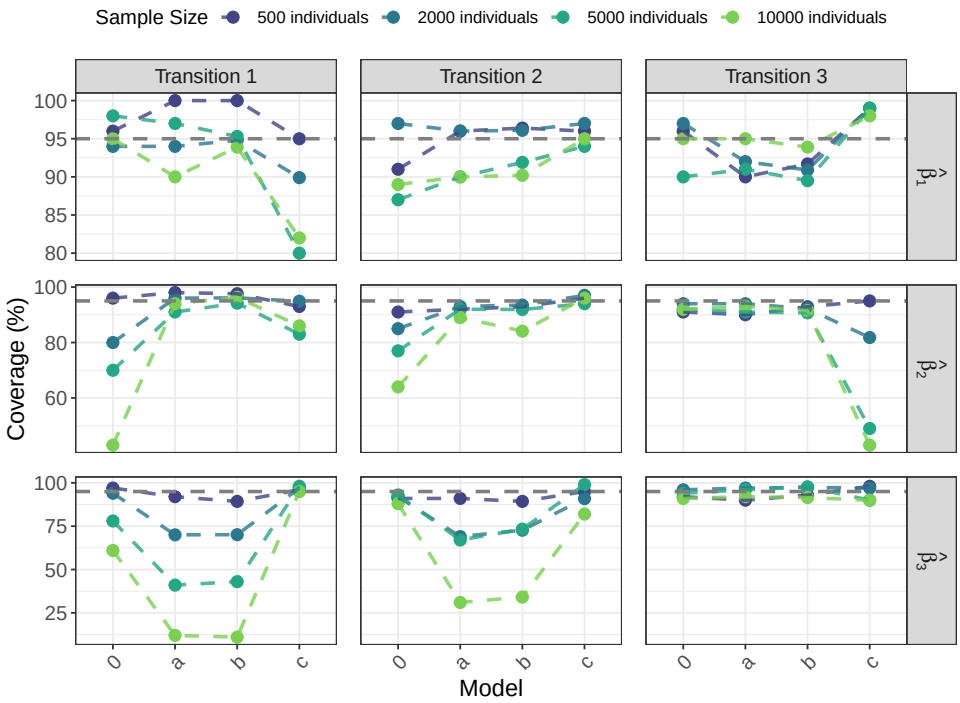


Figure B.5: Coverage of covariate coefficient estimates for different models across different sample sizes for scheme C

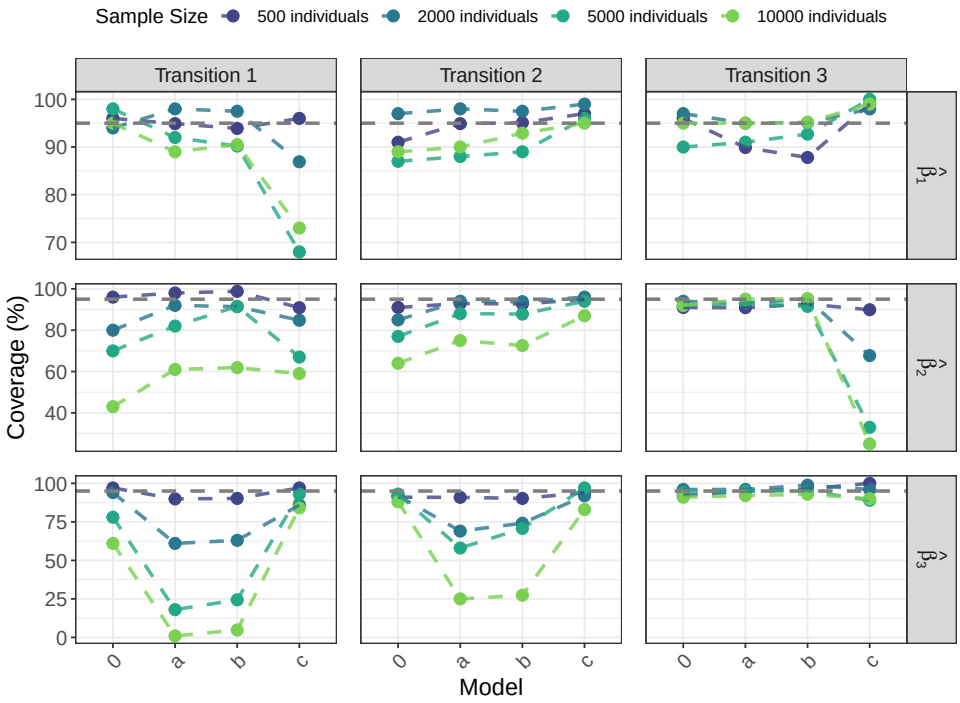


Figure B.6: Coverage of covariate coefficient estimates for different models across different sample sizes for scheme D

