Multilevel Modelling of Electronic Health Records

by

Alessandro Gasparini

DEPARTMENT OF HEALTH SCIENCES

UNIVERSITY OF LEICESTER

THESIS SUBMITTED FOR THE DEGREE OF

DOCTOR *of* PHILOSOPHY

2019

They said an analytics man does not have a heart. But I ran the numbers, and nothing could be farther from the truth.

«Trust the Process.»

Abstract

Multilevel Modelling of Electronic Health Records

by

Alessandro Gasparini

The use of electronic health records (EHRs) is increasingly common in applied research, providing the opportunity to answer more relevant and detailed clinical questions. Among others, assessing the quality of routine care, enabling pragmatic clinical trials, investigating temporal trends and the natural evolution of diseases. The effective use of EHRs in medical research provides several opportunities, but challenges persist.

The principal aim of this Thesis consists of investigating methodological challenges, with focus on the multilevel structure of EHRs. First, I studied shared frailty survival models for clustered survival data and the impact of model misspecification on estimates of risk and heterogeneity. Then, I investigated joint models for longitudinal and survival data and their use to account for the drop-out and observation processes in the analysis of longitudinal data. Drop-out and the timing between observations are likely not independent of the outcome of interest in the settings of EHRs, therefore violating common assumptions of traditional methods. Focussing on the observation process, I compared the joint modelling approach to other methods previously proposed in the literature via Monte Carlo simulation. Lastly, given the use of simulation methods throughout this Thesis, I introduced newly-developed software in R to aid, support, and supplement their analysis.

The results of this Thesis highlight the importance of properly modelling the baseline hazard, frailty distribution, and assessing model fit in shared frailty survival models, as clinically-relevant biases may arise otherwise. Moreover, the joint modelling approach showed superior performance and flexibility when modelling the observation process, with a consistent pattern across all simulated scenarios.

I illustrated the above-mentioned results in practice using real-world data on chronic kidney disease and intensive care medicine, emphasising once again the requirement for appropriate statistical methods that can accommodate the complexities commonly encountered in the settings of EHRs.

Acknowledgements

First and foremost, I would like to start by thanking my main supervisor, Dr Michael Crowther, for guiding me through the years of my PhD. It has been an incredible experience throughout the process, and I am glad I could learn (a lot!) from you. I would also like to thank you for always encouraging me to pursue my research interests, sharing with me some of your Stata black magic, and keeping me involved with teaching opportunities and collaborations.

I would also like to thank my second supervisor, Prof Keith Abrams, for your incredible ability to put things into perspective and always seeing the bigger picture.

To all collaborators I have had the pleasure of working with during my PhD: thank you for your input, feedback, and support - the quality of my work has increased substantially because of you.

To everyone in the Biostatistics and GenEpi research groups and within Health Sciences at large (you are too many to mention): thank you for all the serious and not-so-serious discussions, laughter, fika and sport sessions, and socials. It's been really fun!

To present and past members of 3.06: through supporting each other, silly jokes, whiteboard-worthy discussions, and even some serious conversations every now and then you always cheered me up and made the whole PhD experience so much more enjoyable. Thank you! A special mention goes to Qingning, Richard, and Tasos (in no particular order) for all the out-of-the-office fun activities and pub expeditions. Another special mention goes to Carl for *reminding* us to take a break from work from time to time and for all the high-quality banter.

To Sarwar: thank you for convincing me to come to Leicester with your contagious enthusiasm, and thank you for all the discussions about statistics, technology, sports, and life in general. I wish you all the best with your new job on the dark side!

To Elisa, thank you for being an amazing housemate. I enjoyed so much all the cooking, *one more episode* binge-watching sessions, and all the support and help to survive our PhDs. To Giorgio, Laura, Peppe, Irene, Giacomo, and Elisa too: thank you for welcoming me to Leicester and being *family away from family*. I am grateful I could meet you all and get to enjoy our time in Leicester very much, and I am stoked to see how life is coming together for all of us.

A mamma e papà, ancora una volta mi trovo a ringraziarvi: non sarei mai potuto arrivare dove sono senza di voi, grazie per avermi reso la persona che oggi sono - è tutto merito vostro! Non riuscirò mai a sdebitarmi per l'infinito supporto e amore che mi dimostrate costantemente.

Un saluto speciale a Fede, in questi ultimi anni ti ho visto crescere e sono incredibilmente orgoglioso dell'uomo che sei diventato e degli ottimi risultati che hai ottenuto. Non vedo l'ora di venire a trovarti ovunque le tue avventure ti porteranno!

A tutto il resto della famiglia: non ci vediamo spesso, ma ogni volta è un piacere. So che mi supportate ovunque io sia, e ciò aiuta tantissimo. Grazie, grazie, grazie!

Agli amici di sempre, e non devo nemmeno nominarvi: il tempo passa, ma voi non passate mai. Tra mille avventure ormai siamo diventati grandi, e nonostante la distanza e gli impegni di sempre ogni volta che ci vediamo il tempo sembra non esser mai passato. Sono incredibilmente fortunato ad avere degli amici come voi, e ve ne sono infinitamente grato.

Dulcis in fundo, to Betty. You tried to warn me against coming to Leicester, and I am glad I did not follow your *not-so-subtle* advice! You know I am not good with words, so I will keep it short: thank you for always being there no matter what, and thank you for being just you, the way you are. I am glad we get to share our lives, and I look forward to finding out what life has in store for us.

Table of Contents

List of Figures

List of Tables

1 *Introduction*

1.1 Electronic Health Records in Medical Research

The use of electronic health records (EHRs) is increasingly common in applied epidemiological research. EHRs consist of longitudinal data collected and recorded during the routine delivery of healthcare to patients and stored in a digital format. The following information is generally included, although there may be variations between countries:

- Demographics;
- Biomarkers such as blood pressure, cholesterol levels, etc.;
- Diagnostic codes, stored as part of a physician's diagnosis of a given disease or condition;
- Pharmacy prescription and utilisation data;
- Billing codes, collected e.g. for repayment purposes or resources allocation;
- Any other information arising as part of health care delivery and stored by physicians, nurses, etc.

Further to that, EHRs can be linked to other data sources such as nationwide disease-specific registries to conduct observational and clinical research at a previously unattainable scale; examples of cohorts constructed in such way are CALIBER (CArdiovascular disease research using LInked BEspoke studies and electronic health Records [1], the Stockholm CREAtinine Measurements (SCREAM) project [2], and the Clinical Practice Research Datalink (CPRD) [3]. For instance, CPRD included individual patient data from a network of GP practices across the United Kingdom and linked primary care data to - among other data sources - hospital episode statistics, outpatient data, cancer data from Public Health England, deprivation data, and so on.

The amount of EHR data being digitally generated and collected every day is growing at an unprecedented pace and presents several opportunities to enhance and transform medical practice and research. For instance, the high-resolution data with a complete longitudinal follow-up that can be created by extracting EHRs and linking with other relevant data sources is a feature that is completely unattainable with traditional observational study designs. EHRs cohorts include (potentially) millions of individuals with repeated measures over time, and do not require investigators to wait and collect data for several years before being able to answer the questions that lead to the inception of the study. Further to that, EHRs cohorts allow researchers to examine the entire medical history of patients, potentially including data across all of their lifespans. By doing so, innovative and more detailed questions can be answered: among others, how do diseases evolve over time? Are there noteworthy temporal trends in disease incidence or severity? Is there a sequence of preliminary, minor events leading to more serious adverse outcomes, and if so, can this sequence be detected, and serious outcomes avoided or accurately forecasted? Can treatments be applied earlier, and if so, would that benefit patients?

Another benefit emerging from the use of EHRs cohorts is the possibility to study interventions in real-world settings. For instance, using EHRs it would be possible to study whether treatments are safe in patients that are commonly excluded in clinical trials but end up receiving the drug in clinical practice (e.g. because of lack of guidelines). As an example, a link between medications commonly prescribed to inhibit the secretion of gastric acid (proton pump inhibitors, PPIs) and kidney failure was recently observed [4].

Rare disease can also benefit from the use of EHRs in research. The large amount of data at hand allows including a larger number of events (compared to traditional study designs) and therefore alleviates the methodological constraints of studying rare outcomes. As an example, chickenpox as a risk factor for stroke in children was investigated using UK-based EHRs cohorts [5]; the authors concluded that indeed children experiencing chickenpox were at increased risk of stroke during the subsequent 6 months.

Finally, EHRs can not only improve observational research but also help to enhance clinical trials and to enable pragmatic clinical trials [6]. EHRs can directly inform the design of clinical trials, enable the identification of optimal target populations, and allow the accurate estimation of event rates (as observed in clinical practice). Moreover, data extracted from EHRs could be linked to trial data at a fraction of the cost, increasing the generalisability of trial results to external populations.

However, data sources constructed by extracting EHRs present several challenges that need to be addressed. The challenges arising from the use of EHRs could be broadly classified in methodological, computational, technical, and privacy, data security and consent-related issues. The innovative nature of EHRs-based research yields methodological issues, as traditional methods may not be valid (or their assumptions may be violated) in the settings of EHRs. Some of these issues form the core of this Thesis, as described in more detail in Section 1.2. For instance, new observations are recorded in EHRs every time individuals are visiting their doctor or are attending specialist visits. The existence of EHRs is therefore likely correlated with the underlying disease severity: patients with more severe conditions (or showing early symptoms of a disease) tend to visit their doctor or go to the hospital more often than those with milder conditions (and no symptoms). Their worse disease status is also likely to be reflected in worse biomarker values being recorded as such visits, causing abnormal values of such biomarkers to be over-represented and normal values to be under-represented.

Computational and technical issues arise as the amount of available data often requires ad-hoc storage and access systems, and powerful high-performance cluster computers to run the most demanding analyses. With millions of observations, computationally expensive statistical methods become unfeasible to apply: for instance, methods that require numerical integration (e.g. joint models for longitudinal and survival data) are almost impossible to fit with more than a few thousand individuals, in practice. Technical issues arise as clinical conditions are often not clearly defined in EHRs, and researchers are faced with the challenge of defining which individuals have been diagnosed with a particular condition. The task of defining medical conditions from data is known as phenotyping, and it becomes extremely challenging when multiple sources of data are linked together as in the case of EHRs. Diagnostic codes may be present, but they may be primary or secondary codes; conditions may be recorded at different moments in time, and several related conditions may be covered by single codes. To make the matter worse, there are no standard methodologies to help researchers define, share, and evaluate phenotypes defined from electronic health records, although research in the area is very active; note for instance the eMERGE consortium, a network that aims to combine DNA biorepositories with electronic medical record systems for large scale, high-throughput genetic research in support of implementing genomic medicine [7]. Another technical issue is the definition of the research cohort for a given project, requiring the description of the population from which individuals are sampled, inclusion and exclusion criteria, follow-up, and handling of missing data. In the settings of EHRs, it is necessary to report as well which data sources were included, information on the quality of the administrative records, phenotyping algorithms. The reporting of all of those details can be described to some extent in published manuscripts but is often lacking despite the emergence of reporting guidelines (e.g. the RECORD statement [8]).

Finally, privacy, data security and consent-related issues. This aspect is among the most complex, as regulations and attitude towards it vary noticeably around the world. The security of data is of high priority, but access should not be restricted to the extent that data becomes of limited utility. Distributed analyses of potentially multi-national cohorts have the advantage of allowing data to remain with the individual site, but they require systems to be put in place and appropriate analysis methods. Assuring the security and privacy of patients' data is of primary interest, to preserve the trust of the public in EHRs systems. Informed consent is also problematic, as individuals are not recruited directly, and informed consent cannot be sought on an individual basis. A recent survey in the UK concluded that 91% of respondents expected to be explicitly asked for consent for their identifiable records to be accessed for health provision, research or planning, while half the respondents (49%) did not expect to be asked for consent before their de-identified records were accessed [9].

In conclusion, using EHRs for research purposes has the potential to transform medical research, with several exciting opportunities to answer new questions and apply innovative designs. However, this potential comes at the cost of new challenges that need to be meaningfully addressed. Challenges and opportunities following from EHRs-based research are further described elsewhere [10–13].

1.2 AIMS OF THE THESIS

With this Thesis, I will focus on methodological challenges encountered when analysing EHR data. In particular, I will focus on the multilevel (hierarchical) structure of EHRs and I will investigate statistical methods that can be used to analyse survival and longitudinal data in the settings of EHRs.

To illustrate examples of the multilevel structure of EHRs data and how it fits in the settings of survival and longitudinal data, consider the following examples. In routine health care, patients generally visit their primary care practice as the first line of access to health care (e.g. for more general inquiries). Every time a subject attends a visit, their medical profile is updated, and a new record is added to the database of health records; furthermore, a single individual can attend multiple visits, yielding several repeated records over time. When extracting EHRs from the primary care practices for research use, two hierarchical structures can be identified. First and foremost, individuals are nested within practices: this leads to (potentially) significant homogeneity between individuals registered at a given practice (e.g. people living in the same neighbourhood), and heterogeneity between practices. Assuming the aim of the analysis is the time until the occurrence of an event of interest, it is necessary to take into account the multilevel structure of the survival data being analysed. The second hierarchical structure arises as a consequence of the repeated measures recorded each time an individual is attending a visit at their practice: the measurements from a given individual are correlated (with substantial within-subject homogeneity), while there is heterogeneity between individuals. In this scenario, geographical nesting within e.g. primary care practices still apply.

The work described in this Thesis is two-fold. First, I am investigating how model misspecification affects survival models that take into account the hierarchical structure depicted in the first example above. Then, I focus on the second setting described above and how models for longitudinal data perform in the settings of EHRs. In particular, I am studying how the violation of common modelling assumptions - violations that are likely encountered in the settings of EHRs - affect the analysis of repeated measurements data. Further details are presented in the next Section, where I will be outlining the structure of this Thesis.

1.3 Structure of the Thesis

The Thesis is organised as follows.

First, I introduce the basics of longitudinal and survival data analysis in Chapter 2. Relevant notation that will be used throughout the whole Thesis is introduced, as well as common methodologies used to analyse this kind of data that will form the basis of methodological work described later on.

Then, in Chapter 3 I describe two motivating clinical examples that will be used to illustrate the methodological developments of this Thesis in practice.

Next, in Chapter 4 I will introduce Monte Carlo simulation studies, the rationale for their use, and details on how to plan, run, and analyse them. Monte Carlo simulations are heavily used throughout this Thesis, hence the importance of carefully planning and analysing their results. In the same Chapter I will also introduce open source software that I developed during my PhD to aid the analysis and the dissemination of results from Monte Carlo simulation studies.

Chapter 5 describes methods used to analyse multilevel survival data. Multilevel survival data arises often in EHRs, e.g. as patients are often clustered within GP practices, which are clustered within regions, and so on. This multilevel structure induces correlation between patients within the same cluster, correlation that needs to be accounted for in the analysis. I will focus on survival models with random effects (e.g. survival models with shared frailty terms), and I will illustrate the results of a thorough investigation of the impact of model misspecification in shared frailty survival models. To the best of my knowledge, this is the most extensive investigation on the topic in the current literature.

Chapter 6 introduces the joint modelling of longitudinal and survival data. EHRs often include information on survival and repeated measures of biomarkers recorded over time; joint models for longitudinal and survival data (referred to as *joint models* in short) bring those two components together, allowing them to inform each other. Joint models are also useful when modelling longitudinal data where the assumption of independence between values of the longitudinal outcome and drop-out from the study is violated; such violation is likely to happen with EHRs data and will be discussed in more detail.

Chapter 7 will further investigate the use of joint models to overcome another limitation of models for longitudinal data. Specifically, they assume independence between the value of the longitudinal outcome and the timing of each measurement. In the setting of EHRs, that assumption is likely violated as individuals often visit their doctor (and trigger new observations to be recorded) when they feel unwell; this leads to the overrepresentation of sicker individuals compared to the general population and needs to be accounted for in the analysis.

Finally, I will conclude with a summary of the Thesis and a discussion on future work in Chapter 8. In particular, I will introduce the natural extension of the joint modelling approach described in Chapter 7: a multivariate joint modelling framework that can account for both an informative observation process and informative drop-out at the same time.

2 *General Methods for the Analysis of Survival and Longitudinal Data*

2.1 OUTLINE

This Chapter will introduce the foundations of survival analysis, including defining characteristics of survival data and methods and models commonly used to analyse survival data. Analogously, I will also introduce longitudinal data, its defining characteristics, motivating examples, and methods to analyse repeated measurements data. This chapter will (1) lay foundations and define notation that will be used later on throughout the Thesis, and (2) form the basis for methodological developments of this Thesis, especially Chapters 5 and 7.

2.2 ANALYSIS OF SURVIVAL DATA

The term *survival data* denotes data that measures the time until the occurrence of an event. Examples of events that may be of interest in biomedical research are death, disease onset, disease recurrence. However, survival data is not restricted to biomedical research: for instance, time to failure of a mechanical component is survival data commonly encountered in industrial settings, and time to acceptance of a job offer for an unemployed person is survival data encountered in economics. From now onwards, I will focus on examples in biomedical settings for simplicity. In some cases, the event of interest may be the transition from a state to another, e.g. the transition from alive to dead or the transition from healthy to diseased. The branch of survival analysis that focuses on modelling the probabilities of transitioning from one state to another is named *multi-state modelling* and is outside the scope of this Thesis. Nevertheless, a simple survival setting can be considered as a simplified version of multi-state modelling.

2.2.1 Censoring and Truncation

One of the defining characteristics of survival data is that one has to wait for the event of interest to happen, hence not all study subjects may have experienced the event of interest by the end of the study. Such individuals may experience the events later on, but such knowledge is generally not available. Consequently, the outcome of survival data is generally only observed for a subset of individuals, therefore it is not possible to use standard methodology (such as ordinary linear regression) and it is necessary to use ad-hoc methods.

The fact that observations may be incomplete is denoted *censoring*, and it is illustrated in Figure 2.1. Say I have 5 study subjects recruited over 2 years, allocating a total of up to 10 years of follow-up for each study subject. Individuals are included over time; hence they start being observed at irregular times. I stop observing people after they have been in the study for 10 years: events that happen before that time point are observed (study subjects A, B, C), while individuals event-free at the end follow-up (D, E) are denoted as censored. In particular, this kind of censoring is denoted as *right censoring*, as survival times are cut off on the right side: the true, unobserved survival time is equal to or greater than the observed time. Beside right censoring, survival times can be *left censored* or *interval censored.* Left censoring occurs when the true survival time is less than or equal to the observed survival time, for instance when a subject tests positive for a disease: in other words, if a person is left-censored at time t, it is known that they had an event between time 0 and t, but the exact time of event is unknown. Interval censoring occurs when the true survival time lays within a known time interval: if a subject tests negative for a disease at time t_1 and positive at time t_2, I would know that the true survival time of the subject lays in the interval (t_1, t_2) but I do not know the exact time of disease onset. Further details on censoring are presented in Chapter 1 of Kleinbaum and Klein [14].

Another feature of survival data often confused with censoring is truncation. Similarly to censoring, it is possible to identify different types truncation: left, right and interval truncation; I focus on left truncation which is more common alongside right censoring. Left truncation, also referred to as *delayed entry*, arises when subjects have been at risk

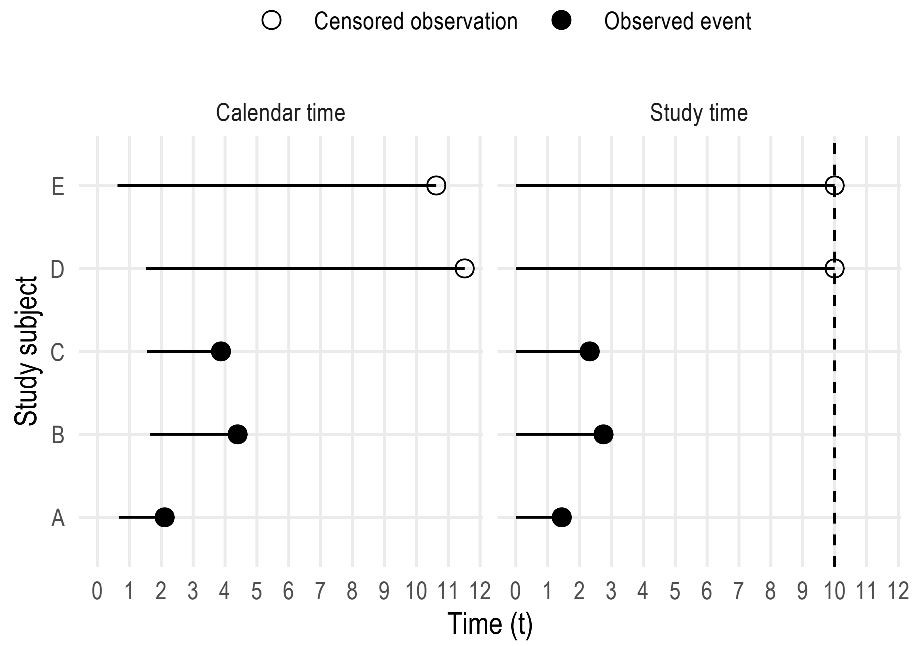


FIGURE 2.1: Example of survival data for 5 subjects; the left panel depicts the survival data on a calendar timescale, while the right panel depicts the data on a study timescale

before entering the study. For instance, assuming that subjects enter a study at time $t_0 > 0$, left truncated subjects are not observed between time 0 and t_0.

It is possible to identify two types of left truncation:

1. The first type of left truncation occurs when the subject has the event before t_0 and thus is not included in the study;
2. The second type of left truncation occurs when the subject survives (loosely speaking) beyond t_0 and is therefore observed in the study.

For instance, delayed entry is common when using age as the time scale in epidemiological studies, as a way of controlling for age rather than adjusting for it at baseline [15, 16]: in this setting, patients become at risk at the age that they e.g. are diagnosed with a given disease rather than when they are first observed. As a consequence of that, analysis methods need to account for delayed entry by conditioning on survival up to entry time. This topic is further discussed elsewhere [14].

If considering the five individuals depicted in Figure 2.1, it is immediately obvious that ad hoc statistical methods are required to analyse this kind of data. The true survival time of subjects D and E is unknown; therefore, it is not possible to calculate the mean, standard error, or apply any statistical test. In Section 2.2.3 I will introduce non-parametric methods that can be applied to survival data, while in Section 2.2.4 I will introduce regression models commonly fitted to survival data.

2.2.2 *Notation*

I denote the random variable for an individual's survival time with T^*; it denotes time and can, therefore, assume any non-negative value. Lower-case t^* represents a realisation of T^* for a given individual. In the case of right censoring, I denote with C the random variable representing censoring time, and c its realisation. The observed time is denoted with $T = \min(T^*, C)$, and its realisation is t. Finally, I denote with $D = I(T^* \leq C)$ the random variable indicating either occurrence of the event of interest or censoring; analogously as before, its realisation is lower-case d.

Next, I will define two of the main quantities of interest in survival analysis, the survival function and the hazard function. They are both functions of the observed time t and are denoted by $S(t)$ and $h(t)$, respectively.

The survival function is the complement of the cumulative distribution function of the observed time T and represents the probability that a given individual survives (loosely speaking) longer than t:

$$S(t) = 1 - F_T(t) = 1 - P(T \leq t) = P(T > t) \tag{2.1}$$

t ranges (theoretically) between zero and infinity, hence the survival function can be plotted as a smooth, continuous function that tends to zero as t goes to infinity. An example survival function is plotted in Figure 2.2, panel A.

The hazard function $h(t)$ is the limit of the probability of the survival time T laying within an interval $[t, t + \Delta_t)$ given that an individual survived up to time t divided by the length of the interval Δ_t, for Δ_t approaching zero:

$$h(t) = \lim_{\Delta_t \longrightarrow 0} \frac{P(t \leq T < t + \Delta_t | T \geq t)}{\Delta_t} \tag{2.2}$$

It represents the instantaneous potential (e.g. risk) for the event to occur within the interval $[t, t + \Delta_t)$ (with $\Delta_t \longrightarrow 0$), given that the individual survived up to time t. The hazard function is always non-negative, it can assume different shapes over time, and it

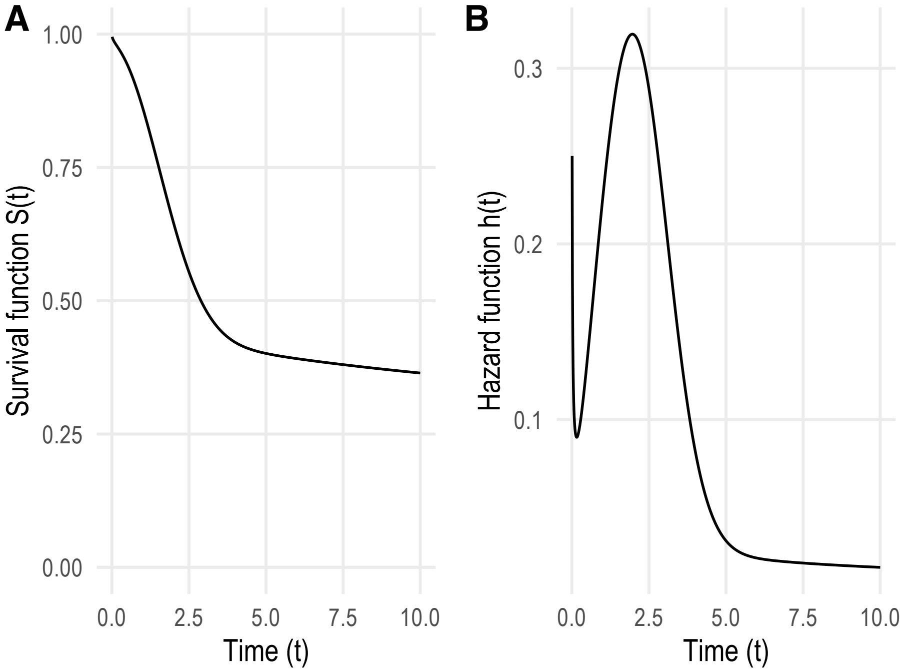


FIGURE 2.2: Example of survival (panel A) and hazard (panel B) function.

has no upper bound. An example hazard function is plotted in Figure 2.2, panel B.

The survival function and the hazard function are strictly related. In fact, there is a clearly defined mathematical relationship between them, and it is possible to derive the form of $S(t)$ when knowing the form of $h(t)$ (and vice versa). Formally:

$$S(t) = \exp\left[-\int_0^t h(u)\ du\right] \tag{2.3}$$

$$h(t) = -\left[\frac{\partial S(t)/\partial t}{S(t)}\right] \tag{2.4}$$

A third quantity of interest strictly related to the survival and hazard functions is the cumulative hazard function $H(t)$. The cumulative hazard function represents the accumulation of hazard $h(t)$ over time, and is defined as

$$H(t) = \int_0^t h(u)\ du; \tag{2.5}$$

it can be expressed in terms of survival function, via the relationships $H(t) = -\log S(t)$ or alternatively $S(t) = \exp(-H(t))$.

The notation presented in this Chapter follows Collett [17], where further details can be found. The hazard function $h(t)$ and the survival function $S(t)$ form building blocks for the methods that I will be presenting in the following Sections and Chapters.

2.2.3 *Non-Parametric Methods*

The survival function can be estimated from survival data using a non-parametric method known as the Kaplan-Meier product-limit method [18]. Let t_i be the distinct failure times, n_{t_i} the number of individuals at risk before time t_i, and d_{t_i} the number of events observed at time t_i. The Kaplan-Meier non-parametric estimate of the survival function is then:

$$\hat{S}(t) = \prod_{i|t_i \leq t} \left(\frac{n_i - d_i}{n_i} \right) \tag{2.6}$$

In practice, the Kaplan-Meier estimate of the survival function appears as a step function as individuals can be observed at discrete times only and not all individuals may experience the event before the end of the study.

It is possible to obtain a standard error for the estimate above using the Greenwood formula [19]:

$$\widehat{\text{Var}}(\hat{S}(t)) = \hat{S}^2(t) \sum_{i|t_i \leq t} \frac{d_j}{n_j(n_j - d_j)} \tag{2.7}$$

Another analysis often of interest when dealing with survival data consists of comparing the survival times obtained from two (or more) groups of individuals. Say for instance that I am interested in comparing the survival function of two distinct groups, e.g. treated and non-treated individuals. It is possible to plot the Kaplan-Meier estimate of the survival function for each group as in Figure 2.3. The two functions largely overlap, showing that the survival function is similar between the two groups.

However, the next step requires using some hypothesis testing procedure to formally assess whether the two survival functions are statistically significantly different. An appropriate test to compare two groups of survival data is the so-called log-rank test [20].

Consider the two groups depicted in Figure 2.3; say there are r distinct death times, denoted as $t_{(1)} < t_{(2)} < \cdots < t_{(r)}$ across the two groups, and that at time t_j there are d_{1j} deaths in group 1 (e.g. treated) and d_{2j} in group 2 (e.g. untreated), for $j = 1, 2, \ldots, r$. Say there are n_{1j} and n_{2j} individuals at risk just before time $t_{(j)}$ for group 1 and 2, respectively.

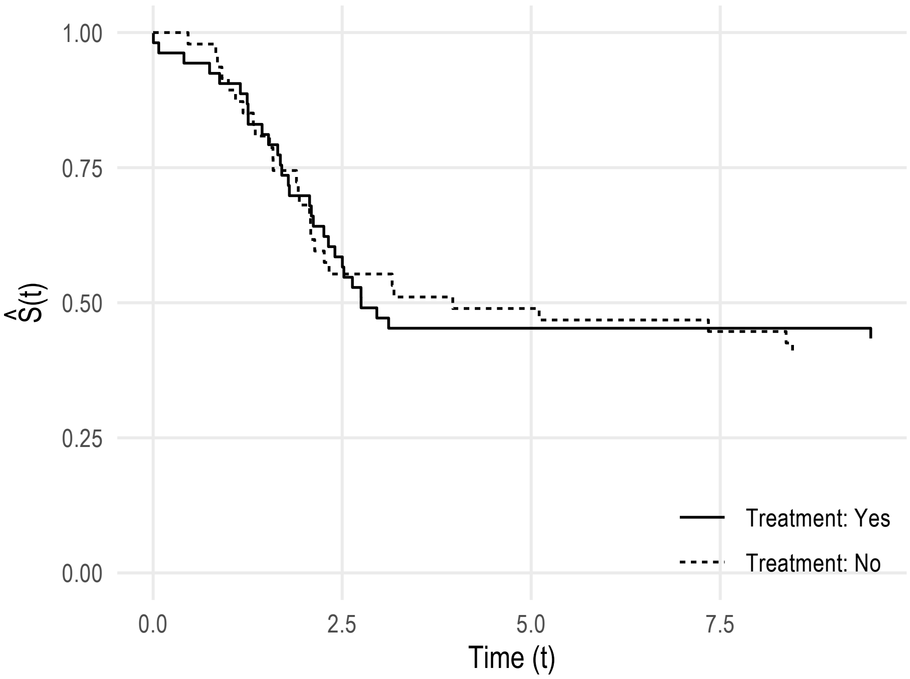


FIGURE 2.3: Kaplan-Meier estimate of the survival function for two study groups, e.g. treated and untreated individuals.

Consequently, there are $d_j = d_{1j} + d_{2j}$ events out of $n_j = n_{1j} + n_{2j}$ individuals at risk at time $t_{(j)}$. Consider the null hypothesis of no difference between the two survival functions; a way of assessing the validity of this hypothesis consists in comparing the observed number of events at each distinct events time in each group against the expected number of events under the null hypothesis. The expected number of events in group 1 at time $t_{(j)}$ under the null hypothesis can be computed as

$$e_{1j} = \frac{n_{1j} d_j}{n_j}, \tag{2.8}$$

and analogously for group 2.

We can then construct the following test statistic:

$$W_L = \frac{U_L^2}{V_L} \sim \chi_1^2, \tag{2.9}$$

where

$$U_L = \sum_{j=1}^{r} (d_{1j} - e_{1j}) \tag{2.10}$$

and

$$V_L = \mathrm{Var}(U_L) = \sum_{j=1}^{r} \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}. \tag{2.11}$$

The test statistic W_L follows a χ^2 distribution with one degree of freedom and summarises the extent to which the observed survival times in the two groups deviate from the expected number of events under the null hypothesis.

In the applied example of Figure 2.3, the observed test statistics is $W = 0.03$. By comparing the observed value of the test statistics with the quantiles of a χ^2_1 distribution I obtain a p-value of 0.86, showing that there is not enough empirical evidence against the null hypothesis of no difference between the two groups.

2.2.4 *Regression Models for Survival Data*

The non-parametric methods presented in Section 2.2.3 can be useful for simple analyses of a single group of individuals, or to compare the survival function of two or more groups. However, more complex scenarios often arise. For instance, several characteristics of the study subjects are routinely recorded, and applying methods such as the log-rank test in those settings becomes increasingly complex. Including several covariates at once becomes unfeasible, as does including continuous covariates without categorising them. To study the relationship between a set of observed covariates and a survival outcome, several approaches based on regression modelling are routinely used.

Regression models used to estimate the effect of covariates on survival times can be broadly classified into two families: accelerated failure time (AFT) and proportional hazards (PH) models.

Using the accelerated failure time notation, the logarithm of time t is expressed as a linear combination of the covariates, yielding a linear model:

$$\log(t_i) = x_i \beta + \varepsilon_i, \tag{2.12}$$

where x_i is a vector of covariates for the i^{th} individual, β is a vector of regression coefficients, and ε_i is a residual component with a given density function. The density of the residual component ε_i determines the model.

Conversely, assuming a proportional hazards model, the covariates affect the hazard function multiplicatively:

$$h_i(t) = h_0(t)\exp(x_i\beta), \tag{2.13}$$

where $h_0(t)$ is a baseline hazard function that can be left unspecified or take a given parametric form. Specifically, the baseline hazard function represents the hazard when all covariates are set to zero (or to the reference level in the case of categorical covariates). Proportional hazards models where the baseline hazard follows a parametric distribution are presented in Section 2.2.4.2, while survival models where the baseline hazard function is left unspecified are presented in Section 2.2.4.1. In addition to that, I will introduce survival models where the baseline hazard is modelled flexibly in Section 2.2.4.3. Finally, I will assume the proportional hazards formulation throughout this Thesis.

2.2.4.1 *Semi-Parametric Survival Models*

The most commonly used regression model for survival data in applied epidemiological research is the Cox model [21], also known as the Cox proportional hazards model:

$$h_i(t) = h_0(t)\exp(x_i\beta) \tag{2.14}$$

As illustrated in the previous Section, this model is a proportional hazards model, analogous to Equation (2.13). Assuming a single binary covariate, the model from Equation (2.14) becomes:

$$h_i(t) = \begin{cases} h_0(t)\exp(\beta) & \text{if } x_i = 1 \\ h_0(t) & \text{if } x_i = 0 \end{cases}$$

Comparing the two groups by taking the ratio of the hazard from the two groups, the resulting hazard ratio (HR) is:

$$HR = \frac{h_0(t)\exp(\beta)}{h_0(t)} = \exp(\beta)$$

Consequently, the regression coefficient β can be interpreted as a log-hazard ratio, that is, the effect of covariates x_i on the hazard; this interpretation generalises to multiple covariates and associated regression coefficients.

Fitting a regression model to the observed survival data requires estimating the regression coefficients β and the baseline hazard function $h_0(t)$. Interestingly, Cox [22] showed that it is possible to make inference on β without needing to estimate the baseline hazard function: the Cox model is therefore regarded as *semi-parametric*, as it is not necessary to specify a parametric form of the baseline hazard to make inference on the regression coefficients.

Estimation of the regression coefficients β is via the maximum likelihood method. Say there are n study subjects that experience r events (with $n - r$ censored observations). Say the r ordered event times are $t_{(1)} < t_{(2)} < \cdots < t_{(r)}$, and there are no ties. The group of individuals at risk at time $t_{(j)}$ are denoted with $R(t_{(j)})$, that is, $R(t_{(j)})$ represents the number of individuals event-free immediately before $t_{(j)}$. Cox showed that the relevant likelihood function for the model above is:

$$L(\beta) = \prod_{j=1}^{r} \frac{\exp(x_{(j)}\beta)}{\sum_{l \in R(t_{(j)})} \exp(x_l \beta)}, \tag{2.15}$$

where $x_{(j)}$ is the covariates vector for the individual experiencing the event at $t_{(j)}$. The likelihood function in Equation (2.15) is not a true likelihood as it does not use of all the observed survival times, censored and uncensored; it depends on the ranking of the observed event times only, and for this reason, it is often referred to as *partial likelihood*.

Supposing the observed data consists of n survival times $(t_1, t_2, \dots, t_n)$ paired with a binary indicator variable d_i that takes the value zero when the i^{th} survival time is censored, one otherwise. In this setting the partial likelihood function of Equation (2.15) can be re-written as follows:

$$\text{partial } L(\beta) = \prod_{i=1}^{n} \left[\frac{\exp(x_i \beta)}{\sum_{l \in R(t_i)} \exp(x_l \beta)} \right]^{d_i}, \tag{2.16}$$

where $R(t_i)$ is the risk set at time t_i. The corresponding partial log-likelihood is:

$$\log \text{ partial } L(\beta) = \sum_{i=1}^{n} d_i \left[x_i \beta - \log \sum_{l \in R(t_i)} \exp(x_l \beta) \right], \tag{2.17}$$

which can be directly maximised using any general-purpose optimiser.

So far I illustrated how to estimate the regression coefficients of a Cox model, i.e. measures that quantify relative risk. However, sometimes it is necessary to estimate the baseline hazard function $h_0(t)$ as well: say for instance that the aim of the analysis requires estimating the hazard or the survival function. Let $\hat{\beta}$ be the estimated regression coefficient from a Cox model; the estimated hazard function for the i^{th} individual is:

$$\hat{h}_i(t) = \hat{h}_0(t) \exp(x_i\hat{\beta}), \tag{2.18}$$

requiring an estimation of the baseline hazard function $h_0(t)$.

An estimate of the baseline hazard function based on the maximum likelihood method was derived by Kalbfleisch and Prentice [23]. Their method requires an iterative scheme; hence an approximation of the baseline hazard function is often used. An estimate of the survival function is given by

$$\tilde{S}_0(t) = \prod_{j=1}^{k} \exp\left(\frac{-d_j}{\sum_{l \in R(t_{(j)})} \exp(x_l\hat{\beta})}\right), \tag{2.19}$$

for $t_{(k)} \leq t < t_{(k+1)}$, $k = 1, 2, \ldots, r-1$. When there are no covariates in the model, the estimator of Equation (2.19) simplifies to

$$\tilde{S}(t) = \prod_{j=1}^{k} \exp(-d_j/n_j), \tag{2.20}$$

which is also known as the Nelson-Aalen estimator of the survival function, a competing estimator to the Kaplan-Meier estimator from Equation (2.6).

From $\tilde{S}_0(t)$ it is possible to obtain an estimate for the cumulative baseline hazard function $\hat{H}_0(t)$:

$$\hat{H}_0(t) = -\log(\tilde{S}_0(t)) = \sum_{j=1}^{k} \left(\frac{d_j}{\sum_{l \in R(t_{(j)})} \exp(x_l\hat{\beta})}\right). \tag{2.21}$$

This estimate is known as the Breslow estimate (or Nelson-Aalen estimate) of the cumulative baseline hazard function.

The estimated baseline hazard function follows as

$$\hat{h}_0(t) = \frac{d_j}{(t_{(j+1)} - t_{(j)}) \sum_{l \in R(t_{(j)})} \exp(x_l\hat{\beta})} \tag{2.22}$$

for the interval between $t_{(j)}$ and $t_{(j+1)}$. It is possible to see that differences in successive values of the estimated cumulative baseline hazard function $\hat{H}_0(t)$ provide an approximation of the baseline hazard function in a given interval.

Further details on the Cox model, including how to compare alternative models, model selection, interpretation of model parameters, and others are included in Chapter 3 of Collett [17]. Methods to assess the fit of a Cox model and test the assumption of proportional hazards are included in Chapter 4 of Collett [17].

2.2.4.2 *Parametric Survival Models*

Proportional hazards models where the baseline hazard function is assumed to follow a parametric distribution are commonly referred to as *parametric survival models*. With the Cox model introduced in Section 2.2.4.1 it is straightforward to estimate measures of relative risk without the need to model the baseline hazard function. However, when the assumption of a particular distribution for the baseline hazard function holds inference will be more precise. In addition to that, parametric survival models provide some advantages compared to the Cox model. Among others:

1. Parametric models provide smooth estimates of the hazard and survival function for any combination of covariates;
2. With parametric models it is possible to obtain any type of estimate as a function of the estimated model parameters;
3. With parametric models it is straightforward to include time-dependent effects;
4. Modelling on different scales and implementing multiple time-scales is supported;
5. It is easier to obtain out-of-sample predictions and extrapolation with parametric survival models compared to the Cox model.

If any of the above is not of interest, though, the Cox model retains the appealing characteristic of allowing to estimate relative risk without specifying the baseline hazard function.

Commonly used distribution functions assumed for the baseline hazard are the exponential, Weibull, and Gompertz distributions. The exponential distribution is assuming the following baseline hazard function:

$$h_0(t) = \lambda, \tag{2.23}$$

with λ a positive parameter that can be estimated from data. Exponential hazard functions with varying values of λ are depicted in Figure 2.4. The corresponding survival function is:

$$S(t) = \exp\left(-\int_0^t \lambda \, du\right) = \exp(-\lambda t) \tag{2.24}$$

The Weibull distribution is assuming the following baseline hazard function:

$$h_0(t) = \lambda \gamma t^{\gamma - 1}, \tag{2.25}$$

with λ and γ positive parameters, also estimated from data. The γ parameter determines the shape of the distribution, adapting to a wide variety of scenarios; further to that, when $\gamma = 1$ the Weibull distribution simplifies to the exponential distribution. Weibull hazard functions with varying shape parameters are depicted in Figure 2.4.

The corresponding survival function is:

$$S(t) = \exp\left(-\int_0^t \lambda \gamma u^{\gamma - 1} \, du\right) = \exp(-\lambda t^{\gamma}) \tag{2.26}$$

Finally, the Gompertz distribution is assuming the following baseline hazard function:

$$h_0(t) = \lambda \exp(\theta t), \tag{2.27}$$

with λ and θ parameters estimated from data. λ is constrained to be a positive parameter, while θ can assume any value yielding hazard functions that are decreasing, stable, or increasing. Analogously as with the Weibull distribution, by altering the θ parameter it is possible to cover a wide variety of scenarios: Gompertz hazard functions are depicted in Figure 2.4. The corresponding survival function is:

$$S(t) = \exp\left(-\int_0^t \lambda \exp(\theta u) \, du\right) = \exp\{-\lambda \theta^{-1}[\exp(\theta t) - 1]\} \tag{2.28}$$

The Gompertz model has been often applied in demography and biological sciences, as the Gompertz distribution was first introduced as a model for human mortality.

The proportional hazard parametric model follows the same formulation outlined in Equation (2.13):

$$h_i(t) = h_0(t)\exp(x_i\beta) \tag{2.29}$$

Assuming e.g. a Weibull distribution for the baseline hazard, the model from Equation (2.29) can be written as:

$$h_i(t) = \lambda\gamma t^{\gamma-1}\exp(x_i\beta) \tag{2.30}$$

The corresponding survival function is:

$$S_i(t) = \exp[-\exp(x_i\beta)\lambda t^{\gamma}] \tag{2.31}$$

Analogous equations can be derived for the remaining distributions as well.

Parametric survival models can be fitted using the maximum likelihood method. The likelihood function (assuming right censoring only, and no delayed entry) can be written as:

$$L(\beta) = \prod_{i=1}^{n} h(t_i)^{d_i} S(t_i) \tag{2.32}$$

The log-likelihood follows as

$$\log L(\beta) = \sum_{i=1}^{n} \left[d_i \log h(t_i) + \log S(t_i)\right], \tag{2.33}$$

and can be maximised using e.g. the Newton-Raphson method.

2.2.4.3 Flexible Parametric Survival Models

I outlined in Section 2.2.4.2 the advantages of parametric regression models compared to the Cox model. However, simple parametric functions (such as those introduced in Section 2.2.4.2) may not be flexible enough to adequately represent the hazard function. For instance, standard parametric functions are monotonic: the hazard function is stable, always increasing or always decreasing. Many real-life datasets have hazards that peak after some time and then decrease: in all of these scenarios, using a simple parametric distribution would not fit the data well.

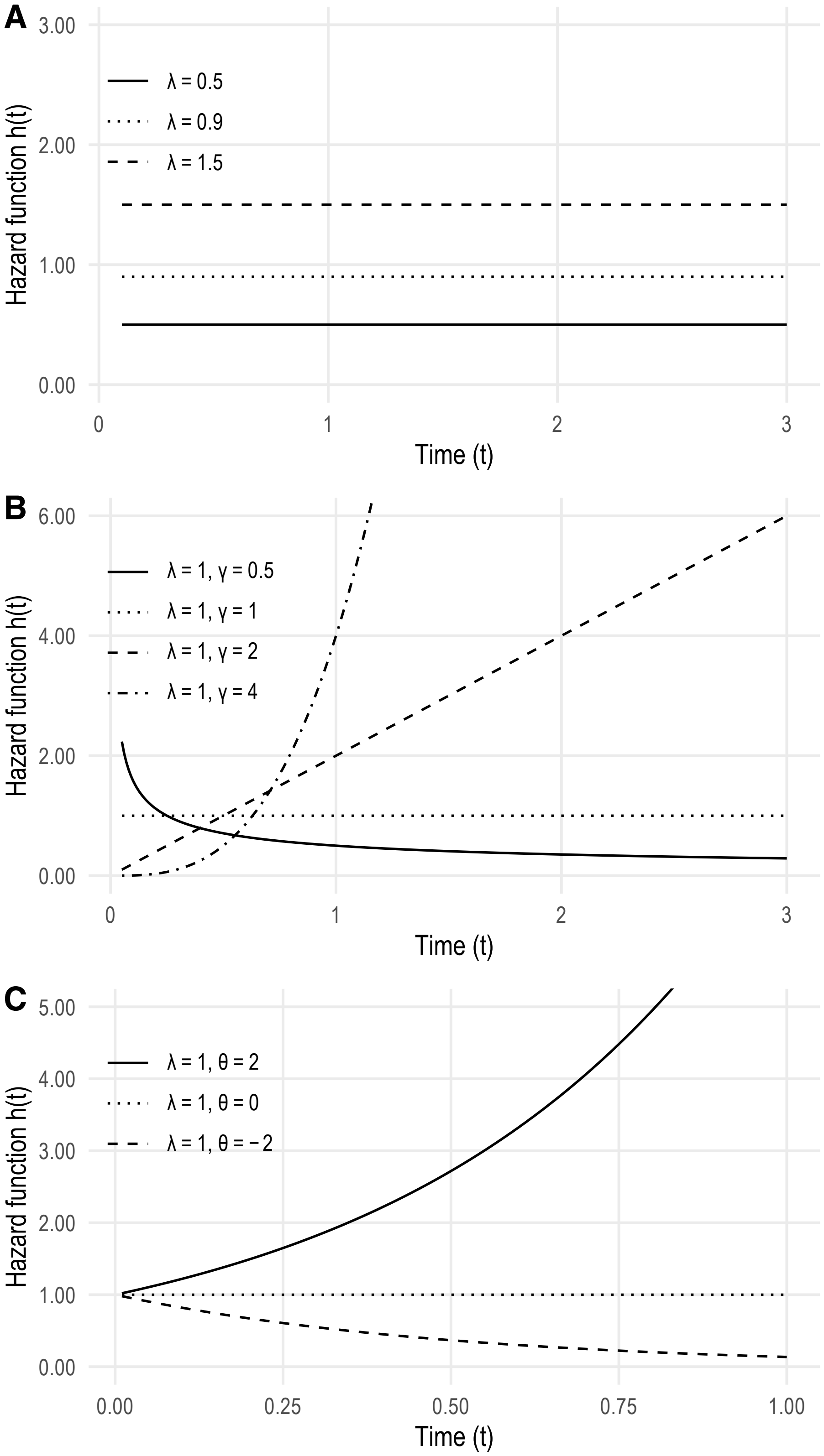


FIGURE 2.4: Hazard functions following an exponential distribution with given λ parameters (panel A), a Weibull distribution with given λ and γ parameters (panel B), or a Gompertz distribution with given λ and θ parameters (panel C)

An alternative to parametric regression models for survival data is given by *Royston-Parmar* (RP) models, also known as *flexible parametric models* [24]. RP models have greater flexibility with respect to the shapes of the survival distributions they can model, as the baseline hazard is modelled using restricted cubic splines. In particular, RP models are formulated by modelling the survival times on the cumulative hazard scale:

$$\log H_i(t) = \log H_0(t) + x_i\beta$$

The baseline log cumulative hazard function can be modelled by using a restricted cubic spline function of log-time, $\log H_0(t) = s(\log t; \gamma)$:

$$\log H_i(t) = \log H_0(t) + x_i\beta = s(\log t, \gamma) + x_i\beta \tag{2.34}$$

It can be shown that the model from Equation (2.34) is also a proportional hazards model. Transforming to the survival scale:

$$S_i(t) = \exp\{-\exp[s(\log t, \gamma) + x_i\beta]\},$$

and consequently to the hazard scale:

$$h_i(t) = \frac{\partial s(\log t, \gamma)}{\partial t} \exp x_i\beta$$

However, this involves calculating the derivatives of the restricted cubic spline functions, which are easy to calculate obtaining closed-form formulæ.

A restricted cubic spline is a cubic spline function that is restricted to be linear before the first knot and after the last knot. Let $s(x)$ be a restricted cubic spline function of x with m internal knots $(k_1, \dots, k_m)$, with boundary knots $k_{\min}$ and $k_{\max}$. Let γ be the parameters of the spline function, and $z_1, \dots, z_{m+1}$ be newly created variable denoted as basis functions of the spline. Formally, the restricted cubic spline function of x is defined as

$$s(x) = \gamma_0 + \gamma_1 z_1 + \cdots + \gamma_{m+1} z_{m+1}, \tag{2.35}$$

where

$$z_1 = x,$$

Table 2.1: Number and location of internal knots for the spline modelling the cumulative baseline hazard function in Royston-Parmar models

Knots	df	Centiles
1	2	50
2	3	33, 67
3	4	25, 50, 75
4	5	20, 40, 60, 80
5	6	17, 33, 50, 67, 83
6	7	14, 29, 43, 57, 71, 86
7	8	12.5, 25, 37.5, 50, 62.5, 75, 87.5
8	9	11.1, 22.2, 33.3, 44.4, 55.6, 66.7, 77.8, 88.9
9	10	10, 20, 30, 40, 50, 60, 70, 80, 90

$$z_j = (x - k_j)_+^3 - \lambda_j (x - k_{\min})_+^3 - (1 - \lambda_j)(x - k_{\max})_+^3,$$

and

$$\lambda_j = \frac{k_{\max} - k_j}{k_{\max} - k_{\min}}$$

When fitting a RP model, it is necessary to choose the number of internal knots and the location of the boundary and internal knots. A sensible choice for the boundary knots $k_{\min}$, $k_{\max}$ is the smallest and largest uncensored survival time. Conversely, for the number and location of internal knots, Royston and Parmar suggested knot positions based on empirical centiles of the distribution of log-time (Table 2.1); these locations are essentially those recommended by Durrleman and Simon for flexible regression models with restricted cubic splines [25].

In practice, the number of knots can be selected empirically by fitting several RP models with varying number of degrees of freedom and then picking the best fitting model using information criteria such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) [26, 27], together with evaluating the behaviour of the model in terms of stability of estimates and convergence properties. However, some authors suggest that estimates of relative and absolute risk obtained by fitting RP models are insensitive to the number of degrees of freedom used for the spline function [28, 29].

Further details on RP and flexible parametric models, including alternative parametrisations (e.g. proportional odds) and practical applications, are presented in Royston and Lambert [30].

2.3 Analysis of Longitudinal Data

The defining characteristic of longitudinal data is that it consists of repeated measurements over time for each study subject. In univariate statistics, each study subject returns a single value for the outcome of interest. Conversely, with longitudinal data, each individual yields a vector of values collected over time. Therefore, with longitudinal data:

1. Each study subject yields a vector of observations;
2. The repeated measurements are collected over time.

Example of longitudinal measurements collected over time are body weight, blood pressure, kidney function, and so on. An advantage of longitudinal studies compared to cross-sectional studies is that it allows distinguishing changes over time within individuals from differences between individuals at baseline. The ability to partition the variation in the outcome in within-individual variation and between-individuals variation is one of the advantages of analysing longitudinal data.

Longitudinal data can be collected prospectively, following subjects over time, or retrospectively, e.g. by extracting historical records. Clinical trials are prospective studies where longitudinal data is collected prospectively, while the analysis of electronic health records is an example of studies where longitudinal data is collected retrospectively. In addition to that, longitudinal data can be collected using a pre-defined observation pattern (e.g. in clinical trials where measurements are taken at pre-defined follow-up times) or as observed (e.g. when extracting EHRs).

The simplest example of a longitudinal study consists of a study with two repeated measurements. Say there are 10 children and their weight is recorded at two time points, as depicted in Figure 2.5. Analysing this longitudinal data it is possible to appreciate the change over time for each study individual (depicted in black) and for the overall population (in grey): although the body weight increases over time for the overall

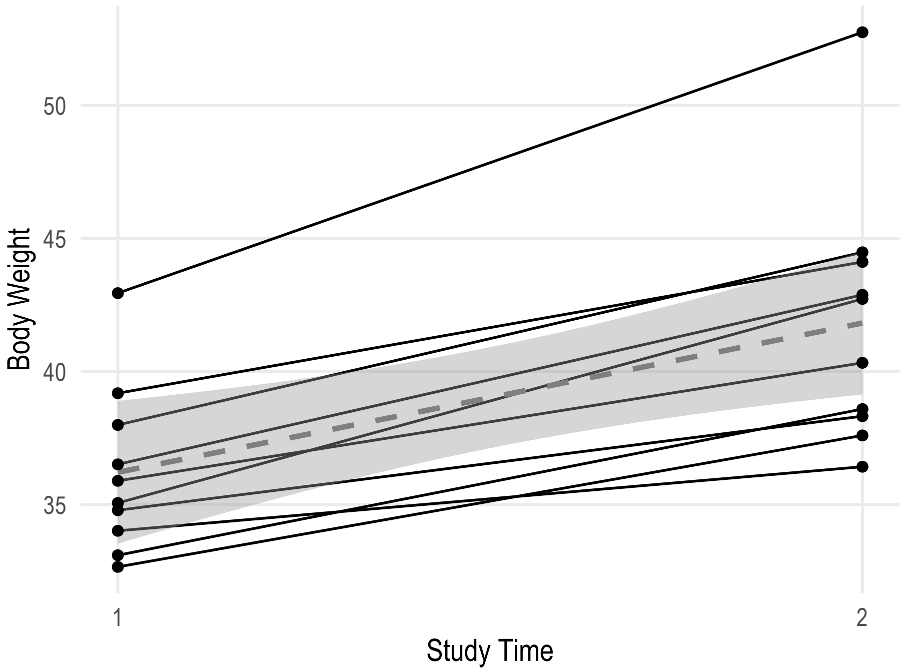


FIGURE 2.5: Example of a longitudinal study with 10 individuals and 2 time points, assuming repeated measurement of body weight. Individual values are depicted in black, with a super-imposed regression line in grey

population, each individual varies in a subject-specific way. This would not be possible by analysing the body weights measured at each time point using a cross-sectional design.

With the previous example, I assumed that every study individual was measured at the same point in time. As I mentioned before, this assumption may not always hold e.g. in analysing EHRs. Generalising the previous example, say a study consists of extracting the EHRs of 10 children born in 2010. Such study is presented in Figure 2.6 using a cohort (i.e. study) time scale: individuals are enrolled in the study when they are first observed, and all consequent measurements are relative to that index date. It is possible to present the same study using a different time scale: the age time scale (Figure 2.6). With the age time scale, individuals are still enrolled when first observed but by plotting each measurement against the age at which it was taken the comparison is fairer, as the observation time does not depend any more on when individuals are first observed.