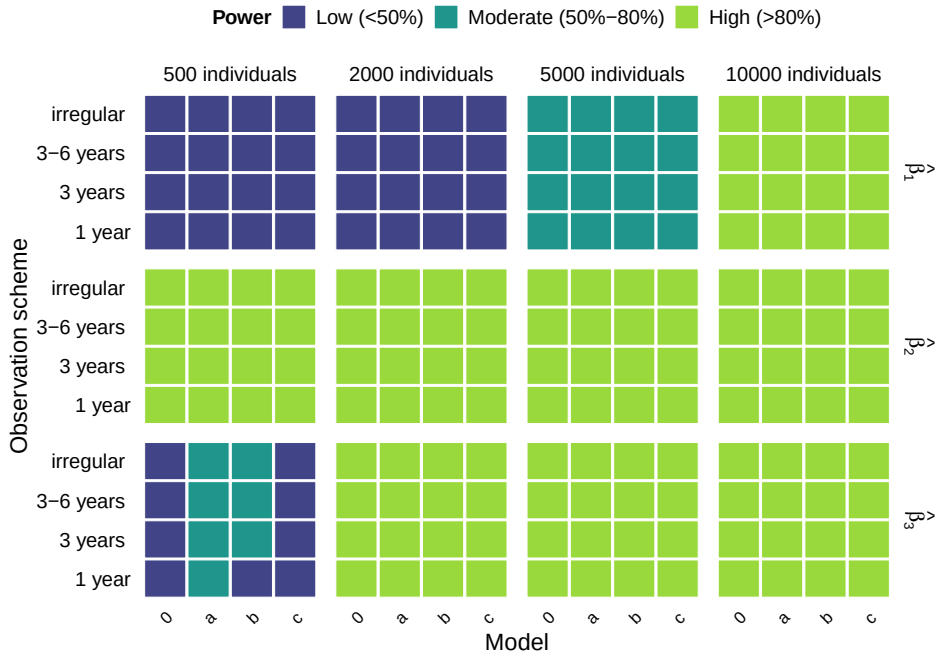


Figure B.7: Categorized Power of models across different observation schemes and sample sizes for transition from Dementia-free state to death

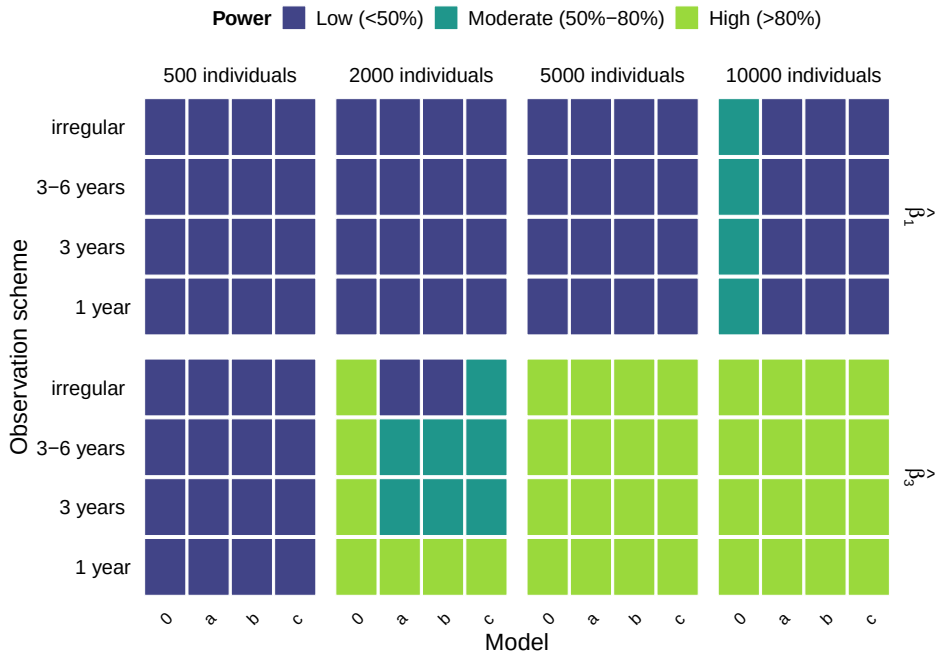


Figure B.8: Categorized Power of models across different observation schemes and sample sizes for transition from Dementia state to death