Finally, the analysis of longitudinal data requires methods that can properly account for the within-subject correlation as the measurements are collected over time for each individual. If the correlation between measurements belonging to the same individual is ignored, then inference can be invalid.

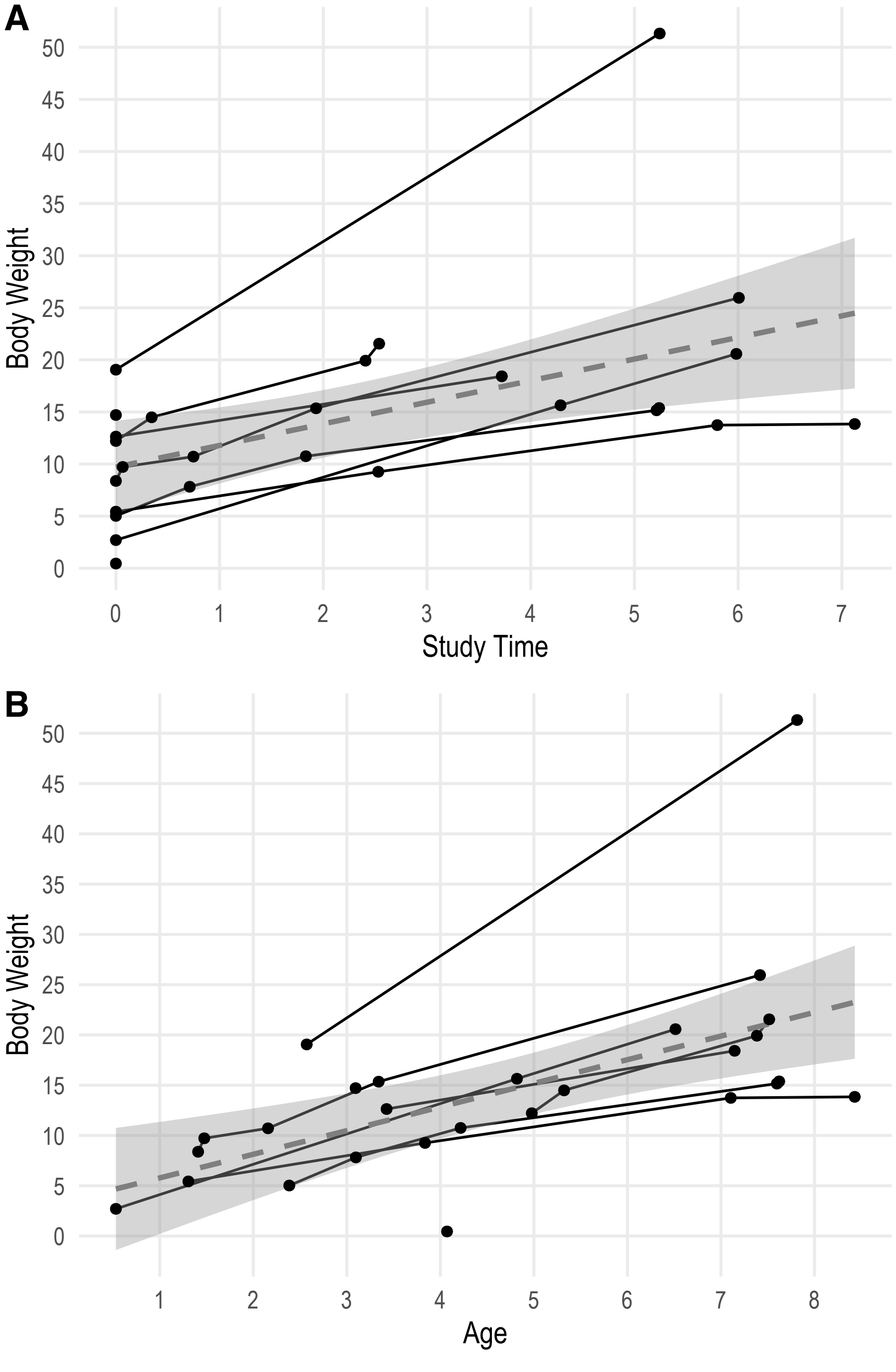


Figure 2.6: Example of a longitudinal study with irregular observation pattern using a study time scale (panel A) or an age time scale (panel B); the grey dashed line is a super-imposed regression line

2.3.1 *Notation*

The notation I will use throughout this Thesis follows that of Diggle *et al.* [31].

Let Y_{ij} be the response variable and $\mathbf{X}_{ij}$ a vector of p covariates observed at time t_{ij}, for observation $j = 1, \ldots, n_i$ and subject $i = 1, \ldots, m$. The mean and variance of Y_{ij} are represented by $E(Y_{ij}) = \mu_{ij}$ and $\mathrm{Var}(Y_{ij}) = v_{ij}$. The repeated outcomes for the i^{th} subject are represented by the vector $\mathbf{Y}_i = \{Y_{1i}, \ldots, Y_{n_i}\}$ with $E(\mathbf{Y}_i) = \mu_i$ and variance-covariance matrix $\mathrm{Var}(\mathbf{Y}_i) = V_i$ of size $n_i \times n_i$. The $\{j, k\}^{\text{th}}$ element of V_i is the covariance between Y_{ij} and Y_{ik}, denoted by $\mathrm{Cov}(Y_{ij}, Y_{ik})$. Finally, R_i represents the $n_i \times n_i$ correlation matrix of $\mathbf{Y_i}$.

2.3.2 *Approaches to Longitudinal Data Analysis*

There are several approaches to analyse longitudinal data. I will introduce two popular approaches: modelling the marginal mean as in a cross-sectional study and modelling the correlation within individuals by assuming that the regression coefficients vary between study subjects. The former approach will be introduced in Section 2.3.3. The latter approach is also known as the random effects approach, and models that include both fixed and random effects are named mixed-effects models; I will introduce the approach and further explain the differences between fixed and random effects in Section 2.3.4.

I will not cover older methods such as repeated-measures analysis of variance (ANOVA) or covariance (ANCOVA) or the analysis of derived variables (e.g. the patient-specific slope). These methods are described and illustrated in practice elsewhere [31–33].

2.3.3 *Modelling the Marginal Mean*

Marginal models focus on estimating the marginal expectation of the outcome, which is modelled separately from the within-person correlation. The general framework for this kind of models is given in Liang and Zeger [34, 35], and it extends generalised linear models to account for the within-person correlation arising from repeated measurements from a given individual.

Within this framework, the mean response for the j^{th} observation of the i^{th} subject μ_{ij} is related to a set of covariates via a link function h:

$$g(\mu_{ij}) = g[E(Y_{ij})] = \mathrm{X}_{ij}\boldsymbol{\beta} \tag{2.36}$$

In addition to that, the relationship between mean and variance is specified as:

$$\mathrm{var}(Y_{ij}) = \phi w(\mu_{ij}),$$

and the within-individual correlation is specified as:

$$\mathrm{corr}(Y_{ij}, Y_{ij'}) = \rho(\alpha).$$

This framework is quite flexible, as it can accommodate a variety of common distributions: for instance, by choosing g to be the identity, logit, and log link function (with appropriate function w) it is possible to model Gaussian, binary, and count data, respectively. These models would retain an interpretation that is analogous to linear, logistic, and Poisson regression models.

The estimation process extends the concept of quasi-likelihood [36] to settings where observations are correlated. In particular, Liang and Zeger proposed the *Generalised estimating equations* (GEE) method. The GEE method consists of solving the following set of estimating equations

$$\sum_{i=1}^{m} \mathbf{D}_i'(\boldsymbol{\beta})\mathrm{V}_i(\boldsymbol{\beta})^{-1}(\mathrm{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})) = 0 \tag{2.37}$$

for $\boldsymbol{\beta}$, where Y_i is the vector of responses for the i$^{\text{th}}$ individual, $\boldsymbol{\mu}_i$ is the modelled response from the model in Equation (2.36), $\mathbf{D}_i = \partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta}$, $\mathrm{V}_i = \mathrm{Var}(\mathrm{Y}_i) = \phi\mathrm{A}_i^{1/2}R(\alpha)\mathrm{A}_i^{1/2}$, and $\mathbf{A}_i = \mathrm{diag}(g(\mu_{i1}, \cdots, \mu_{in_i}))$. $R(\alpha)$ is the correlation structure, such as the independent correlation structure:

$$R(\alpha) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

the exchangeable correlation structure:

$$R(\alpha) = \begin{bmatrix} 1 & \alpha & \alpha \\ \alpha & 1 & \alpha \\ \alpha & \alpha & 1 \end{bmatrix},$$

or the unstructured correlation structure:

$$R(\alpha) = \begin{bmatrix} 1 & \alpha_1 & \alpha_2 \\ \alpha_1 & 1 & \alpha_3 \\ \alpha_2 & \alpha_3 & 1 \end{bmatrix}.$$

Another possibility is the autoregressive of order 1 (AR(1)) correlation structure, where $\text{Corr}(Y_{ij}, Y_{ij+t}) = \alpha^t \; \forall \; t = 1, 2, \dots, t-1$. Several criteria have been suggested for selecting the working correlation structure in GEE models: for instance, the QIC criterion (Quasi-likelihood Information Criterion) has been developed to extend the AIC by replacing the likelihood for the quasi-likelihood [37]. A comprehensive comparison of selection criteria for GEE models is given in Pardo *et al.* [38].

The estimating equations for $\boldsymbol{\beta}$ are solved using a modified Fisher scoring algorithm; Liang and Zeger showed that $\boldsymbol{\beta}$ is consistent even when the correlation structure is misspecified. Analogously, they proposed a robust estimator for the $\text{Var}(\boldsymbol{\beta})$ that is consistent even when the correlation structure is misspecified. Using these estimators it is possible to obtain valid inference under any assumed correlation structure; however, specifying an appropriate model for the correlation structure results in more efficient inference.

2.3.4 *Mixed-effects Models*

GEE models introduced in Section 2.3.3 are considered to be *population-averaged models*, as they model the marginal mean response. Conversely, mixed-effects models assume that the between-patients variability arises from unobserved covariates that are added to the linear predictor in the regression model, accounting for the natural heterogeneity between patients. The unobserved covariates are named random effects, as they are assumed to follow a given distribution; it is often assumed that the random effects follow a normal distribution with mean $\mathbf{0}$ and variance Σ_B.

Beside the unobserved random effects, fixed coefficients are almost always included in a mixed effect model for analysing longitudinal data; in fact, the fixed effects are often the coefficients of primary interest as they represent the effect of changing the value of a covariate on the average longitudinal response.

A mixed-effects model is formalised by modelling the average response, conditional on the random effects $\mathbf{B}$:

$$h[E(Y_{ij}|B_i)] = \mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{Z}_{ij}\mathbf{B}_i, \tag{2.38}$$

where $\mathbf{Z}$ is a vector of covariates with associated random effects $\mathbf{B}$; $\mathbf{X}$ and $\mathbf{Z}$ may overlap. Analogously as before, h is a link function and by appropriately choosing h it is possible to model e.g. Gaussian, binary, and count data. For instance, let's assume h is the identity function; the resulting model is a linear mixed-effects model:

$$E(Y_{ij}|B_i) = \mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{Z}_{ij}\mathbf{B}_i \tag{2.39}$$

The individual-specific model can be written as

$$Y_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{Z}_{ij}\mathbf{B}_i + \epsilon_{ij},$$

where ϵ_{ij} is a within-subject error term (e.g. measurement error), with $\epsilon_{ij} \sim N(0, \sigma^2)$. The random effects $\mathbf{B}$ are assumed to follow a multivariate normal distribution:

$$\mathbf{B} \sim N(\mathbf{0}, \Sigma_B),$$

with Σ_B a variance-covariance matrix that follows a given correlation structure, e.g. one of those introduced in Section 2.3.3.

Assuming a model with a single covariate and a random intercept only, the model above becomes:

$$Y_{ij} = \beta_0 + \beta_1 X_{i1} + b_{i0} + \epsilon_{ij}$$

In brief, it is possible to fit a population-level regression line (for an average individual), and individual-specific regression lines as well; given that I only included a random intercept in the model, the resulting individual-specific lines are parallel. This example is depicted in Figure 2.7 (panel A), using the data described in Figure 2.6; the linear mixed model is:

$$(\text{Body weight})_{ij} = \beta_0 + \beta_1 \text{Time}_{ij} + b_{i0} + \epsilon_{ij}$$

It is possible to appreciate how the individual variation is modelled by including a random intercept, resulting in individual-specific lines that are parallel, as mentioned before.

This example can be further expanded by including a random slope of time as well; the model becomes

$$(\text{Body weight})_{ij} = \beta_0 + \beta_1 \text{Time}_{ij} + b_{i0} + b_{i1} \text{Time}_{ij} + \epsilon_{ij},$$

and it is depicted in Figure 2.7, panel B. Compared to the model of Figure 2.7 (panel A), each individual-specific fitted line has a different slope, allowing the model to capture a larger portion of heterogeneity between individuals.

These two examples illustrate in practice how mixed-effects models are considered to be *subject-specific models*: by modelling the unobserved heterogeneity between individuals and the mix of between- and within-subject data a regression line (if linearity is assumed) that optimally fits the observations of each individual is fitted. Regardless, the models introduced in these two examples are still subject to the modelling assumptions introduced at the beginning of this Section.

Several estimation procedures have been proposed for mixed-effects models. In the settings of linear mixed-effects models, where the link function h is assumed to be the identity function, it is possible to define a maximum likelihood estimation procedure (as described in Chapter 4 of Diggle *et al.* [31]). However, the maximum likelihood estimator is biased for the estimation of variance components. To overcome this limitation, the restricted maximum likelihood (REML) method has been developed. REML takes into account the loss in degrees of freedom resulting from the estimation of the fixed effects and is therefore recommended (especially for small sample sizes) [39]. The downside of using the REML method is that likelihood ratio tests require exactly the same fixed effects specification, and it therefore does not allow the comparison of nested models. In settings where the outcome does not follow a Gaussian distribution, however, the

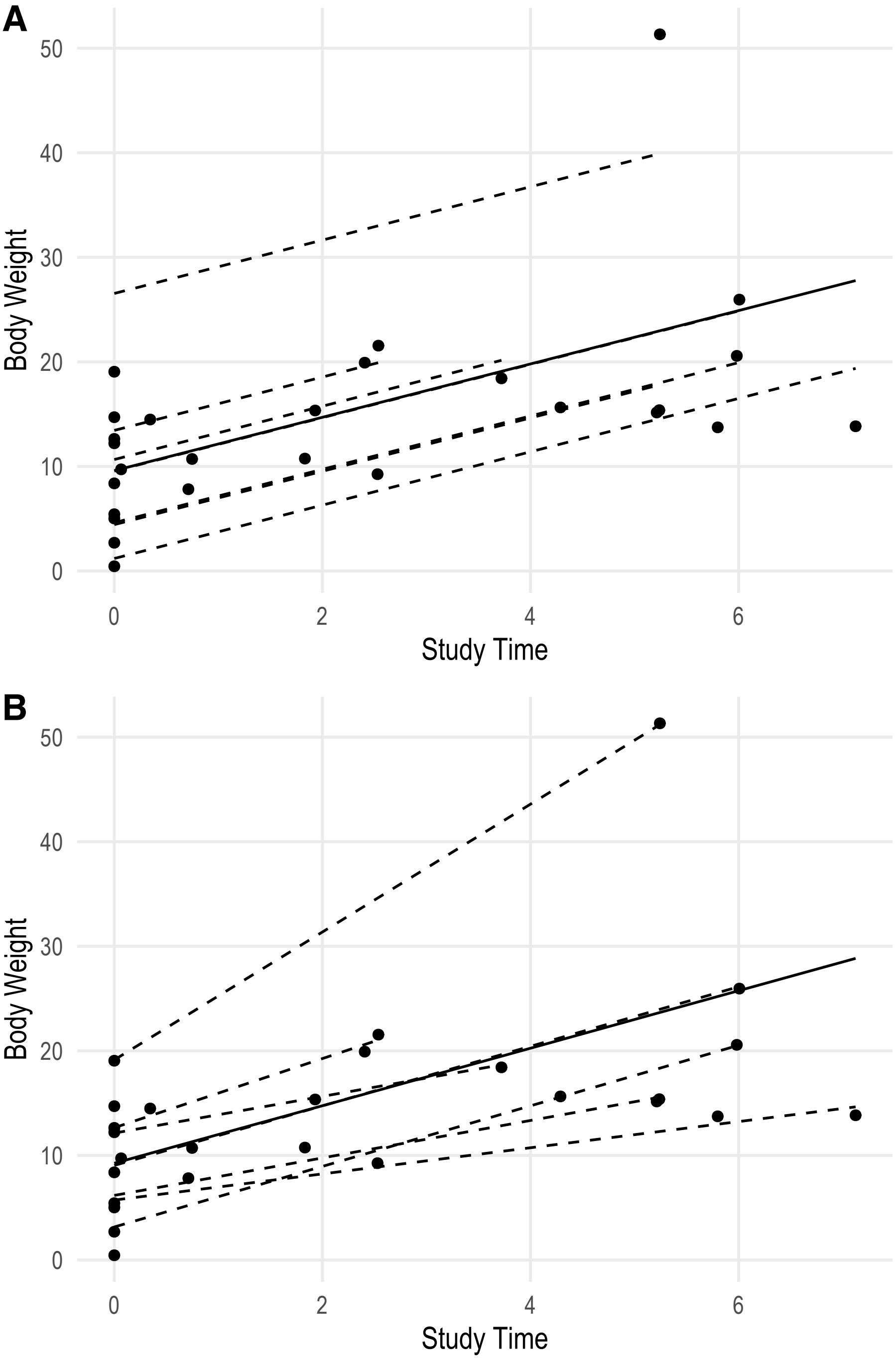


FIGURE 2.7: Population-level (solid black) and subject-specific (dashed) regression lines, fitted using a linear mixed-effects model with a random intercept only (panel A) or a random intercept and slope (panel B)

estimation procedure is more complex; methods such as the expectation-maximisation (EM) algorithm [40] and direct likelihood maximisation [41] can be used instead. Finally, all of the approaches outlined above are from a frequentist point of view; Bayesian approaches to estimation (e.g. using the probabilistic programming language Stan) could be used as well [42].

2.3.5 Marginal Versus Mixed-effects Models

Marginal and mixed-effects models described in Sections 2.3.3 and 2.3.4 have common advantages for the analysis of longitudinal data, but there are notable differences between the methods as well. Both methods are suitable to analyse longitudinal data as they account for the within-subject correlation, although to a different degree. GEE models can accommodate a single level of clustering and treat it as a nuisance parameter not giving an estimate of the heterogeneity; conversely, mixed-effects models can accommodate multiple levels of clustering and do provide estimates of the heterogeneity at each level, allowing explicit modelling of the hierarchical structure.

Another important difference is the interpretation of the regression coefficients. Coefficients of GEE models have a marginal interpretation, e.g. reflect the change in the outcome in the whole population. Coefficients of mixed-effects models have a conditional interpretation, e.g. reflect the change in the outcome for a given individual with a given covariates pattern; as mentioned before, the estimates from mixed-effects models are subject-specific. The choice of marginal versus mixed-effects models, therefore, depends on the aims of the analysis: if the interest is the average effect of covariates on the outcome in a population, then marginal models are the choice. Conversely, if the interest is the subject-specific effect of covariates, mixed-effects models are more appropriate. The fact that parameters from marginal and mixed-effects models require a completely different interpretation shows that the choice between these model families has important consequences and should be reflected upon very carefully.

Despite that, given that the random effects approach yields a fully-specified marginal likelihood, it is possible to derive marginal estimates by averaging of the distribution of the random effects $\mathbf{B}$ (as described in Chapter 16 of Molenberghs and Verbeke [43]). The comparison of marginal and mixed-effects models is discussed in more detail elsewhere [43–45].

3 *Motivating Clinical Examples*

3.1 OUTLINE

The methods investigated throughout this Thesis are illustrated in practice using a variety of datasets that present the critical issues of health care records. In particular, I will utilise data with an underlying hierarchical structure that needs to be accounted for and data with a possible association between longitudinal outcomes and either the drop-out process or the observation process. The datasets used within this Thesis are described in this Chapter.

3.2 THE PRIMARY-SECONDARY CARE PARTNERSHIP TO PREVENT ADVERSE OUTCOMES IN CHRONIC KIDNEY DISEASE (PSP-CKD) STUDY

Kidney disease is defined as an abnormality of kidney structure or function with implications for health; kidney disease is denoted as chronic when the abnormalities persist for 3 months or more [46]. Early diagnosis of chronic kidney disease (CKD) has relevant implications for the health of an individual, as early identification allows prompt treatment and strategic planning. However, early stages of CKD are often asymptomatic, and disease may progress faster leading to kidney failure and consequently needing renal replacement therapy (such as dialysis). At the same time, several comorbid conditions emerge as renal function (defined as estimated glomerular filtration rate, eGFR) worsen, such as diabetes and cardiovascular disease. CKD has no cure, but treatment can help reduce symptoms and slow down disease progression: many individuals with CKD can live long and largely normal lives. Despite that, worldwide data is showing a large and increasing burden of CKD, putting increasing strain on healthcare systems across the world [47].

The *Primary-Secondary Care Partnership to Prevent Adverse Outcomes in Chronic Kidney Disease* (PSP-CKD) study (ClinicalTrials.gov Identifier: NCT01688141) is a cluster randomised controlled pragmatic trial of enhanced chronic kidney disease (CKD) care against usual primary care management [48]. Forty-nine primary care practices in Northamptonshire, United Kingdom, were randomised to either routine care or enhanced care; informed consent was provided at the practice level. Adult individuals with CKD were identified from each practice by using a research version of the web-based CKD management and audit tool IMPAKT [49], and all data was anonymised prior to removal from the primary care practice. Individuals were included if a recorded eGFR value below 60 ml/min/1.73m^2 was found during 5 years before the date of randomisation; eGFR was estimated using the MDRD equation [50]. PSP-CKD investigators concluded that CKD management programs in primary care did not slow disease progression, but improved care and have the potential to decrease cardiovascular disease burden and related costs.

Using data extracted from the PSP-CKD study, I constructed two study datasets for the applied examples of this Thesis: a dataset with a time to event outcome and a dataset with a longitudinal outcome. The dataset with a time to event outcome was constructed by extracting all individuals at randomisation date and following them until the event of interest or December 31th, 2013; the event of interest was kidney failure, defined as the first eGFR measurement below 15 ml/min/1.73m^2. The main exposure was treatment (as defined above), and information on which practice individuals belonged to was extracted as well - alongside with data on demographics (age and gender). The second dataset was constructed by extracting all eGFR measurements recorded on or after randomisation date; eGFR measurements above 90 ml/min/1.73m^2 were discarded to focus on individuals with abnormal kidney function [46]. Besides the longitudinal eGFR measurements, information on treatment and demographics (gender and age at baseline) was extracted.

The first dataset will be used in Chapter 5 to illustrate the analysis of time to event outcomes with observations clustered within groups, in this case the primary care

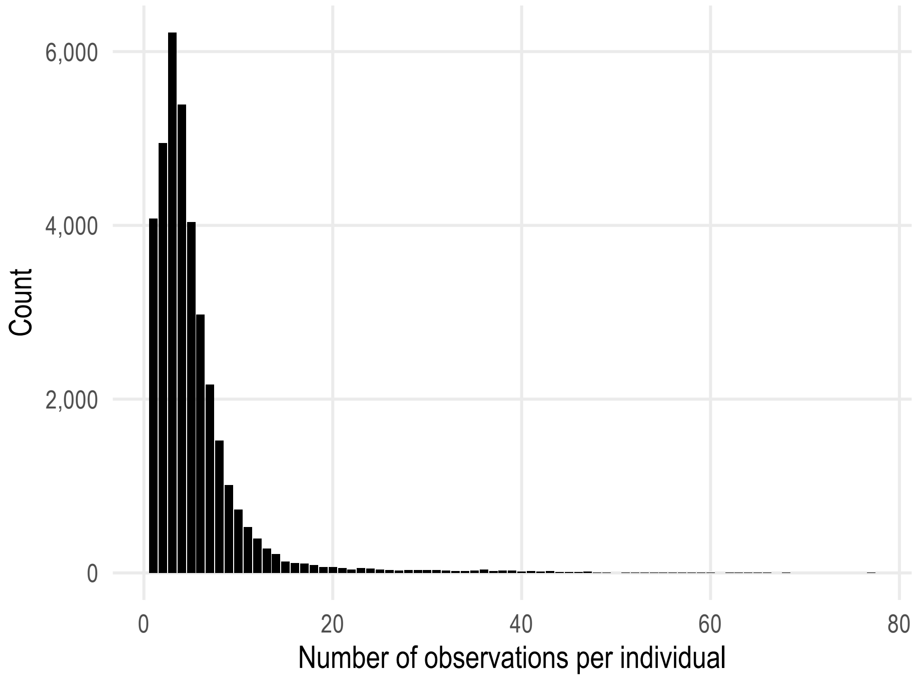


FIGURE 3.1: Distribution of number of observations per individual, dataset constructed by extracting longitudinal data from the PSP-CKD study and used for the applied example of Chapter 7

practices. It included 25,884 individuals clustered within 46 practices; median age was 75.80, and 61.58% were females. 12,724 individuals (49.16%) received enhanced care, the intervention being studied. Median follow-up, estimated via the inverse Kaplan-Meier method [51] and disregarding clustering, was 3.24 years; 458 individuals (1.77%) experienced the event of interest during follow-up.

The second dataset will be used in Chapter 7 to illustrate the analysis of longitudinal data where the timing of each measurement is (potentially) associated with the outcome itself. The dataset consists of 187,671 longitudinal measurements for 35,822 individuals; the median number of observations per individual is 4 (inter-quartile interval: 2 - 6), as depicted in Figure 3.1. 21,674 individuals were females (60.50%), median age was 74.60, and 17,952 individuals received enhanced care (50.11%). The median gap time between observations was 0.61, inter-quartile interval: 0.23 - 1.06.

3.3 The VAsopressin Versus Norepinephrine in Septic Shock Trial (VASST)

Sepsis is the serious complication of an infection that could lead to multiple organ failure and death. In particular, septic shock is the most common cause of death in intensive care units (ICUs) [52, 53]. Intravenous fluids and catecholamines such as norepinephrine are routinely used to resuscitate patients, although they can have significant side effects. Vasopressin, an endogenously released peptide hormone, has been suggested as a potential treatment option for patients with severe septic shock.

The randomised controlled trial of *VAsopressin versus norepinephrine in Septic Shock* (VASST) is a multicentre, randomised, stratified, double-blind trial among patients who had septic shock and were receiving usual care to determine whether the additional treatment with vasopressin decreased mortality. The control group received norepinephrine only as part of usual care. VASST was run between 2001 and 2006 in Canada, Australia, and the United States, and was registered in the ISRCTN registry (Current Controlled Trials number: ISRCTN94845869, [54]). The results of the trial were negative, with vasopressin not showing reduced mortality rates compared to norepinephrine [55].

The data from VASST that I will be using consists of daily measurements of the Sepsis-related Organ Failure Assessment [SOFA] score [56]; survival information will be used as well. The SOFA score consists of six organ subscales (cardiovascular, central nervous system, coagulation, liver, renal, and respiration) that range from 0 to 4, representing no organ dysfunction and a high degree of dysfunction/failure, respectively; the total SOFA score can range therefore from 0 to 24. The SOFA score is commonly analysed as a secondary outcome in sepsis trials; for instance, a systematic review and meta-regression of 87 randomised controlled trials (RCTs) found that treatment-associated changes in SOFA from baseline were reliably and consistently associated with observed mortality [57].

The analysis dataset includes 6,934 SOFA score measurements for 763 individuals, with a median number of observations per individual of 21 (inter-quartile interval: 10 - 29). 389 individuals (50.98%) received vasopressin, while the remaining individuals received

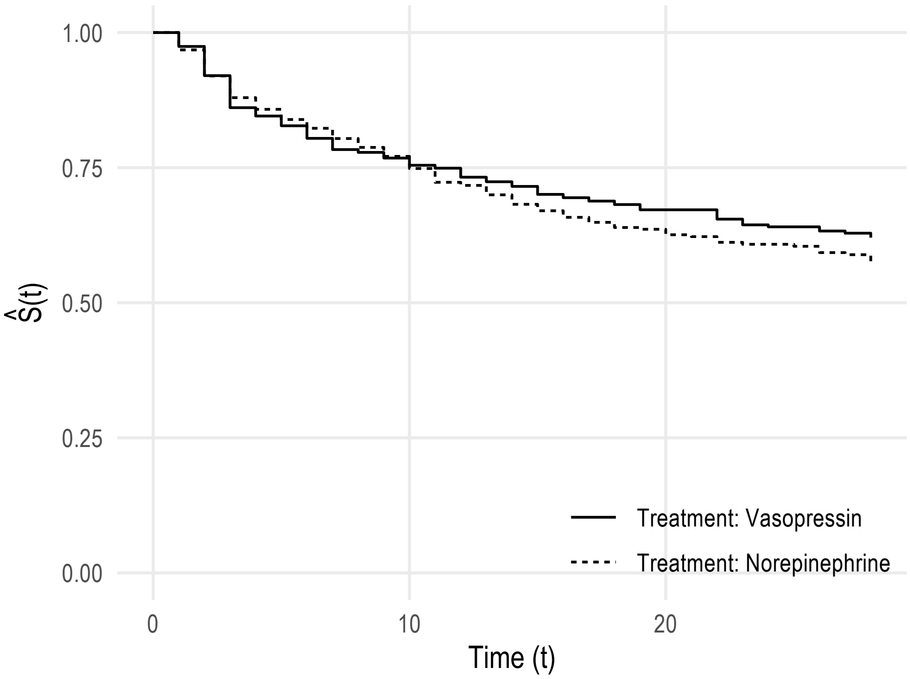


FIGURE 3.2: Kaplan-Meier estimate of the survival function by treatment arm, VASST trial data

norepinephrine.

Median follow-up was 28 days; 284 individuals (37.22%) died during follow-up. The Kaplan-Meier estimate of the survival function by treatment arm is depicted in Figure 3.2: the log-rank test yields a p-value of 0.32, showing that the observed difference in survival between treatment arms is not statistically significant. This result is in agreement with the published results of VASST [55].

An issue with longitudinal data from VASST is informative drop-out: characteristics of septic shock patients that died before the end of the study may be different than the characteristic of patients surviving the whole follow-up; in Chapter 6 I will illustrate how differential drop-out can affect analysis methods and needs to be accounted for. In particular, I will re-analyse data from VASST following the hypothesis that - although the results of the trial showed that vasopressin did not decrease the mortality rate - it may have a differential effect on decreasing the SOFA score compared to norepinephrine.

4 *Monte Carlo Simulation Studies*

4.1 OUTLINE

In this Chapter, I introduce Monte Carlo simulation studies and the rationale for their use; Monte Carlo simulation studies are heavily used throughout this Thesis, especially in Chapters 5 and 7. I outline considerations on the design and analysis of Monte Carlo simulation studies, including summary statistics and their Monte Carlo errors in Section 4.3; I also describe methods that can be used to simulate complex survival data in Section 4.4. In Section 4.5 I introduce open source software that I developed during my PhD to analyse and report Monte Carlo simulation studies. In Section 4.6 I illustrate a case study of designing, running, and analysing a simulation study using the tools introduced in this Chapter. Finally, I conclude the Chapter with a brief discussion in Section 4.7.

4.2 RATIONALE FOR MONTE CARLO SIMULATION STUDIES

Monte Carlo simulation studies are computer experiments based on generating pseudo-random samples from a given probability distribution. Most interestingly, by running this process multiple times it is possible to collect experimental data supporting, for example, the comparison of methods included in a study. Whilst case studies are important and also useful when comparing methods in relative terms, with Monte Carlo simulation studies the *truth* is pre-defined: by doing so, it is possible to compare the results of a Monte Carlo simulation study with the truth to understand e.g. which method included in the study performs best. In other words, Monte Carlo simulation studies represent the only way to understand whether a method provides the *right answer* given that the truth is known in advance. Statisticians usually mean *Monte Carlo simulation study* when they refer to *Simulation study*; throughout this Thesis, I will use the two terms interchangeably.

Simulation studies have several applications and represent an invaluable tool for statistical research nowadays: in statistics, establishing properties of current methods is extremely important to allow their use with confidence. However, sometimes it is very hard (if not impossible) to derive exact analytical properties; large sample approximation is possible but evaluating the goodness of the approximation to finite samples is required. Approximations often require assumptions as well: what are the consequences of violating such assumptions? Monte Carlo simulation studies come to the rescue and can help answer these questions. They can also help to answer questions such as:

- Is an estimator biased in a finite sample?
- Do confidence intervals for a given parameter achieve the desired nominal level of coverage?
- How does a newly developed method compare to an established one?
- What is the power to detect a desired effect size under complex experimental settings and analysis methods?

Simulation studies are used increasingly often in statistical research, due to the increased availability of powerful computational tools (both personal and high-performance cluster computers), the perceived efficacy, and the emergence of specialist courses and tutorial papers on simulation studies [58]. For instance, searching on the database of peer-reviewed research literature Scopus [59] with the simple query string `TITLE-ABS-KEY ("Monte Carlo simulation study")` yields almost 3,000 results with a 15-fold increase during the last 30 years, from 18 documents in 1988 to 277 in 2018 (Figure 4.1).

Despite the increased popularity, simulation studies are often poorly designed, analysed, and reported [58]. Information on data-generating mechanisms (DGMs), number of repetitions, software, estimands are often lacking or poorly reported, making critical appraisal and replication of published studies a difficult task. Another aspect of simulation studies that is often poorly reported or not reported at all is the Monte Carlo error of summary statistics, defined as the standard deviation of the estimated quantity over repeated simulation studies. Monte Carlo errors play an important role in understanding the role of chance in the results of simulation studies and have been shown to be severely underreported [60]. In the next Section I will describe a structured approach that can

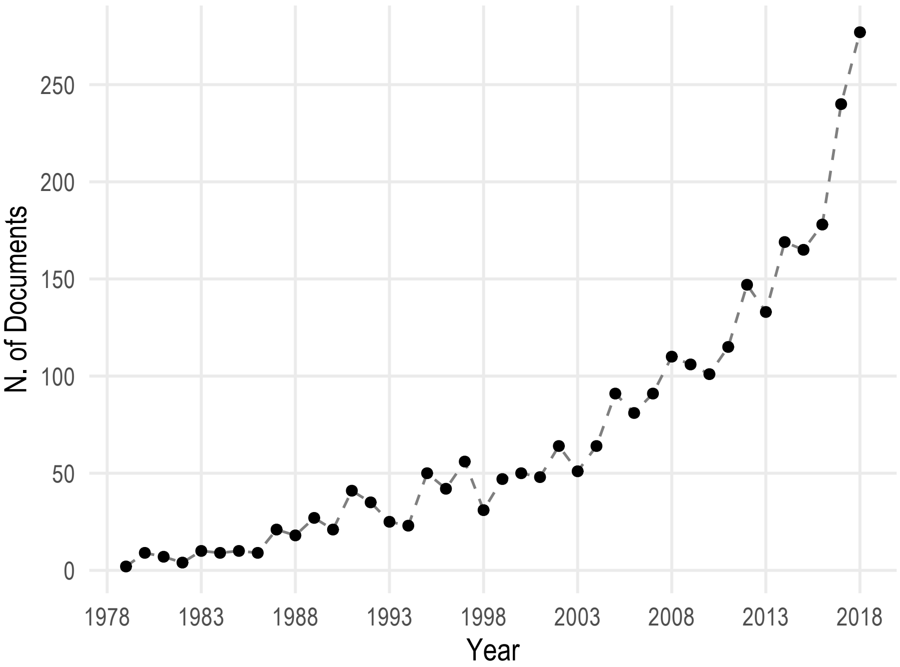


FIGURE 4.1: Number of documents identified from Scopus using the query string `TITLE-ABS-KEY ("Monte Carlo simulation study")` between the years 1979 and 2018

be applied to the design and planning of Monte Carlo simulation study and was first introduced by Morris, White, and Crowther in 2019 [58]. I will also introduce (1) summary statistics and performance measures that can be computed to characterise the behaviour of methods included in a simulation study and (2) Monte Carlo standard errors that are useful to quantify the uncertainty arising from the estimation of performance measures.

4.3 DESIGN AND ANALYSIS OF MONTE CARLO SIMULATION STUDIES

I will begin by introducing some notation. First of all, n_{obs} is the sample size of a given simulated dataset, n_{sim} is the number of replications of the simulation procedure, and $i = 1, \dots, n_{\text{sim}}$ indexes each replication. θ denotes an estimand, $\hat{\theta}$ is its estimator, $\hat{\theta}_i$ is the estimate of θ from the i^{th} replication, and $\bar{\theta}$ is the mean of $\hat{\theta}_i$ across repetitions. $\text{Var}(\hat{\theta})$ is the true variance of $\hat{\theta}$, and $\widehat{\text{Var}}(\hat{\theta}_i)$ is the estimate of $\text{Var}(\hat{\theta})$ from the i^{th} replication.

Next, I will introduce the ADEMP structured approach to planning, designing, and running simulation studies, first proposed by Morris, White, and Crowther [58]. ADEMP stands for *Aims*, *Data-generating mechanisms*, *Estimands*, *Methods*, *Performance measures* and argues that by carefully designing, describing, and reporting each one of these aspects the clarity, reproducibility, and reporting of simulation studies could be greatly improved.

The first step of a simulation study consists of carefully defining the aims of the study; for instance, one may want to study and learn about the precision or efficiency of a new method. Possible aims of a simulation study can be broadly categorised into three groups:

1. Simulation studies that aim to prove that it is possible to apply a method in given settings;
2. Simulation studies that aim to stretch or break methods, e.g. to discover settings in which a given method performs well (or not);
3. Simulation studies that aim to compare the performance of competing methods and identify in which settings a given method is preferable compared to a competing one.

The second step of a simulation study consists of defining the data-generating mechanism(s) (DGMs). DGMs denote the processes used to simulate data, either via a parametric formulation or resampling from an existing dataset. When resampling from a dataset, the true data-generating model remains unknown; conversely, when simulating data from a parametric, user-defined model the truth is known and it becomes feasible to explore a plethora of DGMs. It is important to define several DGMs to ensure coverage of different scenarios and increase the generalisability of the results, and it is important to carefully define DGMs to help to achieve the aims of the study.

The third step of a simulation study consists of defining the estimand of interest. The estimand of interest, denoted with θ, is usually a parameter of the data-generating model (e.g. a regression coefficient) but it could be some other quantity as well (such as measures of predictive ability). However, sometimes a simulation study may not target an estimand and may be targeting a procedure instead, e.g. the choice of an analysis method based on preliminary results on the same data [61, 62].

The fourth step of a simulation study consists of defining the methods being studied; the term *method* is generic, and it refers to anything being studied via simulation. In some simulation studies a single method is sufficient, e.g. when evaluating the feasibility of applying a given method in some settings. In others, however, it is necessary to include at least another method e.g. when comparing several methods to identify which one performs best in some settings. In these settings, it is important to choose a *serious* *contender* (in the words of Morris, White, and Crowther) such as the current gold standard or a method commonly used in practice.

Finally, the fifth step consists of defining the performance measures of interest. Performance measures are, broadly speaking, every numerical quantity that can be used to assess the performance of a method; the choice of performance measures for a given study depends on the aims of the study. An important point that is often overlooked when analysing Monte Carlo simulation studies is that performance measures are estimates themselves and therefore subject to error. Monte Carlo standard errors quantify the uncertainty arising from the estimation of performance measures, and as mentioned before they are severely underreported [60]. I present the most common performance measures including their definition, estimate, and Monte Carlo standard error in Table 4.1.

Briefly describing each performance measure:

- Bias quantifies whether a method targets θ on average. Sometimes the average estimate of $\hat{\theta}_i$ is reported instead, and sometimes relative bias is preferable;
- The empirical standard error measures precision or efficiency of the estimator of θ; it depends only on θ_i and does not require knowledge of θ. Intuitively, the empirical SE estimates the long-run standard deviation of θ_i over the n_{sim} replications;
- Given that the empirical standard error is sometimes hard to interpret, relative precision is often of interest;
- Mean squared error (MSE) is a measure that takes into account bias and variance of $\hat{\theta}$, and appears to be a natural way of integrating both measures into a single summary value;
- The average model-based standard error is the average standard error returned from a given method. It targets the empirical standard error;
- Relative error in model-based standard error compares empirical standard error and model-based standard error, illustrating the performance of the estimator for model-based standard error (in relative terms);
- Coverage probability is defined as the probability that a confidence interval contains the true value θ;
- Bias-corrected coverage probability takes bias into account as a source of under-

or over-coverage, and removes bias from the calculation of coverage probability by targeting $\bar{\theta}$ rather than θ;
- Power describes the performance of a test (and a simulation study) that targets a null hypothesis.

Further details on each performance measure are included in Morris, White, and Crowther [58].

In conclusion, I introduced the ADEMP structured approach for designing and running Monte Carlo simulation studies; the reporting steps outlined by ADEMP should be followed to ensure a simulation study can be easily understood and replicated (if needed).

In addition to the aspects defined above, Monte Carlo simulation studies require careful considerations in terms of computational and programming issues (such as controlling the process that generates pseudo-random numbers and the use of different software packages for different methods); once again, details on these aspects are discussed in more detail elsewhere [58].

Table 4.1: Performance measures commonly used with Monte Carlo simulation studies; this table follows from Morris *et al.* [58]

Performance Measure	Definition	Estimate	Monte Carlo SE of Estimate
Bias	$E(\hat{\theta}) - \theta$	$\frac{1}{n_{\text{sim}}} \sum_{i=1}^{n_{\text{sim}}} \hat{\theta}_i - \theta$	$\sqrt{\frac{1}{n_{\text{sim}}(n_{\text{sim}}-1)} \sum_{i=1}^{n_{\text{sim}}} (\hat{\theta}_i - \bar{\theta})^2}$
Empirical SE (EmpSE)	$\sqrt{\text{Var}(\hat{\theta})}$	$\sqrt{\frac{1}{n_{\text{sim}}-1} \sum_{i=1}^{n_{\text{sim}}} (\hat{\theta}_i - \bar{\theta})^2}$	$\frac{\widehat{\text{EmpSE}}}{\sqrt{2(n_{\text{sim}}-1)}}$
Relative % increase in precision (B vs A)*	$100\left(\frac{\text{Var}(\hat{\theta}_A)}{\text{Var}(\hat{\theta}_B)} - 1\right)$	$100\left(\left(\frac{\widehat{\text{EmpSE}}_A}{\widehat{\text{EmpSE}}_B}\right)^2 - 1\right)$	$200\left(\frac{\widehat{\text{EmpSE}}_A}{\widehat{\text{EmpSE}}_B}\right)^2 \sqrt{\frac{1-\text{Corr}(\hat{\theta}_A, \hat{\theta}_B)^2}{n_{\text{sim}}-1}}$
Mean squared error (MSE)	$E[(\hat{\theta} - \theta)^2]$	$\frac{1}{n_{\text{sim}}} \sum_{i=1}^{n_{\text{sim}}} (\hat{\theta}_i - \theta)^2$	$\sqrt{\frac{\sum_{i=1}^{n_{\text{sim}}} \left[(\hat{\theta}_i-\theta)^2 - \widehat{\text{MSE}}\right]^2}{n_{\text{sim}}(n_{\text{sim}}-1)}}$
Average model-based SE (ModSE)*	$\sqrt{E[\widehat{\text{Var}}(\hat{\theta})]}$	$\sqrt{\frac{1}{n_{\text{sim}}} \sum_{i=1}^{n_{\text{sim}}} \widehat{\text{Var}}(\hat{\theta}_i)}$	$\sqrt{\frac{\widehat{\text{Var}}[\widehat{\text{Var}}(\hat{\theta})]}{4 n_{\text{sim}} \widehat{\text{ModSE}}^2}}$ †
Relative % error in ModSE*	$100\left(\frac{\text{ModSE}}{\text{EmpSE}} - 1\right)$	$100\left(\frac{\widehat{\text{ModSE}}}{\widehat{\text{EmpSE}}} - 1\right)$	$100\left(\frac{\widehat{\text{ModSE}}}{\widehat{\text{EmpSE}}}\right) \sqrt{\frac{\widehat{\text{Var}}[\widehat{\text{Var}}(\hat{\theta})]}{4 n_{\text{sim}} \widehat{\text{ModSE}}^4} + \frac{1}{2(n-1)}}$ †
Coverage	$P(\hat{\theta}_{\text{low}} \leq \theta \leq \hat{\theta}_{\text{upp}})$	$\frac{1}{n_{\text{sim}}} \sum_{i=1}^{n_{\text{sim}}} 1(\hat{\theta}_{\text{low,i}} \leq \theta \leq \hat{\theta}_{\text{upp,i}})$	$\sqrt{\frac{\widehat{\text{Cover}}(1-\widehat{\text{Cover}})}{n_{\text{sim}}}}$
Bias-eliminated coverage	$P(\hat{\theta}_{\text{low}} \leq \bar{\theta} \leq \hat{\theta}_{\text{upp}})$	$\frac{1}{n_{\text{sim}}} \sum_{i=1}^{n_{\text{sim}}} 1(\hat{\theta}_{\text{low,i}} \leq \bar{\theta} \leq \hat{\theta}_{\text{upp,i}})$	$\sqrt{\frac{\widehat{\text{B-E Cover}}(1-\widehat{\text{B-E Cover}})}{n_{\text{sim}}}}$
Rejection % (power of type I error)	$P(p_i \leq \alpha)$	$\frac{1}{n_{\text{sim}}} \sum_{i=1}^{n_{\text{sim}}} 1(p_i \leq \alpha)$	$\sqrt{\frac{\widehat{\text{Power}}(1-\widehat{\text{Power}})}{n_{\text{sim}}}}$

*]Monte Carlo SEs are approximate for *Relative % increase in precision*, *Average ModSE*, and *Relative % error in ModSE* †]$\widehat{\text{Var}}[\widehat{\text{Var}}(\hat{\theta})] = \frac{1}{n_{\text{sim}}-1} \sum_{i=1}^{n_{\text{sim}}} \{\widehat{\text{Var}}(\hat{\theta}_i) - \frac{1}{n_{\text{sim}}} \sum_{j=1}^{n_{\text{sim}}} \widehat{\text{Var}}(\hat{\theta}_j)\}^2$

4.4 Simulating Survival Data

The generation of biologically plausible survival times assumed to follow a given distribution to run Monte Carlo simulation studies is of primary importance. In fact, this Thesis (and especially Chapters 5 and 7) will rely on simulating survival data. In this Section, I will introduce methods that have been proposed to simulate biologically plausible survival times straightforwardly and efficiently.

First, Bender *et al.* described how to apply the inversion method to simulate survival times from a given model with known regression coefficients, covariates, and any non-zero baseline hazard rate following a parametric distribution (exponential, Weibull, or Gompertz) [63]. This method is computationally efficient and easy to implement and can be described as follows. Say I want to simulate survival times from a given proportional hazards survival model:

$$h(t) = h_0(t) \exp(X\beta),$$

where $h_0(t)$ is a given parametric baseline hazard function and X is a vector of regression coefficients with associated vector of regression parameters β. The corresponding cumulative hazard function $H(t)$, survival function $S(t)$, and cumulative distribution function $F(t)$ follows (as defined in Chapter 2):

$$\begin{aligned} H(t|X) &= H_0(t) \exp(X\beta), \text{ with } H_0(t) = \int_0^t h_0(u)\, du \\ S(t|X) &= \exp[-H(t|X)] \\ F(t|X) &= 1 - \exp[-H(t|X)] \end{aligned}$$

Let T be the simulated survival time; by letting

$$F(T|X) = 1 - \exp[-H(T|X)] = u \tag{4.1}$$

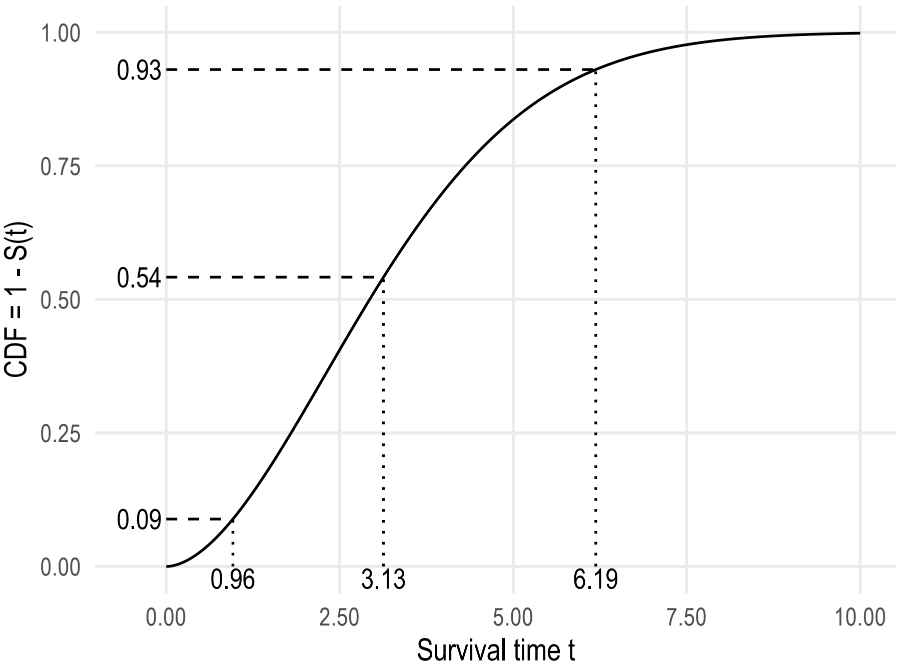


FIGURE 4.2: Example application of the inversion method to simulate survival times

with $u \sim \text{Unif}(0, 1)$, or alternatively:

$$S(T|X) = u \tag{4.2}$$

If $h_0(T) > 0$ then Equation (4.2) can be re-arranged and solved for T as long as $H_0(t)$ can be directly inverted:

$$T = H_0^{-1}\left[-\frac{\log(u)}{\exp(X\beta)}\right] \tag{4.3}$$

Practically speaking, u represents a simulated quantile from a given cumulative distribution function (CDF); simulating a survival time follows by drawing from a uniform distribution and applying Equation (4.3). This approach is illustrated in practice in Figure 4.2: first, three values of u are drawn (on the vertical axis), 0.09, 0.54, and 0.93. The three values have a unique corresponding value of the CDF, consequently yielding a given simulated survival time (on the horizontal axis).

Interestingly, Bender *et al.* showed that it is possible to obtain closed-form formulæ for the three parametric distributions introduced in Chapter 2 (Table 4.2).

To simulate survival data following a given distribution (and a given model formulation) it is then sufficient to be able to draw from a uniform $U(0, 1)$ distribution (e.g. using the `runif()` function in `R`) and then applying the formulæ from Table 4.2; the whole process

Table 4.2: Formulæ to simulate survival times following an exponential, Weibull, or Gompertz distribution

	Exponential	Weibull	Gompertz
Survival time	$T = -\frac{\log(u)}{\lambda \exp(X\beta)}$	$T = -\left(\frac{\log(u)}{\lambda \exp X\beta}\right)^{1/\gamma}$	$T = \frac{1}{\gamma} \log\left[1 - \frac{\gamma \log(u)}{\lambda \exp(X\beta)}\right]$
Hazard function	$h(t) = \lambda \exp(X\beta)$	$h(t) = \lambda \gamma t^{\gamma - 1} \exp(X\beta)$	$h(t) = \lambda \exp(\gamma t) \exp(X\beta)$

does not require complex calculations and is therefore very fast.

One of the limitations of the method outlined above is that it requires the cumulative baseline hazard function to be easily invertible: when that is not the case, it is not possible to obtain closed-form formulæ for the survival time to be simulated. Furthermore, the exponential, Weibull, and Gompertz distribution - being monotonic functions - are not flexible enough to capture the underlying complexities often encountered in clinical data, where turning points in the baseline hazard can be observed (e.g. the hazard could be high early on, decreasing after a few days, and ultimately increasing again to a moderate level).

Crowther and Lambert [64] demonstrated that is is possible to extend the inversion method by using numerical methods to simulate a large variety of more biologically plausible survival times. For instance, Crowther and Lambert illustrate how to simulate from complex baseline hazard functions: their method consists of selecting a baseline hazard function with a closed-form cumulative baseline hazard function $H_0(t)$. Assuming $H_0(t)$ cannot be inverted analytically, a numerical root-finding method can be used to solve Equation (4.2) for t.

In more detail, say I am assuming a two-components mixture Weibull distribution for the baseline hazard. The mixture Weibull distribution, used in standard survival analysis and in models with statistical cure [65, 66], defines the two components additively on the survival scale. The resulting baseline survival function for a two-components mixture is:

$$S_0(t) = \pi S_{01}(t) + (\pi - 1) S_{02},$$

where $S_{01}(t)$ and $S_{02}(t)$ are the survival functions of each component (in this case, the survival function following from a Weibull baseline hazard) and π is the mixing probability ($0 \leq \pi \leq 1$). The survival function can be expanded as

$$S_0(t) = p\exp(-\lambda_1 t^{\gamma_1}) + (1-p)\exp(-\lambda_2 t^{\gamma_2})$$

with λ_1, λ_2 and γ_1, γ_2 being the scale and shape parameters of the Weibull components. The next step consists of introducing covariates in the model:

$$\begin{aligned} S(t) &= S_0(t)^{\exp(X\beta)} \\ &= [p\exp(-\lambda_1 t^{\gamma_1}) + (1-p)\exp(-\lambda_2 t^{\gamma_2})]^{\exp(X\beta)} \end{aligned} \tag{4.4}$$

Then, Equation (4.4) can be substituted into Equation (4.2):

$$[p\exp(-\lambda_1 t^{\gamma_1}) + (1-p)\exp(-\lambda_2 t^{\gamma_2})]^{\exp(X\beta)} - u = 0, \tag{4.5}$$

with u being a value drawn from a $U(0,1)$ distribution. Equation (4.5) cannot be solved analytically for t, hence Crowther and Lambert suggest using a numerical method such as Brent's method [67]. Implementation using standard statistical software is straightforward: in fact, the authors have released their implementation in Stata and a version in R is available as well [68, 69].

Finally, the work by Crowther and Lambert further extend methods to simulate biologically plausible survival times e.g. by allowing the simulation of time-dependent effects, time-dependent covariates, delayed entry, censoring distributions, and even user-defined log-hazard function (for increased flexibility) [68].

4.5 Open Source Software to Analyse Monte Carlo Simulation Studies

In this Section I will describe open-source software that I have developed to aid and support the analysis of Monte Carlo simulation studies. First, I will introduce the R package `rsimsum` in Section 4.5.1; then, I will introduce the Shiny app INTEREST in Section 4.5.2. Both have been released publicly (on the Comprehensive R Archive Network, CRAN, and/or on GitHub) and are openly available for users to test and use. Further to that, an article for `rsimsum` has been published in the Journal of Open Source Software and is included in Appendix A, while a manuscript introducing INTEREST is currently under review in the Journal of Data Science, Statistics, and Visualisation (with a pre-print included in Appendix ??).

4.5.1 `rsimsum`

`rsimsum` is an R package that can be used to analyse Monte Carlo simulation studies and compute summary statistics. The package is published on CRAN, and can be installed from within R as follows:

```
install.packages("rsimsum")
```

In addition to that, the source code and the current development version is published online on GitHub at `https://github.com/ellessenne/rsimsum`.

`rsimsum` is modelled upon a similar package available in Stata, the user-written command `simsum` [70]: to the best of my knowledge, there is no similar package in R.

The aim of `rsimsum` is to help reporting simulation studies, including understanding the role of chance in results of simulation studies: Monte Carlo standard errors are computed and reported by default to the user. The summary statistics supported by `rsimsum` are those included in Table 4.1: bias, empirical and model-based standard errors, relative precision, relative error in model standard error, mean squared error, coverage, bias.