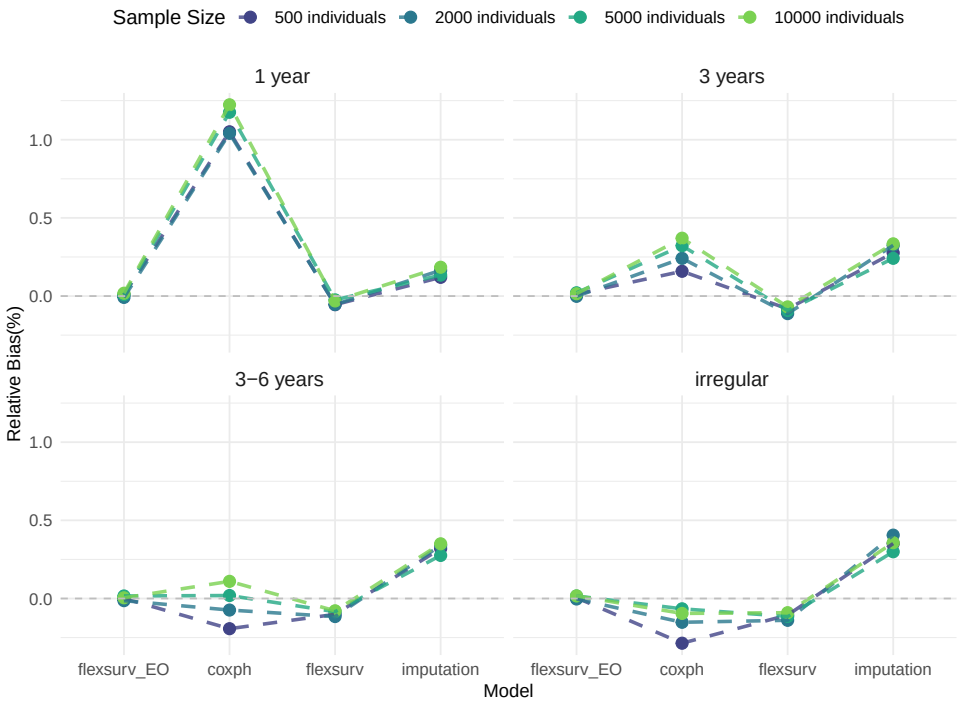


Figure B.9: Relative Bias of the total length of time spent in Dementia-free state computed with different models across different observational schemes and sample sizes