The main function of `rsimsum` is called `simsum` and can handle simulation studies with a single estimand of interest at a time. The arguments of the `simsum` function are:

```
args(rsimsum::simsum)

# function (data, estvarname, true, se, methodvar = NULL, ref = NULL,
#     by = NULL, ci.limits = NULL, dropbig = FALSE, x = FALSE,
#     control = list())
# NULL
```

- `data` is a `data.frame` in which variable names are interpreted. It has to be in tidy format [71];
- `estvarname` is the column in `data` that contains the point estimates;

- `true` is the true value of the estimand, e.g. θ;
- `se` is the column in `data` that contains the standard error of the point estimate;
- `methodvar` is the column in `data` that contains the methods to compare. For instance, methods could be the models compared within a simulation study. This argument is not required, e.g. when a simulation study included a single method only;
- `ref` specifies the reference method against which relative precision will be calculated. Only useful if `methodvar` is specified;
- `by` is a vector of column names that defines the data-generating mechanisms. As with `methodvar`, this argument is not required e.g. if there is a single data-generating mechanism;
- `ci.limits` is a numeric vector of length 2 specifying the limits (lower and upper) of confidence intervals used to calculate coverage. This feature is experimental;
- `dropbig` specifies whether point estimates or standard errors beyond the maximum acceptable values should be dropped. Defaults to `FALSE`, and acceptable values can be set via the `control` argument;
- `x` defines whether the final dataset used to calculate the summary statistics (i.e. after all pre-processing steps) should be returned or not. Defaults to FALSE, as the size of the returned object increases considerably;
- `control` is a list of arguments that control the default behaviour of `simsum`. For instance:
  - `mcse`, whether to calculate Monte Carlo standard errors. Defaults to `TRUE`;
  - `level`, the significance level used for coverage, bias-eliminated coverage, and power. Defaults to 0.95;
  - `df`, whether to use robust critical values from a t distribution with df degrees of freedom when calculating coverage, bias-eliminated coverage, and power. Defaults to NULL, in which case a Gaussian distribution is used;
  - `na.rm`, whether to remove point estimates or standard errors where either (or both) is missing. Defaults to `TRUE`;
  - `char.sep`, a character utilised when splitting the input dataset data. Generally, this should not be changed;
  - `dropbig.max`, specifies the maximum acceptable absolute value of the point estimates, after standardisation. Defaults to 10;

  - `dropbig.semax`, specifies the maximum acceptable absolute value of the standard error, after standardisation. Defaults to 100
  - `dropbig.robust`, specifies whether to use robust standardisation (using median and inter-quartile range) rather than normal standardisation (using mean and standard deviation). Defaults to `TRUE`, in which case robust standardisation will be used for `dropbig`.

`rsimsum` also provides a function to analyse Monte Carlo simulation studies with multiple estimands: `multisimsum`.

```
args(rsimsum::multisimsum)
```

```
# function (data, par, estvarname, true, se, methodvar = NULL,
#     ref = NULL, by = NULL, ci.limits = NULL, dropbig = FALSE,
#     x = FALSE, control = list())
# NULL
```

The arguments of `multisimum` are mostly the same of `simsum`, with two exceptions: first, an argument named `par` has been added, identifying the column in `data` that defines estimands. Second, the `true` argument now has to be a named vector, defining the true value of each estimand being analysed.

Finally, an important step of reporting a simulation study consists in visualising the results. `rsimsum` implements the `autoplot` method for `simsum` and `multisimsum` objects, exploiting the R package `ggplot2` [72] to produce a set of opinionated data visualisations useful to quickly and easily explore the results of simulation studies. Supported plots are: scatter plots for method-wise comparison of point estimates (or standard errors), Bland-Altman plots for method-wise comparison of point estimates (or standard error), ridgeline plots, forest plots, lolly plots, zipper plots, heat maps, and nested loop plots [73]. The latter is a visualisation type that aims to include all results in a single plot by ordering all simulation scenarios and arranging them consecutively on the horizontal axis; the summary statistics of interest is then included on the vertical axis, with a line for each method included in the simulation study. The first three plot types allow comparing the point estimates (or standard errors) between methods, allowing e.g. the identification of serial trends (when a method constantly produces standard errors that are larger than those of a comparator). Conversely, the remaining plots are useful to plot summary

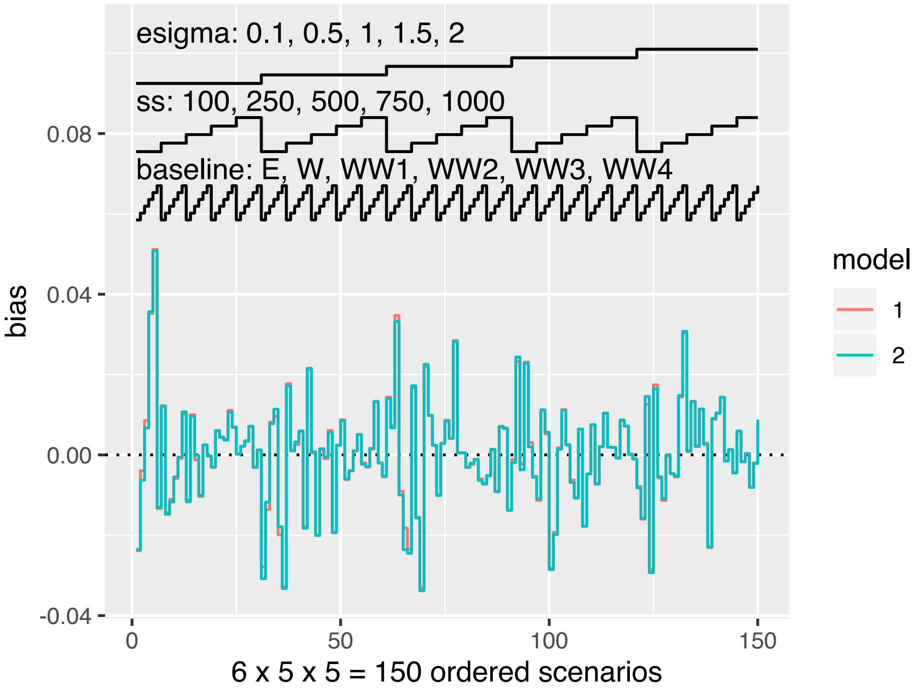


FIGURE 4.3: Nested loop plot for the comparison of bias across all data-generating mechanisms, example dataset included in `rsimsum` on survival modelling

statistics such as bias and coverage probability; uncertainty in the estimation of summary statistics can be included as well, resulting in (some) plots including confidence intervals based on Monte Carlo standard errors. The `autoplot` method returns `ggplot2`-type objects, which can be combined with additional components and easily styled to fit taste and requirements. All the aforementioned plots are presented in practice in the case study of Section 4.6; however, nested loop plots are not really relevant with a small number of scenarios, e.g. in the above-mentioned case study (with only 2 distinct data-generating mechanisms). Therefore, I will use an example dataset included in `rsimsum` to illustrate nested loop plots here: in particular, the example data originates from a simulation study on modelling survival data across 150 distinct scenarios (more details can be obtained by typing `help("nlp", package = "rsimsum")` in the R console). An example nested loop plot (e.g. for bias) is depicted in Figure 4.3.

4.5.2 *INTEREST*

The possibility of replicating, reproducing and independently verifying results from scientific studies is a fundamental aspect of science [74]; as a consequence, several reporting guidelines have emerged (e.g. CONSORT and STROBE [75, 76]). Despite similar calls for harmonised reporting to allow for greater reproducibility in the area of computational science [77] and several articles advocating for more rigour in specific aspects of simulation studies [70, 78–82], design and reporting guidelines for simulation studies are lacking. Morris, White, and Crowther introduced the ADEMP framework described in Section 4.3, aiming to fill precisely that gap [58]: in their review, they outline several ways of reporting results that they observed: including results in text for small simulation studies, tabulating and plotting results, and even the nested-loop plot for fully factorial simulation studies with several data-generating mechanisms [73]. They conclude by arguing that *there is no correct way to present results, but we encourage careful thought to facilitate readability, considering the comparisons that need to be made*. An exciting opportunity to aid the understanding of data visualisations consists of adding a layer of interactivity allowing users to adjust the output to fit their requirements, as outlined by Spiegelhalter *et al.* [83]; the recent advent of tools such as Data-Driven Documents (D^3) [84] and R Shiny [85] has further facilitated the development of interactive visualisations. The increased availability of powerful computational tools not only contributed to the rise in popularity of simulation studies, it also allowed researchers to simulate an ever-growing number of data-generating mechanisms and include several estimands and methods to compare: up to 6×10^{11}, 32, and 33, respectively, in the review or Morris, White, and Crowther [58]. With a large number of data-generating mechanisms, estimands, and methods the analysis and report of results of a simulation study becomes cumbersome. For instance:

1. What results should be the main results to focus on?
2. Which estimands and methods should be included in tables and plots?
3. How could plots or tables illustrate several data-generating mechanisms at once?

To solve this problem, I developed INTEREST (an acronym for *INteractive Tool for Exploring REsults from Simulation sTudies*), an interactive web app to analyse results of Monte Carlo simulation studies. INTEREST requires first uploading a dataset with results from a simulation study; then, it computes summary statistics and creates a variety of tables and plots automatically. The user can vary data-generating mechanisms, estimands, and methods: tables and plots are updated automatically. I will describe the implementation details and features and functionality of INTEREST in the next few Sections.

4.5.2.1 *Implementation of INTEREST*

INTEREST was developed using R [86] and the Shiny framework [85]. Shiny is a framework in R that allows building interactive web apps using R code: the resulting applications can be hosted on a web page, embedded in reports and dashboards, or just run as stand-alone apps. Further details on Shiny are available online (`https://shiny.rstudio.com`). The front-end of INTEREST has been built using the `shinydashboard` package [87]; `shinydashboard` is based upon `AdminLTE` [88], an open-source admin control panel built on top of the Bootstrap framework [89]. The back-end functionality of INTEREST is mostly implemented through the `rsimsum` package (described in Section 4.5.1), with ad-hoc additions for more advanced functionalities. By separating the front-end and the back-end, long-term maintainability should be easier.

INTEREST is available as an online application and as a stand-alone version for offline use. The online version is hosted at `https://interest.shinyapps.io/interest/` and can be accessed via any web browser on any device (desktop computers, laptops, tablets, smartphones, etc.). The stand-alone offline version can be obtained from GitHub at `https://github.com/ellessenne/interest/` and can be run on any desktop computer and laptop, with the only requirement being a functioning installation of R. INTEREST can be installed locally by typing the following commands in the R console:

```
## Install the 'remotes' package if not available
# install.packages("remotes")
remotes::install_github("ellessenne/interest")
```

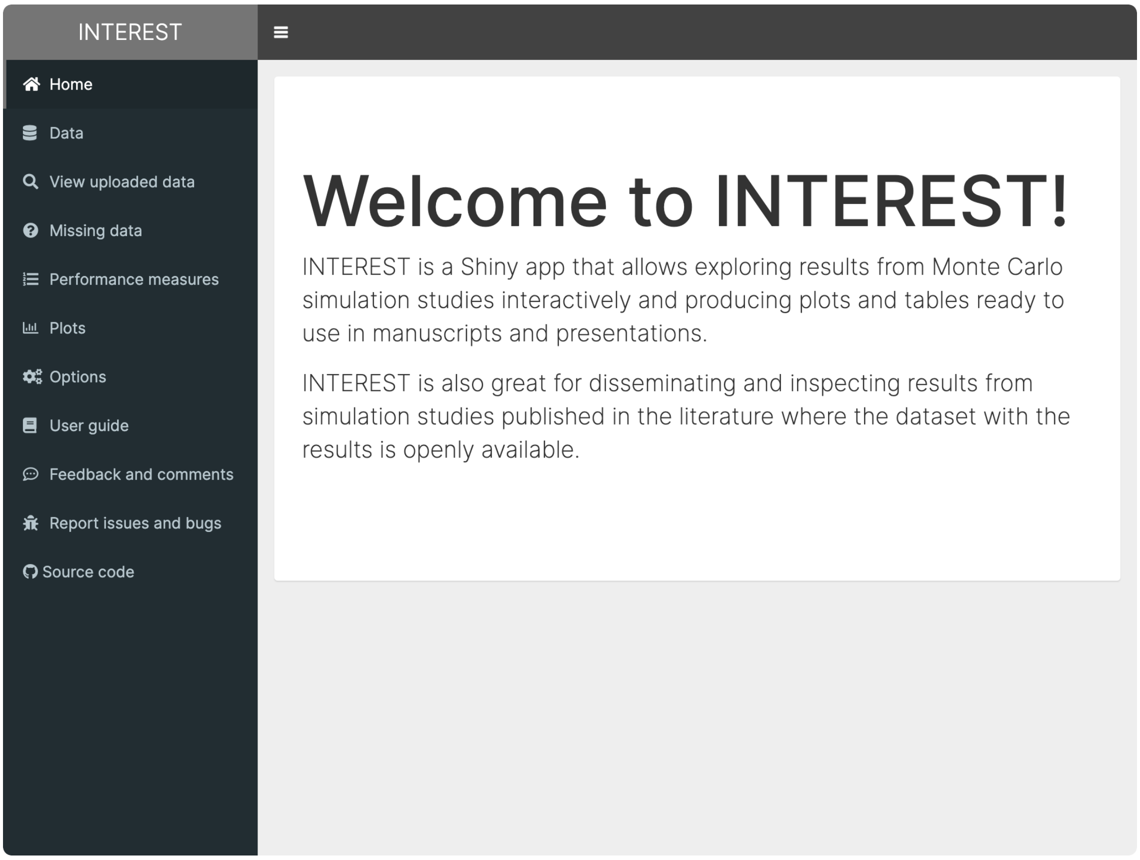
Then, the web app can be launched by typing the following code in the R console:

```
library(interest)
interest()
```

This will launch INTEREST using the default web browser, ready to be utilised.

4.5.2.2 *Features and Functionality of INTEREST*

The main interface of INTEREST is introduced in Figure 4.4. The interface is divided into two sections: the main body on the right-hand side with a sidebar on the left. The body contains the controls and the input/output of the web app; the sidebar contains menu



Welcome to INTEREST!

INTEREST is a Shiny app that allows exploring results from Monte Carlo simulation studies interactively and producing plots and tables ready to use in manuscripts and presentations.

INTEREST is also great for disseminating and inspecting results from simulation studies published in the literature where the dataset with the results is openly available.

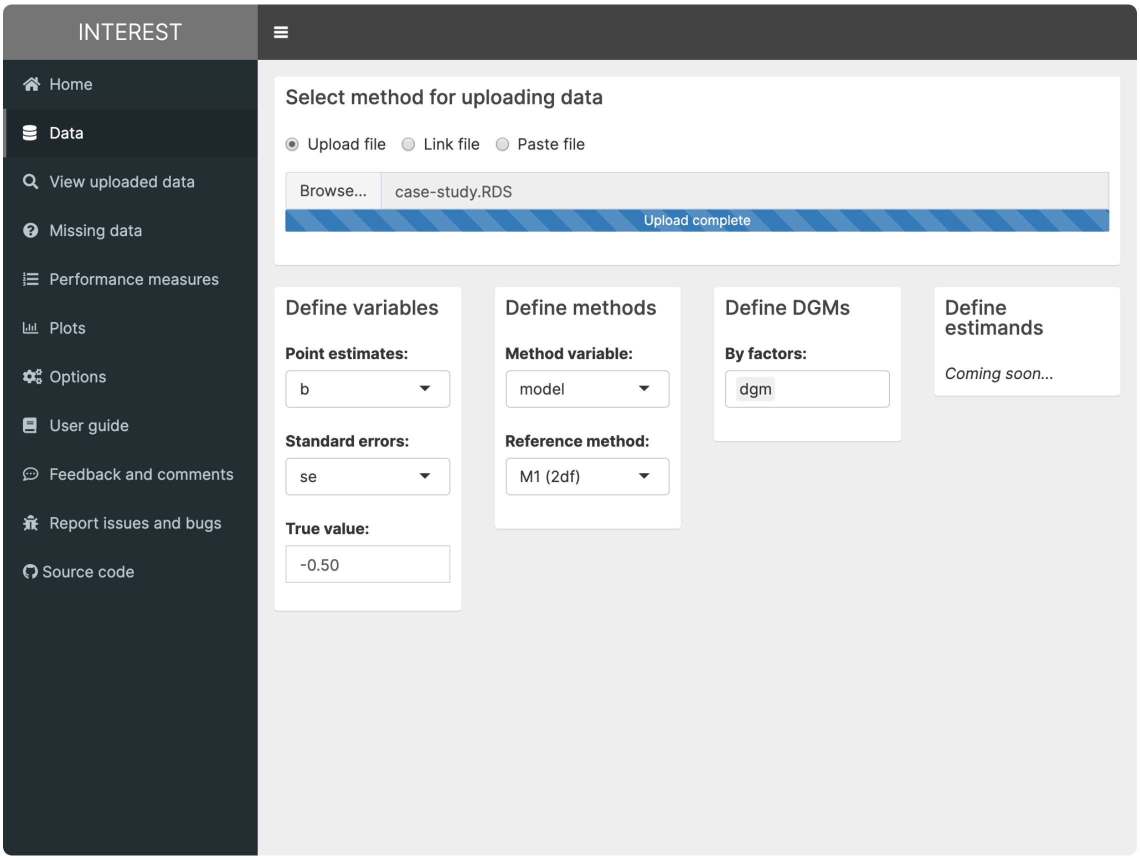
FIGURE 4.4: Homepage and main interface of the INTEREST web app

items that behave like tabs, allowing to navigate the various sections of INTEREST.

The use of INTEREST starts by providing a tidy dataset (variables form columns, observations are in rows [71]) with results from a simulation study via the *Data* tab from the sidebar (Figure 4.5). A dataset can be provided to INTEREST in three different ways:

1. The user can upload a dataset. The uploaded file can be a comma-separated file (`.csv`), a Stata dataset (`.dta`), an SPSS dataset (`.sav`), a SAS dataset (`.sas7bdat`), or an R serialised object (`.rds`); the format will be inferred automatically from the extension of the file, and the whole process is transparent to the user; further to that, the autodetection is case-insensitive. It is also possible to upload compressed files (formats allowed: `.gz`, `.bz2`, `.xz`, or `.zip`) that are automatically decompressed;
2. The user can provide a URL link to a dataset hosted elsewhere. All considerations relative to the file format from the previous point are also valid here;
3. Finally, the user can paste a dataset (e.g. from Microsoft Excel) in a text box. The pasted data is assumed to be tab-separated.

Once a dataset has been uploaded via one of the three methods outlined, the user will



Select method for uploading data

● Upload file ○ Link file ○ Paste file

case-study.RDS

Define variables

Point estimates: b

Standard errors: se

True value: -0.50

Define methods

Method variable: model

Reference method: M1 (2df)

Define DGMs

By factors: dgm

Define estimands

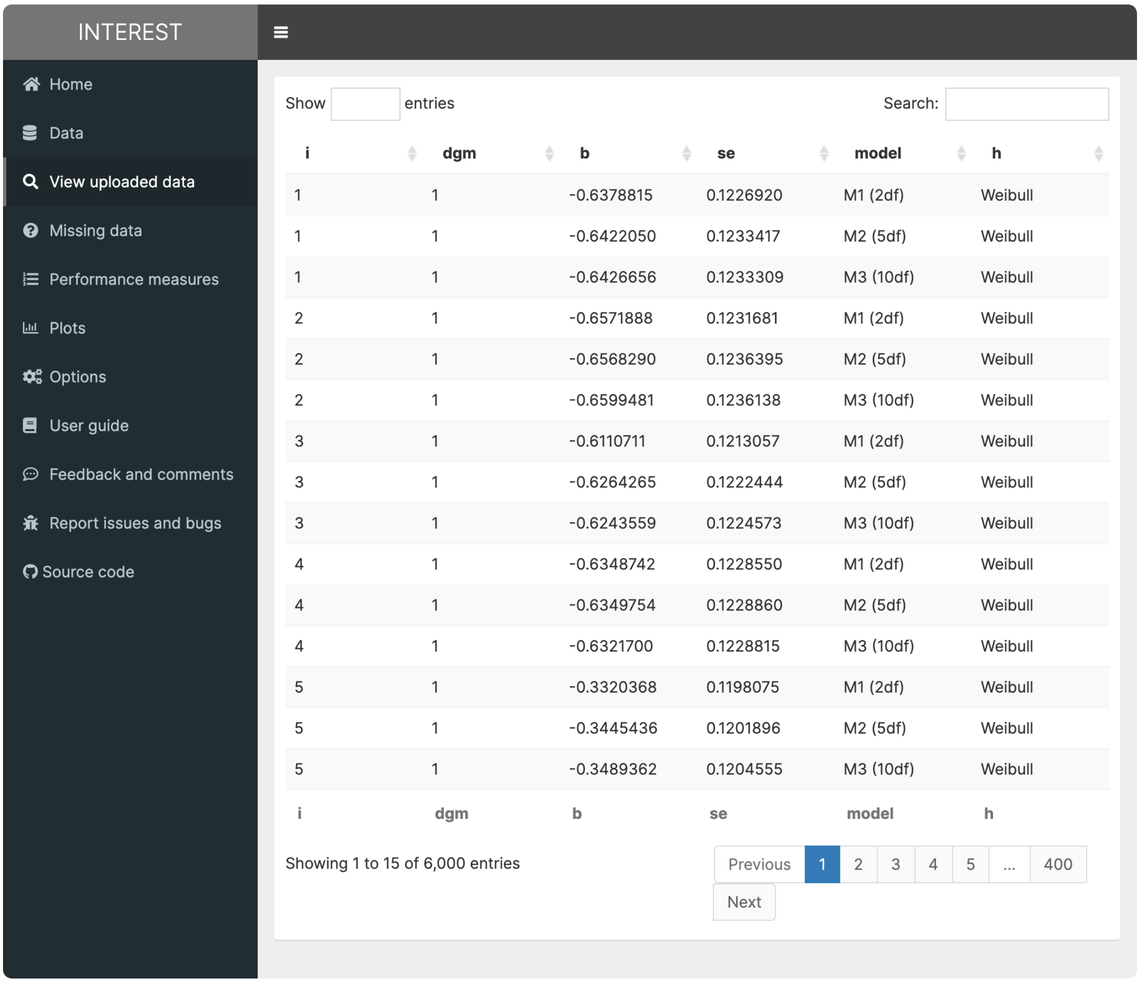
Coming soon...

FIGURE 4.5: App interface to upload and define data for INTEREST

have to define the variables required by INTEREST and some optional variables. The names of each column (i.e. variable) from the uploaded dataset automatically populate a set of select list inputs to assist the user. A variable defining a point estimate from the simulation study and a variable representing the standard error of such estimates are required, and the user has to define the true value of the estimand of interest as well. Additionally, a user can define a variable representing methods being compared with the current simulation study (and choose the default one), and one or more variables defining data-generating mechanisms (e.g. sample size, true correlation, true baseline hazard function for survival models, etc.).

The *View uploaded data* tab in INTEREST displays the dataset uploaded by the user using an interactive table that can be sorted and filtered at will by the user (Figure 4.6). As a good practice, it is recommended to check that the uploaded dataset is correct before continuing with the analysis and any visual exploration.

INTEREST includes a section for exploring missingness of estimates and/or standard errors from each replication of a simulation study. Missing values need to be carefully



i	dgm	b	se	model	h
1	1	-0.6378815	0.1226920	M1 (2df)	Weibull
1	1	-0.6422050	0.1233417	M2 (5df)	Weibull
1	1	-0.6426656	0.1233309	M3 (10df)	Weibull
2	1	-0.6571888	0.1231681	M1 (2df)	Weibull
2	1	-0.6568290	0.1236395	M2 (5df)	Weibull
2	1	-0.6599481	0.1236138	M3 (10df)	Weibull
3	1	-0.6110711	0.1213057	M1 (2df)	Weibull
3	1	-0.6264265	0.1222444	M2 (5df)	Weibull
3	1	-0.6243559	0.1224573	M3 (10df)	Weibull
4	1	-0.6348742	0.1228550	M1 (2df)	Weibull
4	1	-0.6349754	0.1228860	M2 (5df)	Weibull
4	1	-0.6321700	0.1228815	M3 (10df)	Weibull
5	1	-0.3320368	0.1198075	M1 (2df)	Weibull
5	1	-0.3445436	0.1201896	M2 (5df)	Weibull
5	1	-0.3489362	0.1204555	M3 (10df)	Weibull
i	dgm	b	se	model	h

Showing 1 to 15 of 6,000 entries

FIGURE 4.6: App interface to inspect the dataset uploaded to INTEREST

explored and handled at the initial stage of any analysis, and may originate as a software failure (in which case the code should be made more robust to ensure fewer or no failures). Conversely, missing data may arise as a consequence of characteristics of the simulated data, yielding to non-convergence of the estimation procedures. In other words, missing values may not be missing completely at random [58, 90].

The missing data functionality is based on the R package `naniar` [91], and can be accessed via the *Missing data* tab. It comprises visual and tabular summaries; missing data visualisations available in INTEREST are the following:

- Bar plots of number (or proportion) of missing values by method and data-generating mechanism (if defined). Number and proportion of missing values are produced for each variable included in the data uploaded to INTEREST;
- A plot to visualise missing data in the whole dataset;
- A scatter plot with missing status depicted with different colours; in order to be able to plot missing values, they are replaced with values 10% lower than the minimum value in that variable. This plot allows identifying trends and patterns between variables in missing values (e.g. all estimates with a very large point estimate have a missing standard error);
- A heat map with methods on the X-axis and the remaining variables on the Y-axis, with the colour fill representing the percentage of missingness in each tile.

Each plot can be further customised and exported (e.g. for use in slides and reports); more details are discussed later on when introducing the plotting functionality. Finally, INTEREST computes and outputs a table with the number, proportion, and the cumulative number of missing values per variable stratifying by method and data-generating mechanisms; the table can be easily exported in LaTeX format for further use.

Summary statistics are computed automatically as soon as the user defines the required variables in the *Data* tab, and displayed in the *Performance measures* tab (Figure 4.7). The summary statistics computed by INTEREST are all the summary statistics supported by `rsimsum`, and are included in Table 4.1; Monte Carlo standard errors are computed and returned by default. Estimated summary statistics are tabulated for a given data-generating mechanism, which can be selected by the user; furthermore, results can be exported in two ways:

1. Export the table in LaTeX format, e.g. for use in reports and articles. It is possible to customise the caption of the table;
2. Export summary statistics as a dataset, e.g. for use in other software packages. It is possible to export the table of summary statistics as displayed in INTEREST or in tidy format, and in a variety of formats: comma-separated (`.csv`), tab-separated (`.tsv`), R (`.rds`), Stata (`.dta`), SPSS (`.sav`), or SAS (`.sas7bdat`).

INTEREST can also automatically produce a variety of plots to visualise results from simulation studies. Plots produced by INTEREST can be categorised into two broad groups: plots of estimated values (and standard errors) and plots of summary statistics. The app interface and sample plots are depicted in Figures 4.8 and 4.9.