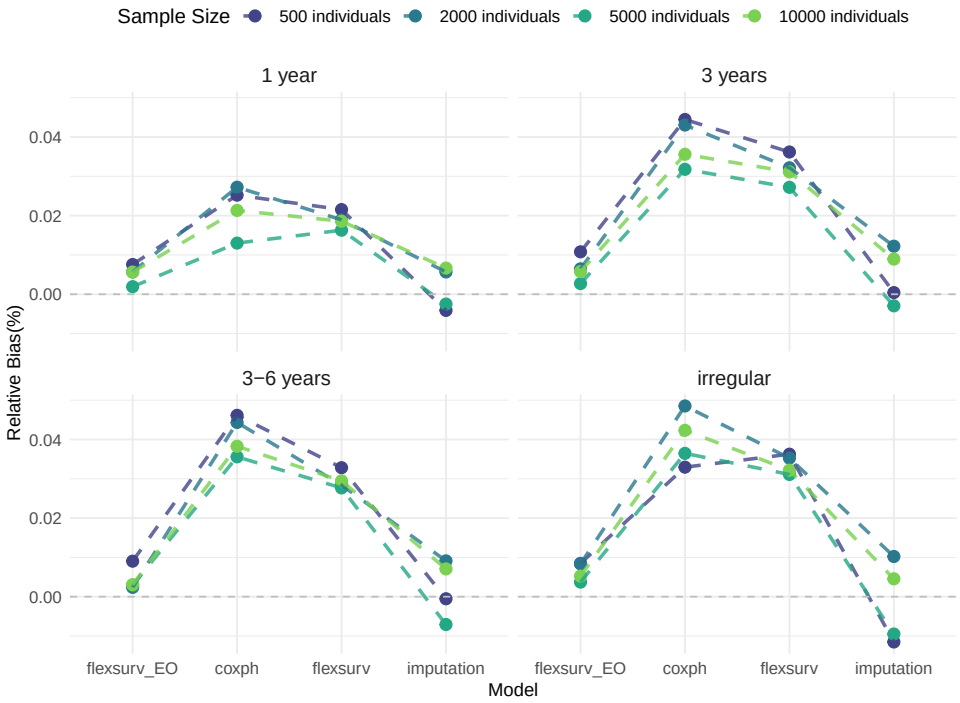


Figure B.10: Relative Bias of the total length of time spent in Dementia state computed with different models across different observational schemes and sample sizes

List of Figures

List of Tables

Acknowledgements

Desidero esprimere la mia sincera gratitudine alla professoressa Francesca Ieva per aver riposto fiducia nelle mie capacità e per avermi affidato un progetto stimolante che mi ha introdotto al campo della statistica biomedica. Questo percorso non sarebbe stato lo stesso senza l'esperienza come visiting researcher presso l'Aging Research Center di Stoccolma, e sono profondamente rinconoscente per aver avuto l'opportunità di vivere quest'esperienza in modo così completo e immersivo.

Un sincero ringraziamento va anche a tutto il dipartimento ARC, con cui ho avuto il piacere di collaborare per alcuni mesi. Mi avete trasmesso l'amore per la ricerca e reso questo percorso estremamente arricchente. In particolare, Caterina, grazie per avermi sempre incoraggiata a seguire la mia strada e portare avanti le mie idee, correggendomi con attenzione e guidandomi nei momenti di incertezza.

Alla mia mamma, che ha sempre creduto in me e nel mio futuro, dandomi la forza di scegliere ogni giorno la strada per costruire il domani che desidero. Grazie per non avermi mai fatto mancare nulla.

A mio zio, non so se sia tu la ragione del mio amore per la matematica, ma sicuramente sei stato colui che me lo ha fatto scoprire e coltivare. Il tuo sostegno è sempre stato prezioso.

Ai miei nonni, per il loro amore incondizionato e per avermi insegnato il valore del sacrificio.

A Lino, la mia costante. I chilometri che ci separano sono tanti, eppure sei sempre stato presente nelle mie giornate. Mi hai insegnato a fermarmi, guardare indietro e apprezzare la strada percorsa. Condividere i traguardi con te non ha prezzo.

A Fra, la mia anima gemella. Ti sono grata immensamente per ogni lamentela che hai ascoltato.

Agli Handicapponi, voi siete casa. Il gruppo di amici che c'è sempre stato e che vorrò sempre accanto a me. Isa, Gnesu e Ross, le mie donne, grazie per avermi insegnato che non c'è niente che ci può buttare giù.

A tutti i colleghi che sono diventati molto di più. Grazie per aver reso questi cinque anni indimenticabili. Ricordo con un grande sorriso le gomitate scambiate prendendo appunti nelle aule troppo affolatte, le pause caffe, gli aperitivi in Piazza Leo e le sessioni infinite nella grigia Bovisa.

A Richi, per essere stato al mio fianco fin dal primo giorno. Per aver trasformato ogni mio problema in una sfida da affrontare insieme e per avermi aiutata a crescere, sotto ogni aspetto.

Ai miei amici dell'Erasmus, che mi hanno fatto scoprire un mondo nuovo. Wero, Ola e Chris, ricordo di avervi osservate con ammirazione, chiedendomi come faceste ad essere così intraprendenti. Da voi ho imparato che non ci sono limiti agli interessi da coltivare e che nulla insegna più di uscire dalla propria comfort zone.

Ai miei amici di Stoccolma, che sono stati la mia famiglia quando più ne avevo bisogno, senza che nemmeno me ne rendessi conto.

Chi mi conosce sa che le parole non mi mancano mai, eppure forse non le ho usate abbastanza per esprimere quanto vi sia grata. Un grazie immenso a tutti voi per aver contribuito a rendermi la persona che sono oggi.