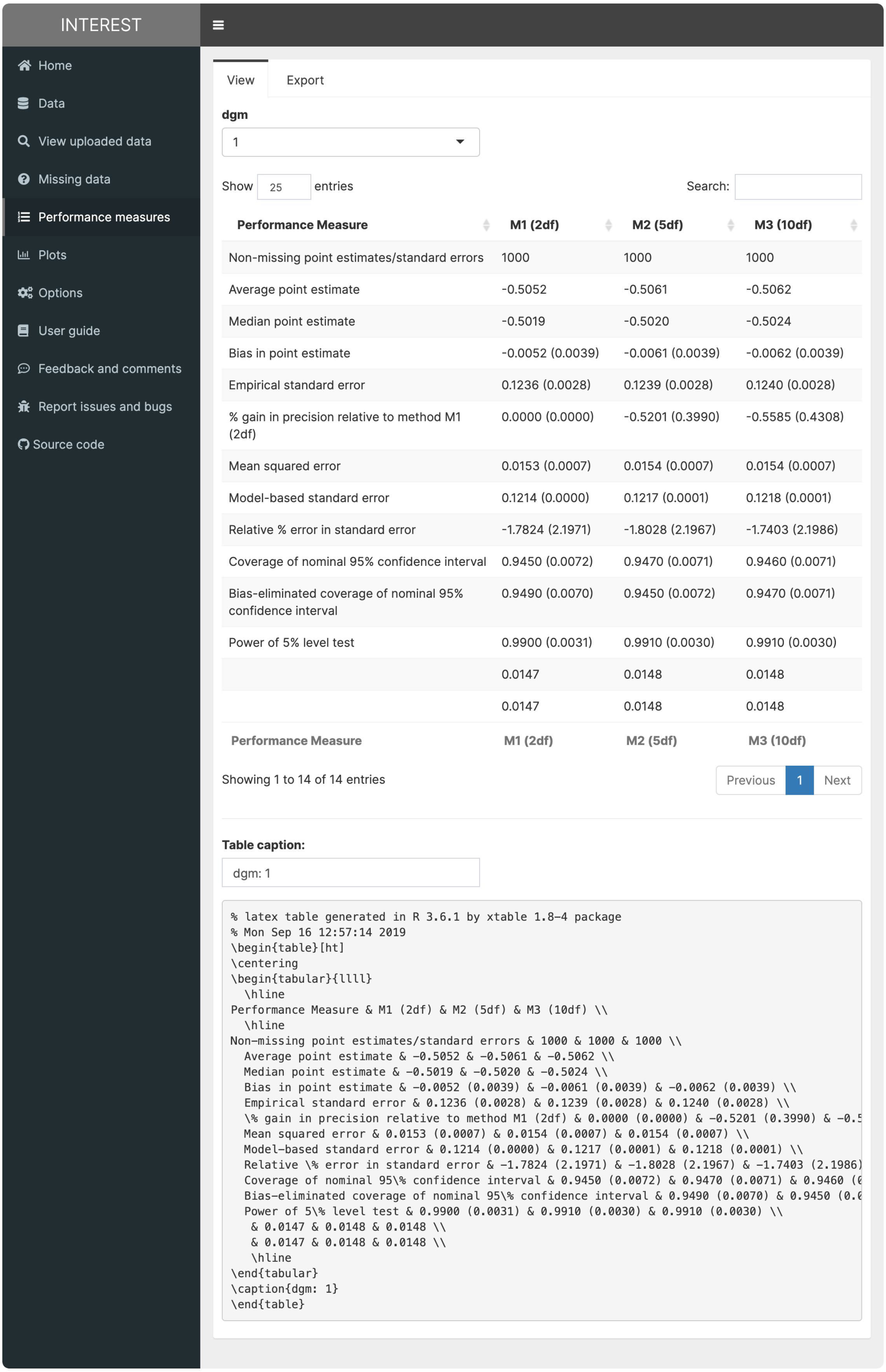
Plots for estimated values and standard errors are:

- Scatter plots with a method-wise comparison of point estimates (or standard errors);
- Bland-Altman plots with a method-wise comparison of point estimates (or standard errors);
- Ridgelines plots with the method-wise comparison of the distribution of point estimates (or standard errors).

Conversely, the following plots are supported for performance measures:

- Plots of summary statistics with confidence intervals based on Monte Carlo standard errors. There are two types of this plot: forest plots and lolly plots;
- Heat plots of summary statistics: these plots are mosaic plots where the factor on the x-axis is represented by methods (if defined) and the factor on the y-axis is represented by a DGM, as selected by the user;
- Zipper plots to visually explain the summary statistic coverage by plotting the confidence intervals directly. This visualisation is described in more detail elsewhere [58].

Further to that, plots can be customised and exported for use in manuscript, reports, presentations via the *Options* tab (Figure 4.10). In terms of customisation, it is possible to define custom labels for the x-axis and the y-axis and to change the overall appearance of the plot by applying one of the predefined themes.



dgm

1

Show 25 entries

Performance Measure	M1 (2df)	M2 (5df)	M3 (10df)
Non-missing point estimates/standard errors	1000	1000	1000
Average point estimate	-0.5052	-0.5061	-0.5062
Median point estimate	-0.5019	-0.5020	-0.5024
Bias in point estimate	-0.0052 (0.0039)	-0.0061 (0.0039)	-0.0062 (0.0039)
Empirical standard error	0.1236 (0.0028)	0.1239 (0.0028)	0.1240 (0.0028)
% gain in precision relative to method M1 (2df)	0.0000 (0.0000)	-0.5201 (0.3990)	-0.5585 (0.4308)
Mean squared error	0.0153 (0.0007)	0.0154 (0.0007)	0.0154 (0.0007)
Model-based standard error	0.1214 (0.0000)	0.1217 (0.0001)	0.1218 (0.0001)
Relative % error in standard error	-1.7824 (2.1971)	-1.8028 (2.1967)	-1.7403 (2.1986)
Coverage of nominal 95% confidence interval	0.9450 (0.0072)	0.9470 (0.0071)	0.9460 (0.0071)
Bias-eliminated coverage of nominal 95% confidence interval	0.9490 (0.0070)	0.9450 (0.0072)	0.9470 (0.0071)
Power of 5% level test	0.9900 (0.0031)	0.9910 (0.0030)	0.9910 (0.0030)
	0.0147	0.0148	0.0148
	0.0147	0.0148	0.0148
Performance Measure	M1 (2df)	M2 (5df)	M3 (10df)

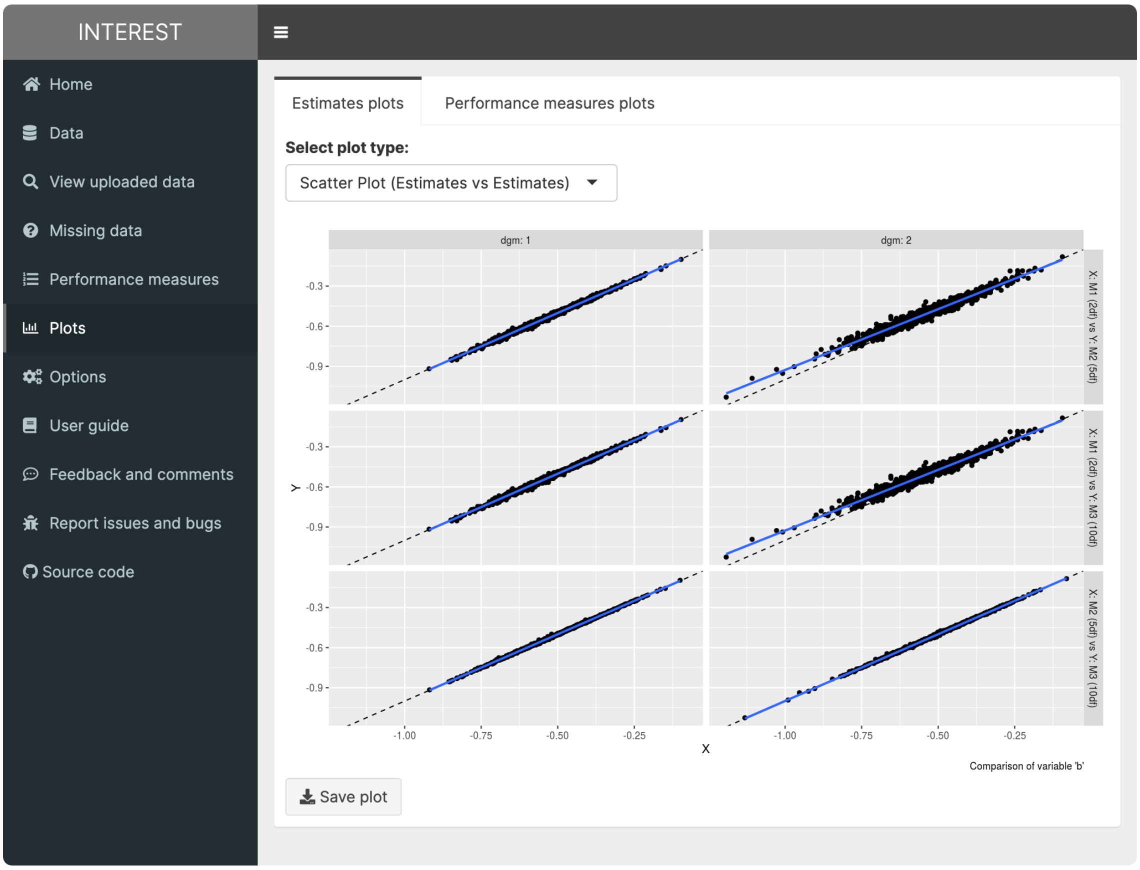
Showing 1 to 14 of 14 entries

Table caption:

dgm: 1

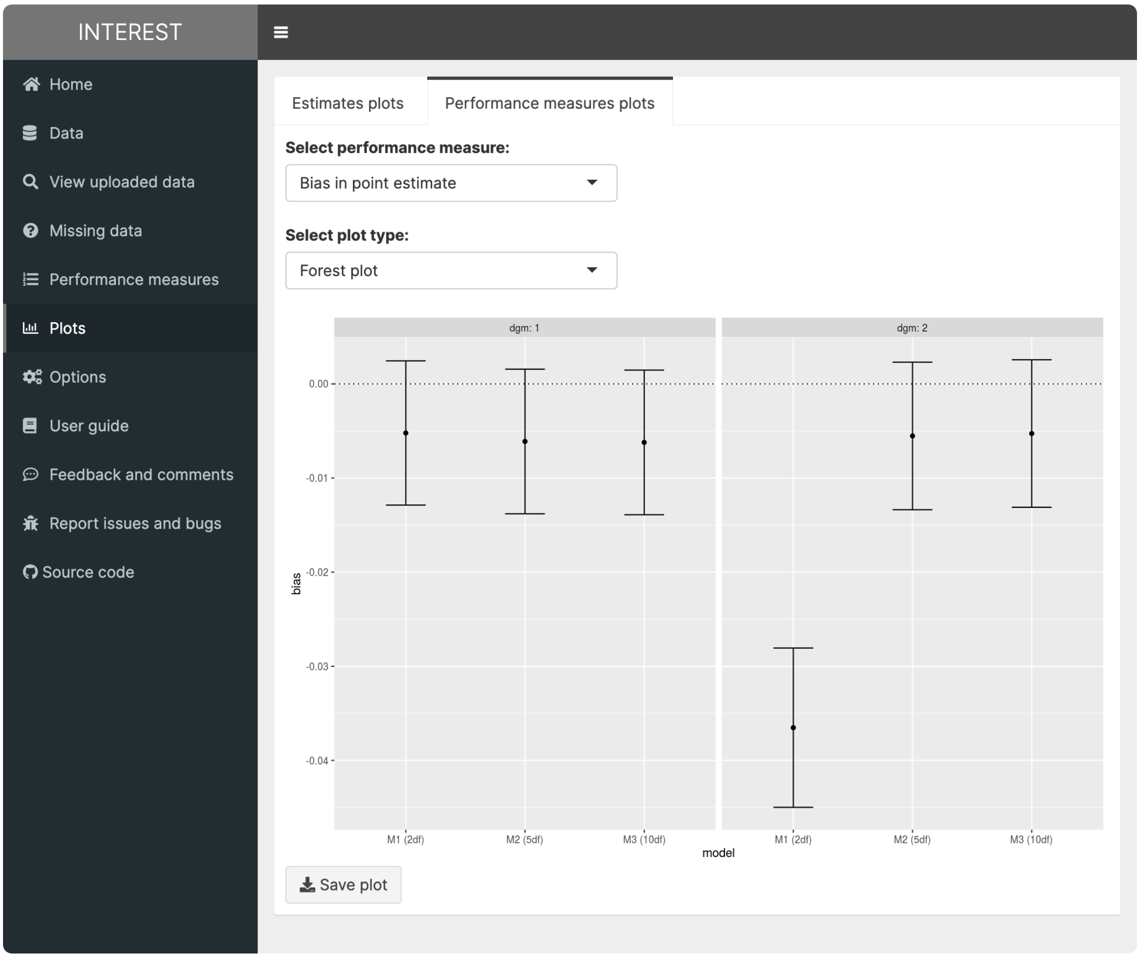
```
% latex table generated in R 3.6.1 by xtable 1.8-4 package
% Mon Sep 16 12:57:14 2019
\begin{table}[ht]
\centering
\begin{tabular}{llll}
  \hline
Performance Measure & M1 (2df) & M2 (5df) & M3 (10df) \\
  \hline
Non-missing point estimates/standard errors & 1000 & 1000 & 1000 \\
  Average point estimate & -0.5052 & -0.5061 & -0.5062 \\
  Median point estimate & -0.5019 & -0.5020 & -0.5024 \\
  Bias in point estimate & -0.0052 (0.0039) & -0.0061 (0.0039) & -0.0062 (0.0039) \\
  Empirical standard error & 0.1236 (0.0028) & 0.1239 (0.0028) & 0.1240 (0.0028) \\
  \% gain in precision relative to method M1 (2df) & 0.0000 (0.0000) & -0.5201 (0.3990) & -0.5
  Mean squared error & 0.0153 (0.0007) & 0.0154 (0.0007) & 0.0154 (0.0007) \\
  Model-based standard error & 0.1214 (0.0000) & 0.1217 (0.0001) & 0.1218 (0.0001) \\
  Relative \% error in standard error & -1.7824 (2.1971) & -1.8028 (2.1967) & -1.7403 (2.1986)
  Coverage of nominal 95\% confidence interval & 0.9450 (0.0072) & 0.9470 (0.0071) & 0.9460 (0
  Bias-eliminated coverage of nominal 95\% confidence interval & 0.9490 (0.0070) & 0.9450 (0.0
  Power of 5\% level test & 0.9900 (0.0031) & 0.9910 (0.0030) & 0.9910 (0.0030) \\
   & 0.0147 & 0.0148 & 0.0148 \\
   & 0.0147 & 0.0148 & 0.0148 \\
   \hline
\end{tabular}
\caption{dgm: 1}
\end{table}
```

FIGURE 4.7: App interface to explore and export summary statistics of a Monte Carlo simulation study using tabular representations



Select plot type: Scatter Plot (Estimates vs Estimates)

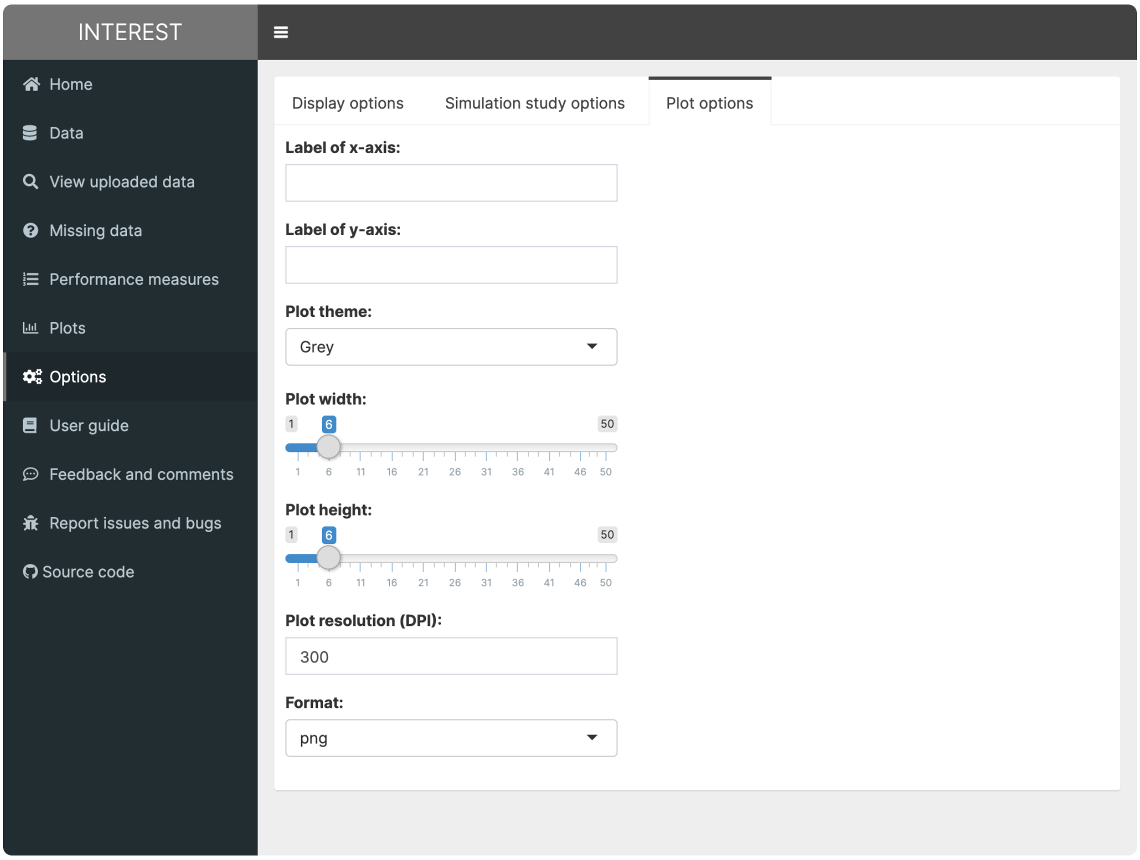
FIGURE 4.8: App interface to plot estimates (or standard errors), with a sample scatter plot of point estimates that is produced by INTEREST



Select performance measure: Bias in point estimate

Select plot type: Forest plot

FIGURE 4.9: App interface to plot summary statistics, with a sample forest plot for bias that is produced by INTEREST



Label of x-axis:

Label of y-axis:

Plot theme: Grey

Plot width: 6

Plot height: 6

Plot resolution (DPI): 300

Format: png

FIGURE 4.10: App interface to customise plot appearance and exporting options

In terms of exporting plots, it is possible to define the width, height, and resolution of the plot to export, and the format of the file to export. To suit a wide variety of possible use cases, INTEREST supports several image formats: among others, `.pdf`, `.png`, `.svg`, and `.eps`.

In conclusion, INTEREST allows researchers to upload a dataset with the results of their Monte Carlo simulation study obtaining summary statistics in a quick and straightforward way. This is very appealing, especially with simulation studies with several data-generating mechanisms where it could be confusing to investigate all scenarios at once. Using the app it is possible to vary data-generating mechanisms and obtain updated tables and plots in real-time, therefore allowing to quickly iterate and take into consideration all possible scenarios.

One of the intended usage scenarios for INTEREST consists of supplementing reporting of simulation studies. This is especially useful with large simulation studies, where it is most cumbersome to summarise all results in a manuscript: it is common to include in the main manuscript only a subset of results for the sake of brevity. The remaining results are then relegated to supplementary material, web appendices, or not published at all - undermining dissemination and replicability of a study.

Being now common practice to publish the code produced to run a simulation study, one could publish the dataset with the results alongside the code used to obtain it. That dataset could then be uploaded to INTEREST by readers, who could then explore the full results of the study as they wish. Given the ubiquity of web services like GitHub (`https://github.com`) and data-sharing repositories such as Zenodo (`https://zenodo.org/`), this practice is encouraged and strongly recommended.

As outlined above, Monte Carlo simulation studies are too often poorly analysed and reported [58]. Given the increased use in methodological statistical research, I believe that INTEREST could improve reporting and disseminating results from simulation studies to a large extent. To the best of my knowledge, there is no similar application readily available for researchers to use.

4.6 A Case Study Using Flexible Parametric Survival Models

In this Section I will present a case study that involves planning a simulation study using the ADEMP structure introduced in section 4.3 and analysing its results using the software introduced in Section 4.5. For illustrative purposes, I will mimic the settings of a published Monte Carlo simulation study [28].

The *aim* of the study consists of investigating the robustness of flexible parametric survival models (as introduced in Section 2.2.4.3) to the modelling choice of how many degrees of freedom are used to model the baseline hazard. It is therefore possible to broadly categorise the aims of this simulation as (1) comparing competing methods and (2) evaluating if and when they break.

I simulate survival data for 300 study subjects using the approaches outlined in Section 4.4, with a binary covariate (e.g. a treatment variable) simulated from a Bernoulli distribution with success probability equal to 50%. The binary covariate has an associated coefficient (e.g. a log treatment effect) of -0.50. Censoring is simulated by applying administrative censoring after 10 years of follow-up. The shape of the baseline hazard function will

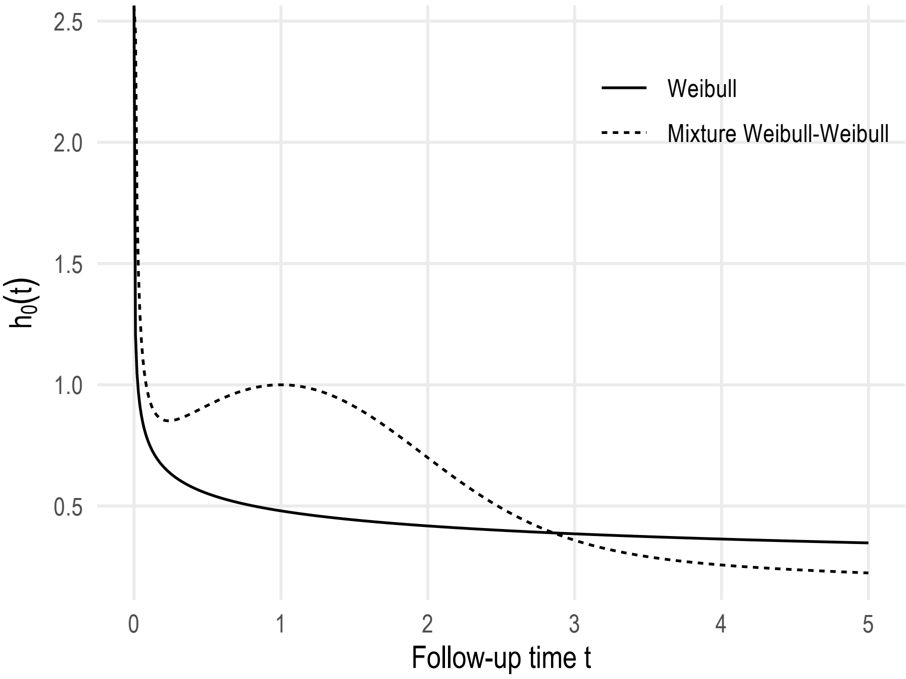


FIGURE 4.11: Baseline hazard functions assumed for the simulated scenario of the case study on flexible parametric survival models. The Weibull baseline hazard is assuming $\lambda = 0.60, \gamma = 0.80$, while the mixture Weibull-Weibull baseline hazard is assuming $\lambda = \{1.00, 1.00\}, \gamma = \{1.50, 0.50\}, \pi = 0.50$

vary, to assess whether the complexity of the underlying hazard function affects the robustness of the models. Specifically, I simulate two scenarios: a monotonic baseline hazard function simulated from a Weibull distribution with $\lambda = 0.60$ and $\gamma = 0.80$, and a baseline hazard function with turning points simulated from a mixture Weibull distribution with $\lambda = \{1.00, 1.00\}$, $\gamma = \{1.50, 0.50\}$, and $\pi = 0.50$ (Figure 4.11). This leads to 2 distinct *data-generating mechanisms*.

The *estimand* of interest is the regression coefficient θ associated with the binary covariate, i.e. the log treatment effect (as an estimate of relative risk).

The *methods* included in this comparison are flexible parametric survival models following the formulation of Royston and Parmar [24], as introduced in Section 2.2.4.3. The fitted models only vary in terms of the number of degrees of freedom used to model the baseline hazard: 2, 5, and 10, respectively.

Finally, the *performance measures* of interest are bias, mean squared error, and coverage probability. Rutherford *et al.* run 1,000 replications of each simulated scenario, which would yield an expected Monte Carlo error for bias of 0.01 or lower assuming a variance of the estimated regression coefficient $\text{Var}(\hat{\theta}) \leq 0.1$. Analogously, the expected Monte

Carlo standard error for coverage, assuming a worst-case scenario of coverage = 0.50, would be 0.02. Therefore, I run 1,000 replications per scenario as well.

The simulation study is coded using R, and all the code required to replicate the analysis is included in Appendix C.

After running the simulation study, its results can be analysed using the software introduced in Section 4.5 as follows.

First, I load the `rsimsum` R package:

```
library(rsimsum)
```

If `rsimsum` is not installed, it can be installed with:

```
install.packages("rsimsum")
```

The dataset with the results of the simulation study contains 6,000 rows and 6 columns and it is named `db`; describing the content of the dataset:

```
dplyr::glimpse(db)
```

```
# Observations: 6,000
# Variables: 6
# $ i     <int> 1, 1, 1, 2, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 5, 6, 6, 6, 7, 7, 7...
# $ dgm   <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
# $ b     <dbl> -0.6378815, -0.6422050, -0.6426656, -0.6571888, -0.6568290, -...
# $ se    <dbl> 0.1226920, 0.1233417, 0.1233309, 0.1231681, 0.1236395, 0.1236...
# $ model <fct> M1 (2df), M2 (5df), M3 (10df), M1 (2df), M2 (5df), M3 (10df),...
# $ h     <fct> Weibull, Weibull, Weibull, Weibull, Weibull, Weibull, Weibull...
```

The `dgm` column identifies the two data-generating mechanisms, with `1` identifying the scenario with a Weibull baseline hazard function and `2` the other scenario with a mixture Weibull-Weibull baseline hazard function.

Then, I use the `rsimsum::simsum` function to define the simulation study to analyse. I set the argument `x = TRUE` as it will be required for plotting later on:

```
simstudy <- rsimsum::simsum(
  data = db,
  estvarname = "b", se = "se", true = -0.50,
  methodvar = "model", by = "dgm",
  x = TRUE
)
simstudy
```

```
# Summary of a simulation study with a single estimand.
#
# Method variable: model
#   Unique methods: M1 (2df), M2 (5df), M3 (10df)
#   Reference method: M1 (2df)
#
# By factors: dgm
#
# Monte Carlo standard errors were computed.
```

Printing the `simstudy` object returns some information on the characteristics of the simulation study being analysed. For instance, the column that defines the methods included in this comparison is `method`, and the possible methods are `M1 (2df)`, `M2 (5df)`, `M3 (10df)`. The column that defines the two data-generating mechanisms is `dgm`. The flexible parametric model with 2 df (denoted as `M1 (2df)`) was automatically selected as the reference method, given that no `ref` argument was passed to the `rsimsum::simsum` function.

The next step consists in summarising the simulation study, using the `summary` method. I select the following summary statistics via the `stats` argument: the median estimate (`thetamedian`), the median squared standard error (`se2median`), bias (`bias`), mean squared error (`mse`), and coverage probability (`cover`):

```
summary(simstudy,
  stats = c("thetamedian", "se2median", "bias", "mse", "cover")
)
```

```
# Values are:
#   Point Estimate (Monte Carlo Standard Error)
#
```

```
# Median point estimate:
#  dgm M1 (2df) M2 (5df) M3 (10df)
#    1  -0.5019  -0.5020   -0.5024
#    2  -0.5326  -0.4992   -0.4986
#
# Median variance:
#  dgm M1 (2df) M2 (5df) M3 (10df)
#    1   0.0147   0.0148    0.0148
#    2   0.0149   0.0147    0.0147
#
# Bias in point estimate:
#  dgm         M1 (2df)         M2 (5df)        M3 (10df)
#    1 -0.0052 (0.0039) -0.0061 (0.0039) -0.0062 (0.0039)
#    2 -0.0365 (0.0043) -0.0055 (0.0040) -0.0053 (0.0040)
#
# Mean squared error:
#  dgm        M1 (2df)        M2 (5df)       M3 (10df)
#    1 0.0153 (0.0007) 0.0154 (0.0007) 0.0154 (0.0007)
#    2 0.0200 (0.0010) 0.0160 (0.0008) 0.0160 (0.0008)
#
# Coverage of nominal 95% confidence interval:
#  dgm        M1 (2df)        M2 (5df)       M3 (10df)
#    1 0.9450 (0.0072) 0.9470 (0.0071) 0.9460 (0.0071)
#    2 0.9230 (0.0084) 0.9520 (0.0068) 0.9530 (0.0067)
```

The median estimate for the parameter of interest is close to the true value of -0.50 across all models and simulated scenarios, with the exception of the model with 2 degrees of freedom in the DGM with a mixture Weibull-Weibull baseline hazard: in that setting, the median estimate resulted to be considerably lower than the more flexible models (with a higher number of degrees of freedom). The model-wise comparison of point estimates from each model using scatter plots can be obtained via the `autoplot` method and requesting the `type = "est"` plot:

```
autoplot(simstudy, type = "est")
```

The resulting plot is depicted in Figure 4.12. Analogously, as outlined in Section 4.5.1, `rsimsum` can produce Bland-Altman and ridgeline plots (Figures 4.13 and 4.14, respectively) for point estimates.

```
autoplot(simstudy, type = "est_ba")
```

```
autoplot(simstudy, type = "est_ridge")
```

`rsimsum` can also produce similar plots for the estimated standard errors. For instance, a scatter plot is included as Figure 4.15, showing how the estimated standard errors from each model-scenario combination are very similar: the estimated regression line (the blue line) overlaps quite well with the diagonal line (dashed black line). This plot can be produced as follows:

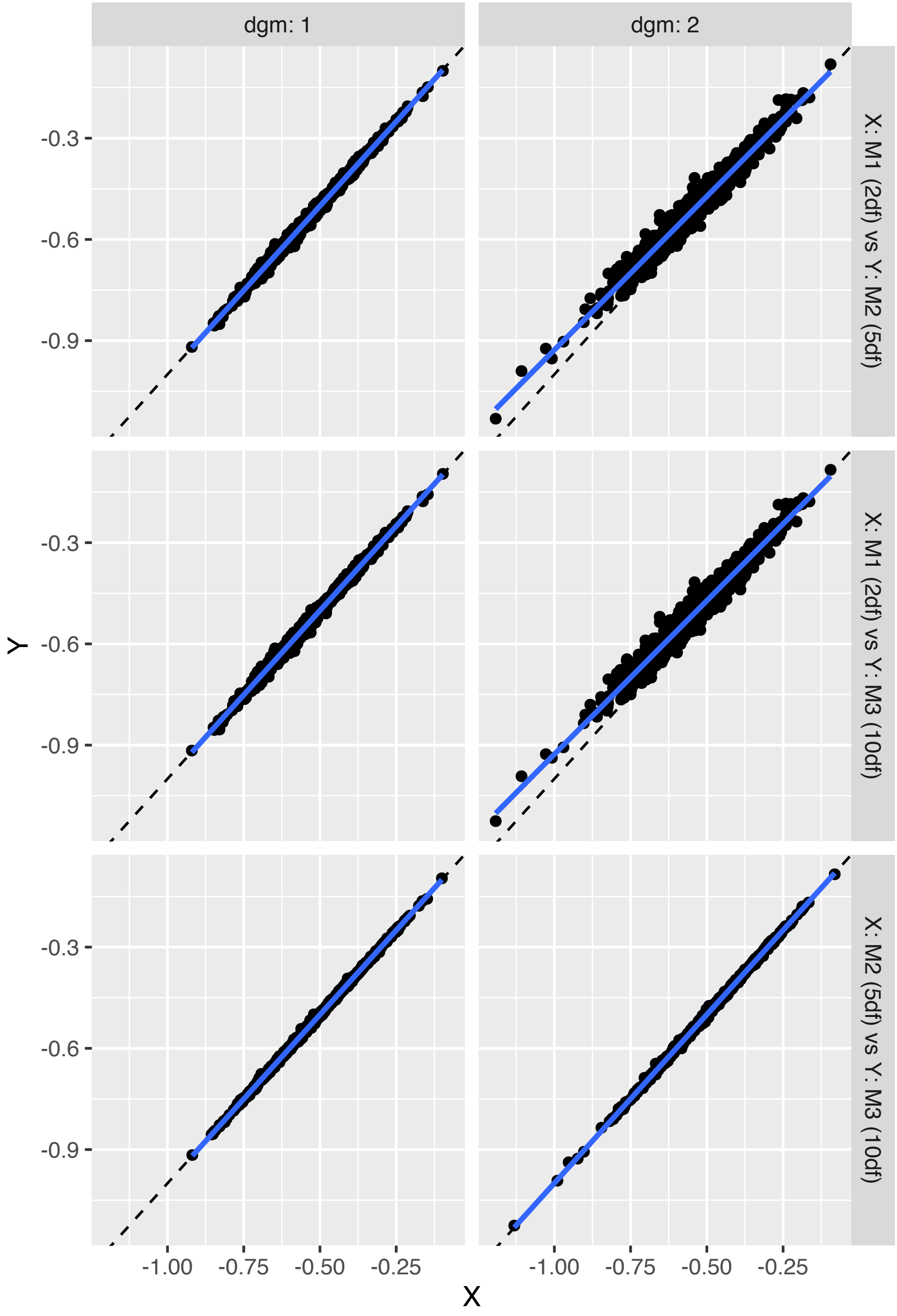
```
autoplot(simstudy, type = "se")
```

Next, the estimated bias by method and data-generating mechanism can be found in the summary above and can be plotted using `rsimsum` as follows.

```
p1 <- autoplot(summary(simstudy), type = "forest", stats = "bias")
p2 <- autoplot(summary(simstudy), type = "lolly", stats = "bias")
cowplot::plot_grid(p1, p2,
  labels = LETTERS,
  ncol = 1,
  align = "hv",
  axis = "lrtb"
)
```

By analysing the forest plot (Figure 4.16, Panel A) or the lolly plot (Figure 4.16, Panel B) it is possible to identify the scenarios where bias is significant, as the uncertainty in estimating the summary statistic is incorporated through confidence intervals based on Monte Carlo standard errors. In particular, it is possible to appreciate how the flexible parametric model with 2 degrees of freedom yields significant bias when the true baseline hazard followed a mixture Weibull-Weibull distribution.

The remaining summary statistics can be plotted analogously, and are included as Figures 4.17 and 4.18:



Comparison of variable 'b'

FIGURE 4.12: Method-wise comparison of point estimates from each simulated scenario by data-generating mechanism using scatter plots, case study using flexible parametric survival models.

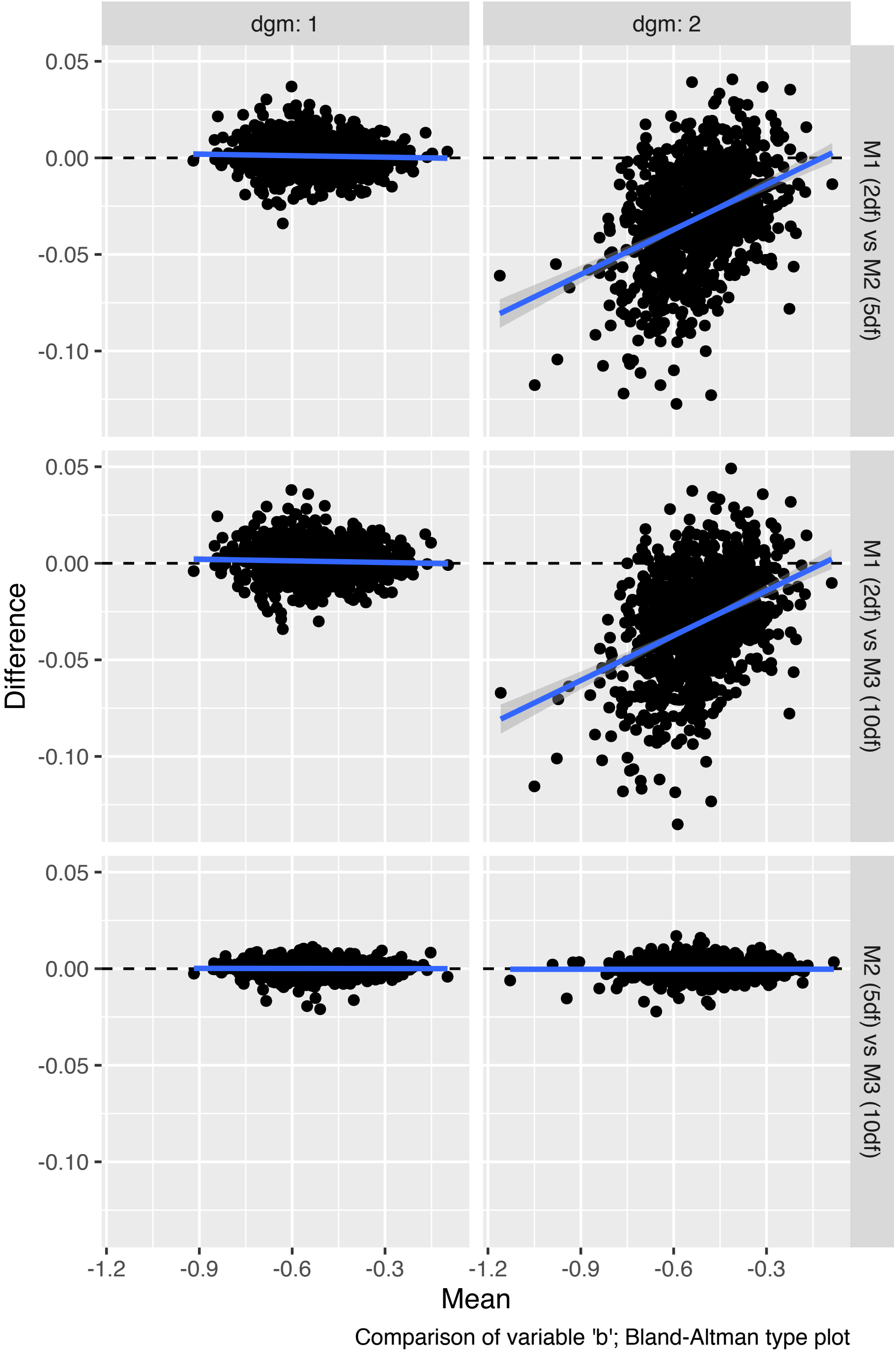


FIGURE 4.13: Method-wise comparison of point estimates from each simulated scenario by data-generating mechanism using Bland-Altman plots, case study using flexible parametric survival models.

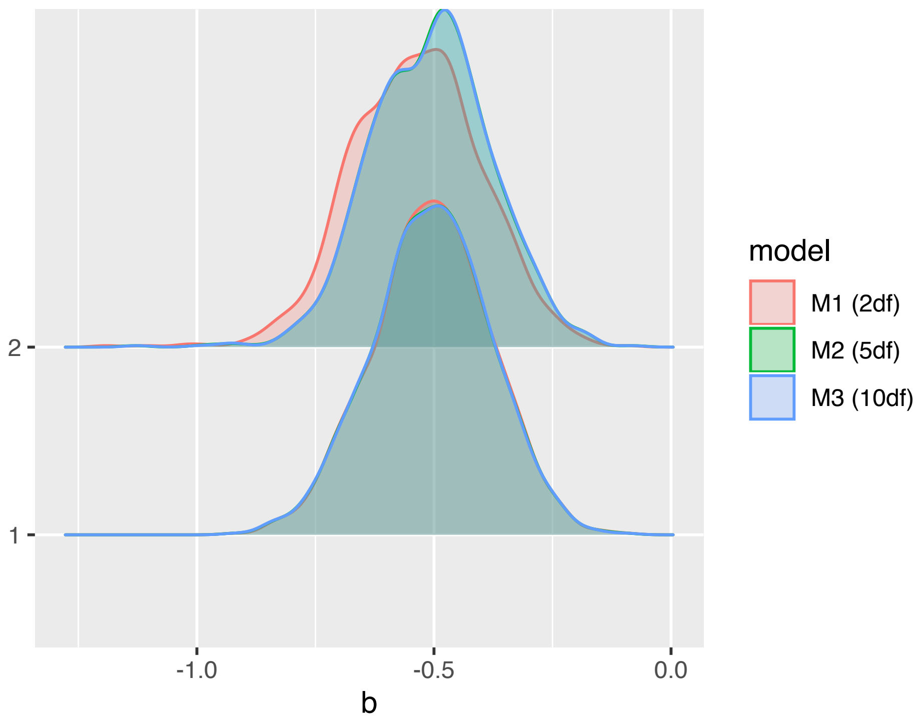


FIGURE 4.14: Method-wise comparison of point estimates from each simulated scenario by data-generating mechanism using ridgeline plots, case study using flexible parametric survival models.

```
p3 <- autoplot(simstudy, type = "forest", stats = "mse")
p4 <- autoplot(simstudy, type = "lolly", stats = "mse")
cowplot::plot_grid(p3, p4,
  labels = LETTERS,
  ncol = 1,
  align = "hv",
  axis = "lrtb"
)

p5 <- autoplot(summary(simstudy), type = "forest", stats = "cover")
p6 <- autoplot(summary(simstudy), type = "lolly", stats = "cover")
cowplot::plot_grid(p5, p6,
  labels = LETTERS,
  ncol = 1,
  align = "hv",
  axis = "lrtb"
)
```

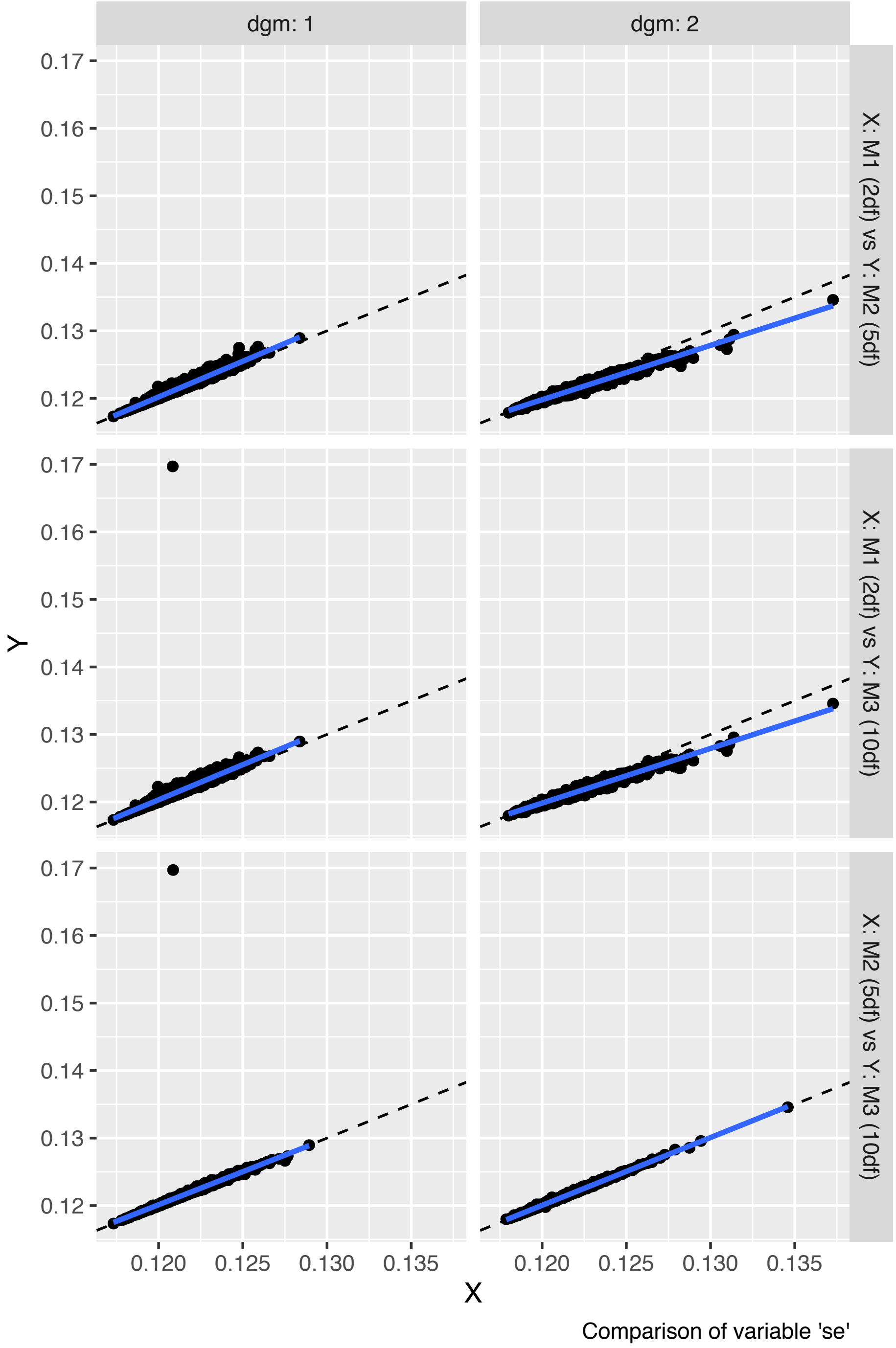


Figure 4.15: Method-wise comparison of estimated standard errors from each simulated scenario by data-generating mechanism using scatter plots, case study using flexible parametric survival models.

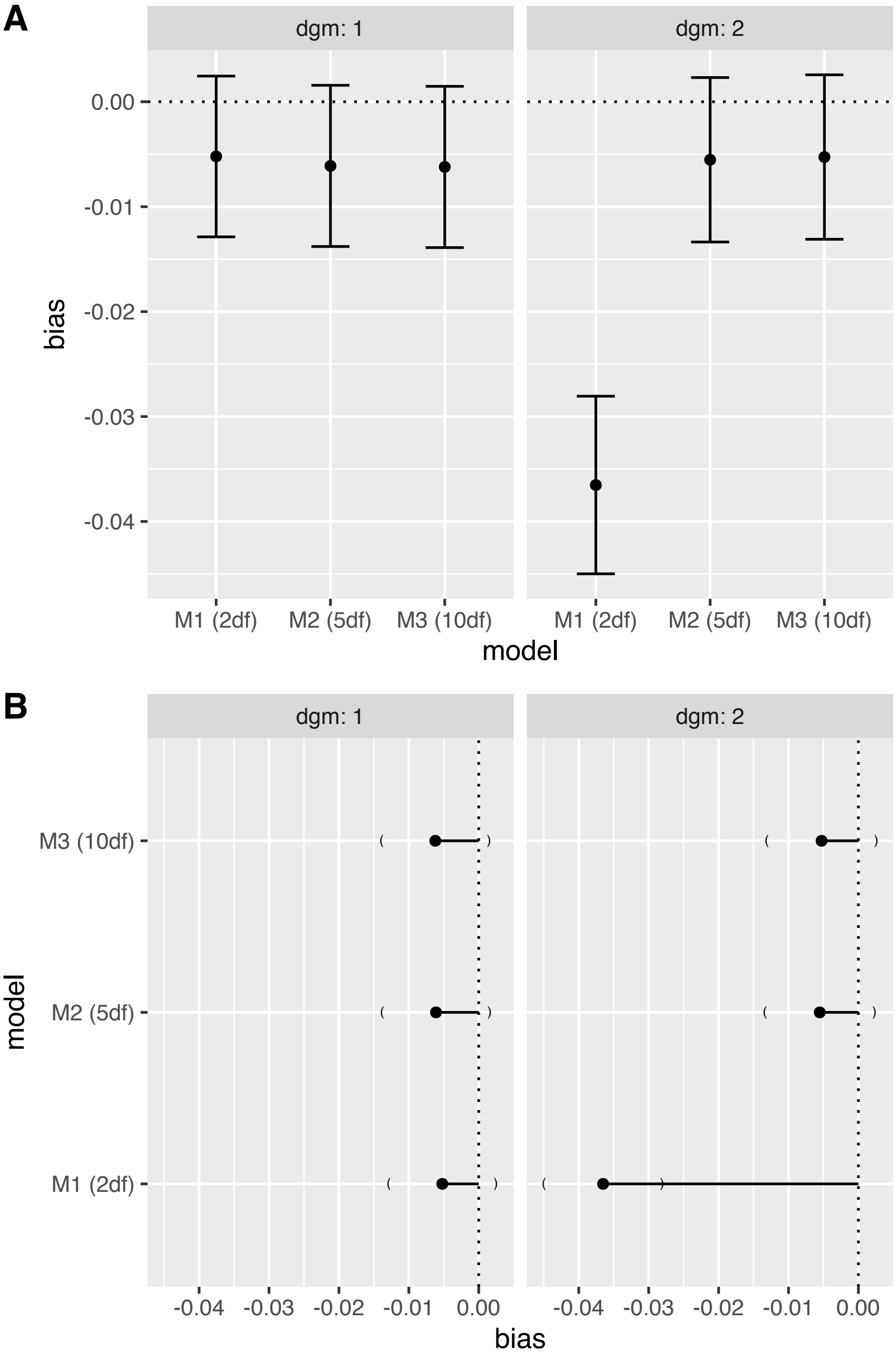


FIGURE 4.16: Method-wise comparison of bias across all data-generating mechanisms, case study using flexible parametric survival models. Panel A is a forest plot and Panel B is a so-called lolly plot. Both plots can be produced by `rsimsum` using the `autoplot` method.

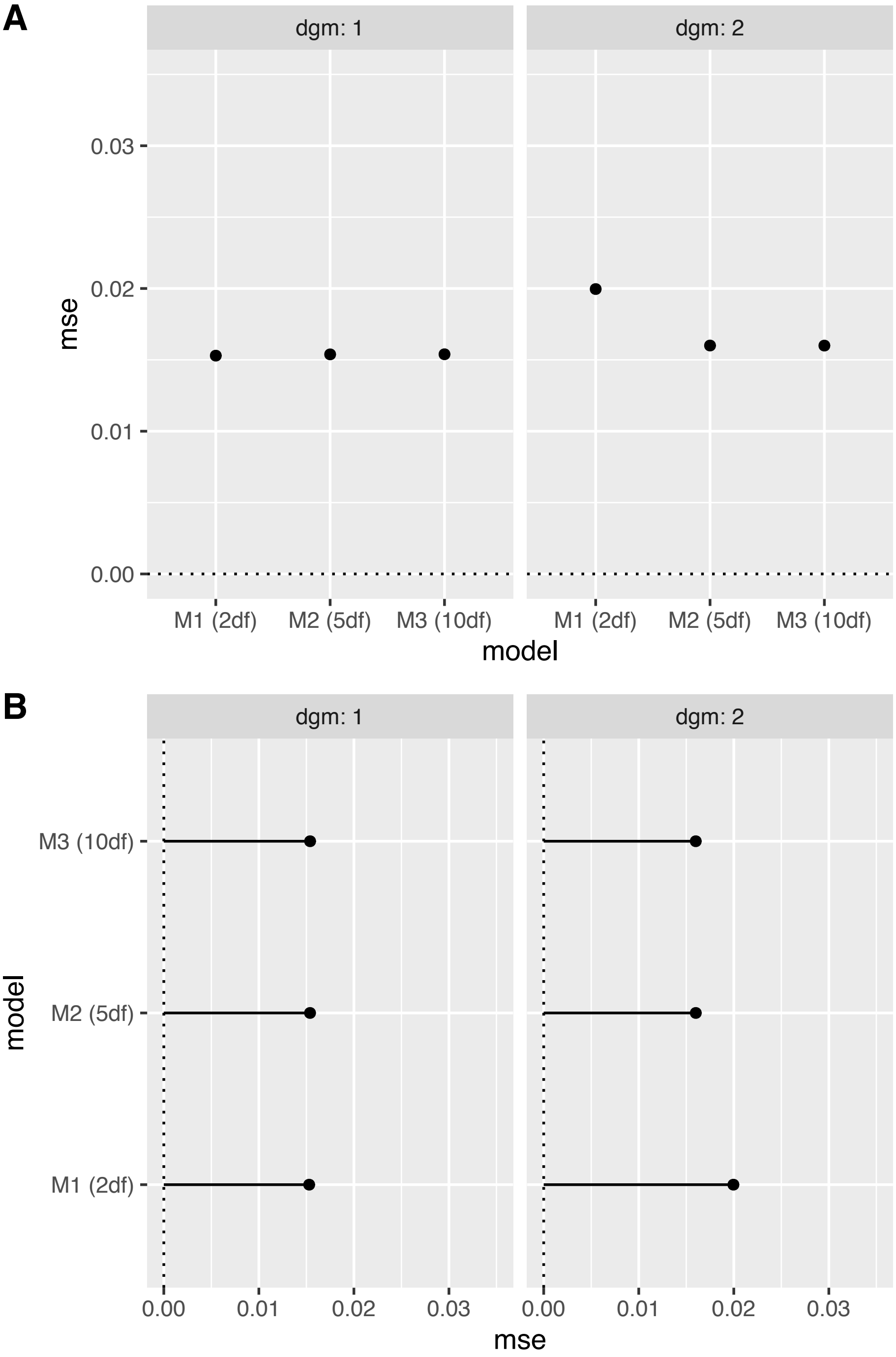


FIGURE 4.17: Method-wise comparison of mean-squared error (MSE) across all data-generating mechanisms, case study using flexible parametric survival models. Panel A is a forest plot and Panel B is a so-called lolly plot. Both plots can be produced by `rsimsum` using the `autoplot` method.

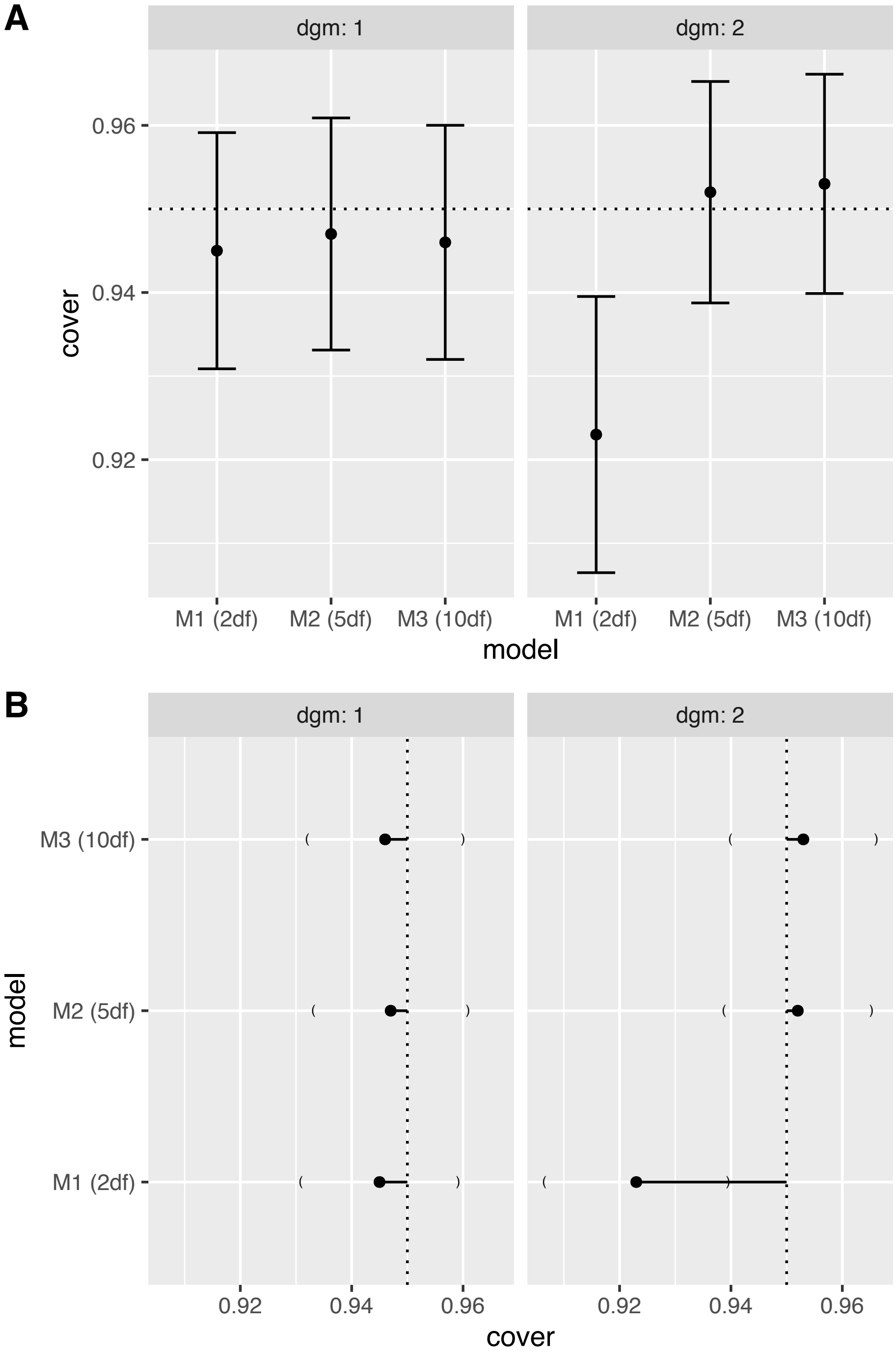


Figure 4.18: Method-wise comparison of coverage probability across all data-generating mechanisms, case study using flexible parametric survival models. Panel A is a forest plot and Panel B is a so-called lolly plot. Both plots can be produced by `rsimsum` using the `autoplot` method.

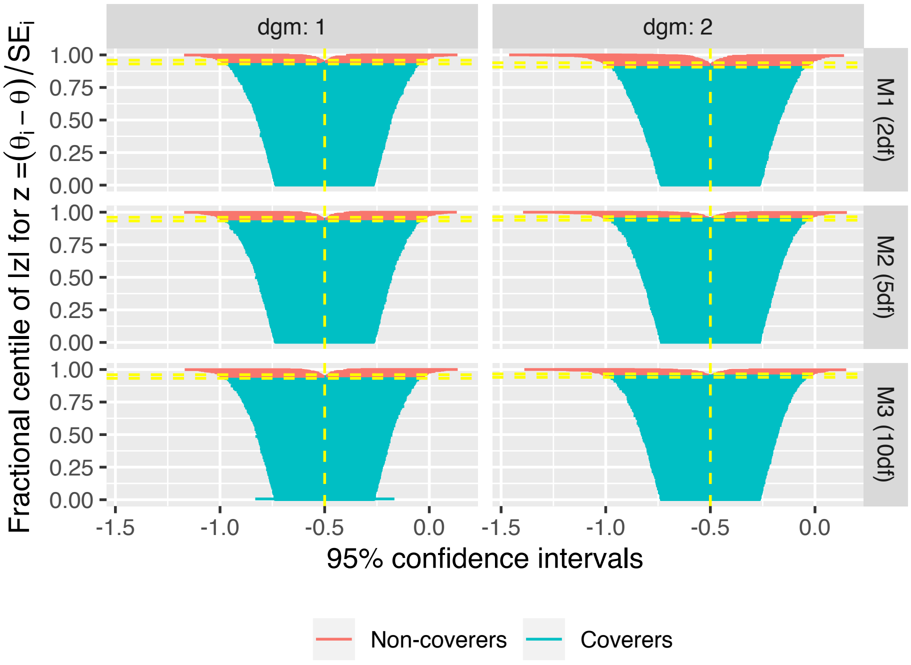


FIGURE 4.19: Zipper plot for the comparison of coverage probability across all data-generating mechanisms, case study using flexible parametric survival models.

Finally, `rsimsum` can also produce zipper plots to present coverage probability (Figure 4.19), and heat maps for any summary statistic (e.g. MSE, as in Figure 4.20).

```
autoplot(simstudy, type = ”zip”)
```

```
autoplot(simstudy, type = ”heat”, stats = ”mse”)
```

Interestingly, all plots obtained via the `autoplot` method can be easily customised by adding other `ggplot2` features. For instance, replacing the default colour palette from the heat map of Figure 4.20 with the viridis colour palette [92] and replacing the default `ggplot2` theme with a minimalistic theme with no background annotations is straightforward:

```
autoplot(simstudy, type = ”heat”, stats = ”mse”) +
  viridis::scale_fill_viridis() +
  ggplot2::theme_minimal()
```

The resulting plot is included as Figure 4.21.

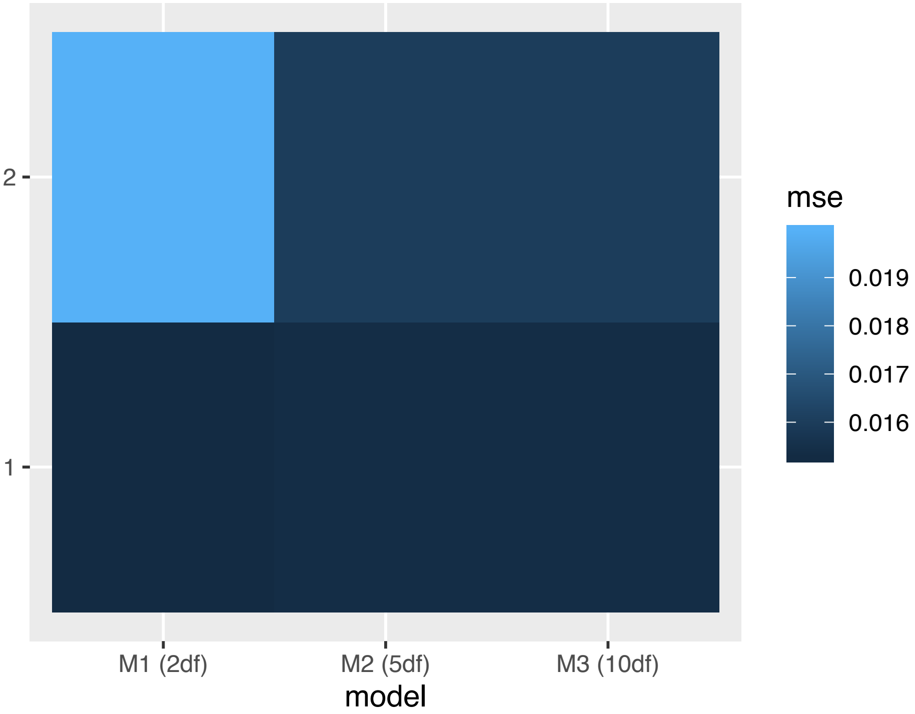


Figure 4.20: Heat map comparing mean squared error across all data-generating mechanisms, case study using flexible parametric survival models.

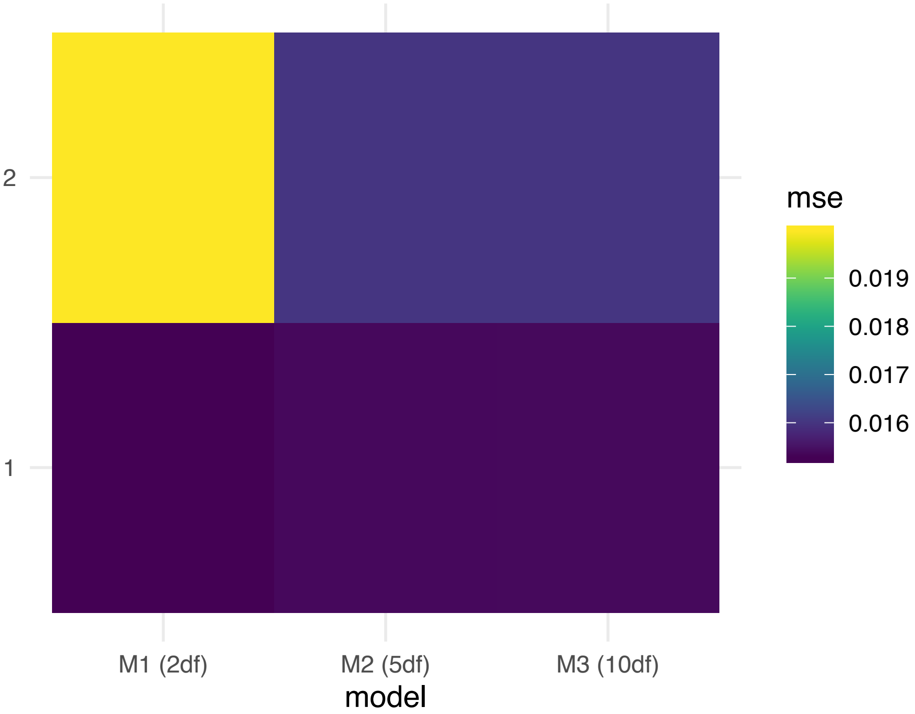


Figure 4.21: Heat map comparing mean squared error across all data-generating mechanisms with customised style, case study using flexible parametric survival models.

4.7 DISCUSSION

In this Chapter, I introduced Monte Carlo simulation studies and motivated their use as an invaluable tool for statistical and biostatistical research. I introduced the ADEMP structured approach proposed by Morris, White, and Crowther [58] and described each component, including the most important performance measures; a structured approach to simulation studies is crucial to guarantee clarity when reporting the results of a study and enhance reproducibility.

Then, I described approaches that can be used to simulate survival data from a given hazard function, including both simple, monotonic baseline hazards and complex formulations with turning points. These methods will be used in Chapters 5 and 7 extensively.

Finally, I introduced the `rsimsum` package and the INTEREST Shiny app. `rsimsum` and INTEREST have been developed during my PhD to support the analysis of Monte Carlo simulation studies and enhance the reporting of their results. Both packages provide an easy-to-use and intuitive option that automates most of the calculations required when computing performance measures. Most importantly, `rsimsum` vastly reduces the amount of error-prone coding required to analyse a simulation study. Monte Carlo standard errors are computed and returned by default, allowing researchers to quantify the degree of uncertainty in the estimation of the performance measures of interest. `rsimsum` has been used throughout this Thesis, especially when analysing the results of the Monte Carlo simulation studies of Chapters 5 and 7.

`rsimsum` and INTEREST are open-source software, and their source code is available from GitHub. `rsimsum` is also available from the Comprehensive R Archive Network (CRAN), while INTEREST can be accessed on-line at `https://interest.shinyapps.io/interest/` or downloaded from GitHub for offline use.

5 *Multilevel Survival Data Analysis*

5.1 OUTLINE

In this Chapter, I introduce the topic of survival data with a multilevel hierarchical structure. First, in Section 5.2 I describe examples of multilevel survival data and how a hierarchy may arise. I describe examples of clustering relevant to the settings of EHR data and focus on the case of recurrent events data (a special case of clustering with observations nested within individuals). Second, I describe methods that can be used to analyse multilevel survival data in Section 5.3 and 5.4, where I describe methods for the analysis of recurrent events data and frailty survival models that can account for broader clustering settings, respectively. Then, I investigate the impact of model misspecification in shared frailty survival models in Section 5.5 via Monte Carlo simulation. In particular, I focus on misspecification of the baseline hazard and/or the distribution of the frailty, and I investigate relative and absolute measures of risk and measures of heterogeneity. The material included in this Section has been published in Statistics in Medicine and is included in Appendix D [93]. In Section 5.6 I apply the methods included in the comparison from Section 5.5 to data from PSP-CKD (described in Section 3.2) to illustrate the impact of model misspecification in practice. Finally, I conclude the Chapter in Section 5.7 with a discussion.

5.2 MULTILEVEL SURVIVAL DATA

Multilevel survival data occurs frequently in a variety of research areas, and especially with EHRs. A hierarchy in the analysis dataset can arise as the result of

1. Clustering, arising from groups of individuals that share common features such as

genetic traits or environmental factors;

2. Events that occur multiple times, commonly termed *recurrent events*.

Clustering becomes apparent when considering geographical clusters: for instance, individuals included in EHR data can be divided into groups such as hospitals or primary care units. The case of recurrent events is a special case of clustering where the clustering unit is the individual: in many biomedical studies the event of interest can occur more than once per study subject. Examples of recurrent events are: admissions to the hospital, falls in elderly patients, migraines, cancer recurrences, infections, bleeding events, and so on. An example of recurring events data for ten study subjects is presented in Figure 5.1: individuals can experience several events, denoted with a solid black dot, until the end of the study (or censoring, denoted by an empty dot).

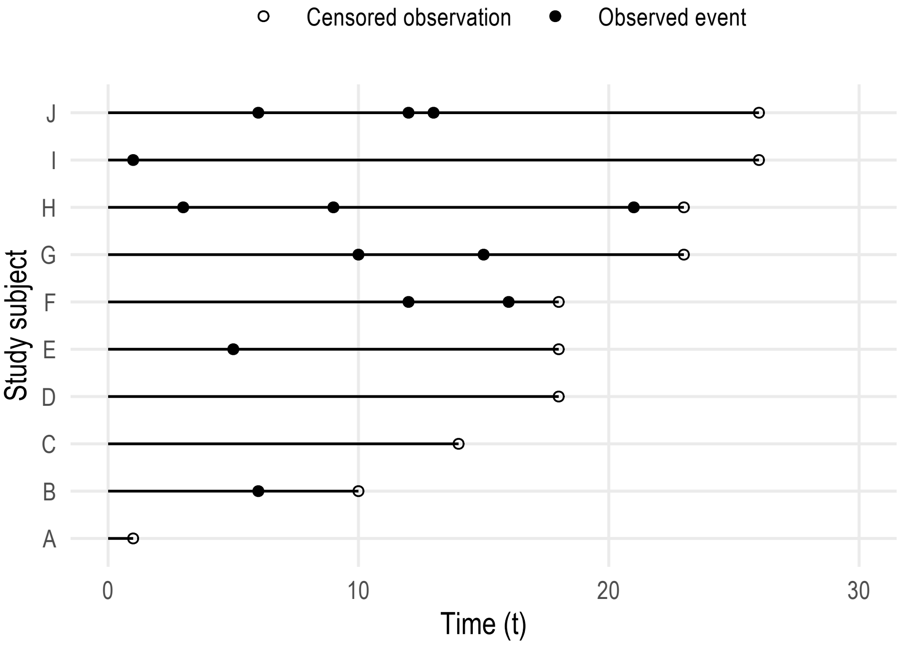


FIGURE 5.1: Example of recurrent events survival data

Conventional regression models for the analysis of survival data (as described in Section 2.2) assume independence between observations. However, the hierarchical structure that I just described yields clusters that include subjects (observations) that are likely correlated - thus violating the above-mentioned assumption of independence. Regression models for multilevel survival data allow analysing survival data that exhibit a given multilevel structure while accounting for clustering, as described in the following Sections.

5.3 MODELLING RECURRENT EVENTS DATA

Within-subject correlation is a key feature of recurrent events data, alongside the ordering of the events and the fact that each subject can be at risk for only one event at a time.

Traditional methods applied to recurrent events data are not wrong per se, but they do not make use of all available information and require strong assumptions. For instance:

1. Logistic regression with *event or not* as outcome ignores time and all events after the first one;
2. Count data models (Poisson or negative binomial regression) model the number of events over time, and the total exposure time can be included in the model as an offset. However, the time between events is ignored;
3. A traditional Cox model ignores all events after the first one by modelling time to the first event only.

Several methods have been proposed in the literature to model recurrent events data. Broadly speaking, they can be categorised in two families: marginal models that account for the correlation by using a robust sandwich-type estimator for the variance-covariance matrix, and methods based on random effects. In particular, the Andersen-Gill (AG) model [94] and the Prentice, Williams and Peterson (PWP) model [95] are considered marginal models. Conversely, the frailty approach is based on the inclusion of a subject-specific random effect that models the within-subject correlation [96]. These methods differ in assumptions, data layout for analysis, and interpretation of the estimated model coefficients e.g. some methods assume that future events depend on the present only (Markov assumption), some models assume dependency via shared random effects, and some models assume that the order of events is important.

5.3.1 The Andersen-Gill Model

The AG model is probably the most often applied model for recurrent events data and generalises the Cox model by modelling the time between recurrent events on the total timescale, i.e. since entry in the study. The total timescale is then split in smaller gap times, where the starting time of a new event is the ending time of the preceding one

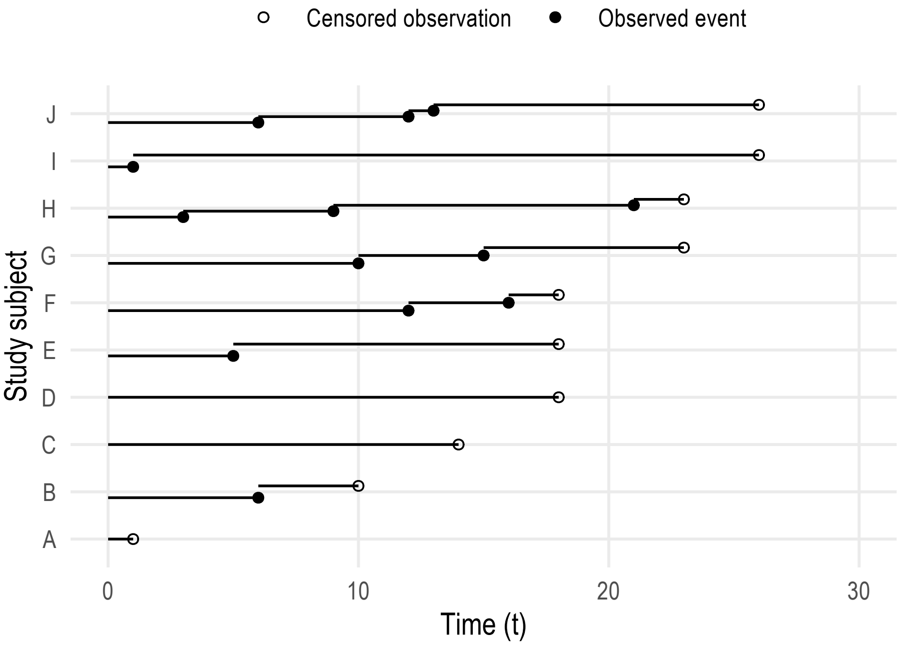


Figure 5.2: Illustration of time splitting in the Andersen-Gill model for recurrent events

(Figure 5.2). The AG model assumes a common baseline hazard function for all events and estimates a global parameter for the covariates included in the model; the underlying assumption is then that the instantaneous risk to experience an event is not affected by whether the previous event occurred or not. The dataset for analysis needs to be in start-stop notation, e.g. each subject has a distinct row for each occurrence of the event of interest (or censoring). Thus, a single patient contributes more than one piece of information depending on the number of individually observed events. Applying the Cox model is then straightforward, using standard statistical software: a Cox model is fitted using a cluster-robust sandwich variance-covariance matrix, where each study individual identifies a cluster. The robust estimator takes into account the correlation between observations originating from the same study subject. The AG model is appropriate when the main aim of the analysis is on the overall effect of covariates on the rate of occurrence of a given event, and when the above-mentioned assumption is believed to hold.

5.3.2 *The Prentice, Williams and Peterson Model*

The PWP model is also a generalisation of the Cox model (although it can be applied to parametric and flexible parametric models as well) and is based on ordering multiple events by stratification based on the prior number of events. This is the main difference

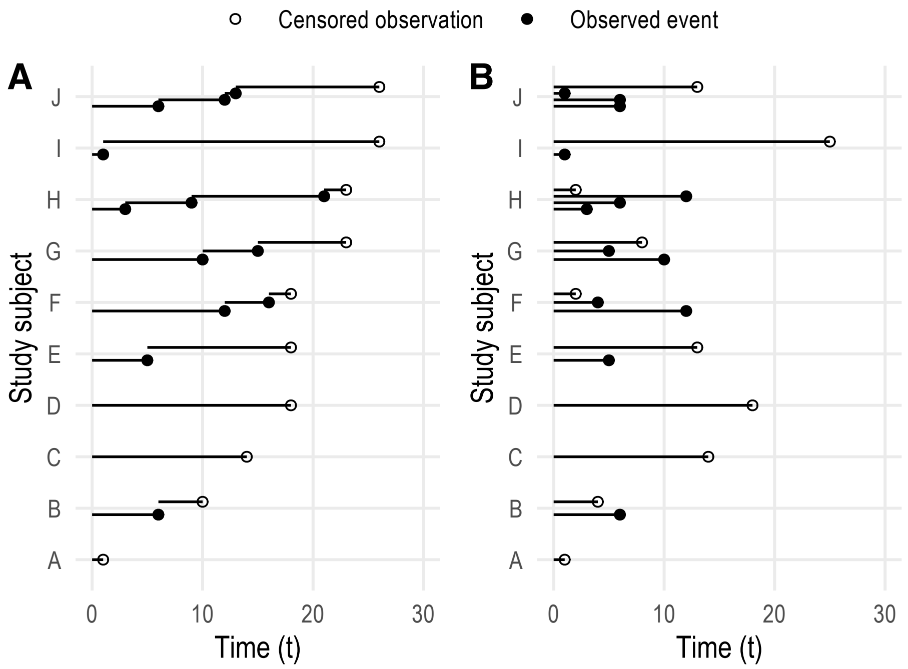


Figure 5.3: Illustration of time splitting in the Prentice, Williams and Peterson model for recurrent events data. Panel A illustrates the total time scale approach (PWP-TT), while panel B illustrates the gap time scale approach (PWP-GT)

with the AG model: with the PWP model, the analysis is stratified by each event. This means that only individuals that experience the first event are at risk for the second one, only individuals that experience the second event are at risk for the third one, and so on. The baseline hazard function is therefore allowed to vary between each subsequent event, and it is possible to include stratum-specific effects in the model as well.

Interestingly, the PWP model can be defined on two time scales: the total time scale (PWP-TT), and the gap time scale (PWP-GT). The time scale of the PWP-TT model is the same in the AG model, that is, the starting time of a new event is the ending time of the one before. Conversely, the time scale of the PWP-GT model differs as the time index is reset at the occurrence of each event. This means that in the PWP-GT model the starting time of a new event is always zero: hence, the gap times approach is also termed *resetting the clock*. A visual comparison of the two approaches is included in Figure 5.3.

Compared to the AG model, the PWP model is preferred when the effects of covariates can vary between subsequent events as well as the baseline rate of occurrence of the recurrent events process. For instance, this may occur when dealing with viral infections, as each individual develops immunity after the first event which greatly modifies the risk of experiencing the event again.

5.3.3 *The Frailty Model*

The final approach that I describe for modelling recurrent events data is the random effects approach. Survival models with an individual-specific random intercept are also known as shared frailty models, where the random effect induces dependence among the recurrent event times for a given individual. The random intercept (also known as frailty) has a multiplicative effect on the hazard and it describes the excess risk (frailty) for a subject compared to the population average. In other words, the frailty term accounts for the unmeasured heterogeneity between individuals in the analysis. Finally, the frailty model could use either the total time scale or the gap time scale (as described for the PWP model in Section 5.3.2, Figure 5.3), with interpretation of the results that varies accordingly. This class of models for multilevel survival data is described in more detail in Section 5.4 and forms the main focus of this Chapter.

5.4 FRAILTY SURVIVAL MODELS

The frailty approach was first proposed by Vaupel *et al.* and Lancaster to model heterogeneity among individuals in univariate data [97, 98]; the resulting frailty survival models are commonly denoted as *univariate frailty survival models*.

However, as described before, it is common to encounter clustered survival data where the overall study population can be divided into heterogeneous clusters of homogeneous observations. With such data, the outcome variable is generally recorded at the lowest hierarchical level while covariates can be measured on units at any level of the hierarchy. As a consequence, survival times of individuals within a cluster are likely to be correlated and need to be analysed as such. Unfortunately, covariates that contribute to explaining the heterogeneity between clusters are often not measured, e.g. for practical reasons. The frailty approach aims to account for the unobserved heterogeneity by including a random effect that acts multiplicatively on the baseline hazard and is shared within a cluster. The univariate frailty approach was extended by Hougaard to accommodate clustered survival data, as in the settings of EHRs [99, 100]; in fact, when considering repeated event-times or clustered data the shared frailty approach yields survival times that are conditionally independent given the frailty [101]. In other words, the presence of this random effect explains the dependence in the sense that had the frailty been known, the survival times would have been independent. The resulting models are commonly referred to as *shared frailty survival models* and are described in Section 5.4.1.

Several extensions of frailty survival models have been developed. For instance Rondeau *et al.* included two nested frailty terms that allow multiple levels of clustering [102], and developed additive frailty models that allow studying both heterogeneity across trials and treatment-by-trial heterogeneity [103]. They also developed joint frailty models for recurrent events and a dependent terminal event to jointly study the evolution of the two processes or account for violations of the proportional hazards assumption [104–106]. Finally, most of these methods assume independence of the frailty terms; Ha *et al.* further relaxed that assumption by developing frailty models that can incorporate correlated frailty effects and/or individual-specific frailty terms within the h-likelihood framework [107].

5.4.1 *Shared Frailty Survival Models*

Shared frailty survival models are defined by a frailty term that introduces a multiplicative effect α_i on the hazard:

$$h_{ij}(t|\alpha_i) = \alpha_i h_0(t) \tag{5.1}$$

for the j^{th} observation in the i^{th} cluster; specifically, individuals within the same i^{th} cluster share the frailty effect α_i.

The frailty term is chosen to have a distribution $f(\alpha)$ with expectation $E(\alpha) = 1$ and variance $\text{Var}(\alpha) = \sigma^2$. $\text{Var}(\alpha)$ is interpretable as a measure of heterogeneity across the population in baseline risk: as σ^2 increases the values of α_i are more dispersed, with greater heterogeneity in $\alpha_i h_0(t)$. Underlying assumptions of this model are: the frailty is time-independent, and it acts multiplicatively on the underlying baseline hazard function.

Introducing observed covariates into the model from Equation (5.1) and inducing proportional hazards:

$$h_{ij}(t|\alpha_i) = \alpha_i h_0(t) \exp(x_{ij}\beta) = \alpha_i h(t|x_{ij}), \tag{5.2}$$

with x_{ij} and β covariates and regression coefficients, respectively.

Any distribution or functional form can be assumed for $h_0(t)$ [108], or it is possible to leave it unspecified altogether yielding a semi-parametric Cox model with random effects [109, 110]. Advantages and disadvantages of modelling the baseline hazard in frailty survival models are the same described in Section 2.2.4 in the settings of standard regression models for survival data.

Given the relationship between hazard and survival function, it can be shown that the individual survival function conditional on the frailty is:

$$S_{ij}(t|\alpha_i) = S_{ij}(t)^{\alpha_i} \tag{5.3}$$

The cluster-specific contribution to the likelihood (assuming no left truncation for simplicity) is obtained by calculating the cluster-specific likelihood conditional on the frailty, consequently integrating out the frailty itself:

$$L_i = \int_A L_i(\alpha_i) f(\alpha_i) \, d\alpha, \tag{5.4}$$

with $f(\alpha)$ the distribution of the frailty, A its domain, and $L_i(\alpha_i)$ the cluster-specific contribution to the likelihood, conditional on the frailty. The cluster-specific contribution to the likelihood is

$$L_i(\alpha_i) = \alpha_i^{D_i} \prod_{j=1}^{n_i} S_{ij}(t_{ij})^{\alpha_i} h_{ij}(t_{ij})^{d_{ij}}, \tag{5.5}$$

with $D_i = \sum_{j=1}^{n_i} d_{ij}$.

Different choices for the frailty distribution are possible, as described in more details by Hougaard [111, 112]. Assigning a probability distribution implies that the frailty can be integrated out of the likelihood function. After this integration, the likelihood can be maximized in the usual way if an explicit form exists. Otherwise, more sophisticated approaches such as numerical integration are required.

The Gamma distribution is widely used, being mathematically very convenient; the inverse Gaussian distribution is also common. A main difference between the two is that a Gamma frailty yields a time-independent heterogeneity, while an inverse Gaussian frailty yields heterogeneity that decays over time, making the population more homogeneous as time goes by; in general, the relative shapes of the individual and population hazard functions could differ greatly because of the frailty effect. Additionally, Hougaard presents a family of distributions with infinite mean, such as the reciprocal Gamma distribution and the positive stable distribution. It is possible to use a log-normal frailty as well; however, that leads to analytically intractable formulæ and additional computational complexity.

Assuming that the frailty α has a Gamma distribution is practical: it has the appropriate range $(0, \infty)$ and it is mathematically tractable. A Gamma distribution with parameters a and b has density

$$f(x) = \frac{x^{a-1} \exp(-x/b)}{\Gamma(a) b^a}; \tag{5.6}$$

by choosing $a = 1/\theta$ and $b = \theta$ the resulting distribution has expectation 1 and finite variance θ. In these settings, the model is analytically tractable: the population survival function takes the form

$$S(t) = [1 - \theta \log(S(t))]^{-1/\theta} \tag{5.7}$$

with the likelihood following by substitution:

$$L_i = \left[\prod_{j=1}^{n_i} h_{ij}(t_{ij})^{d_{ij}} \right] \frac{\Gamma(1/\theta + D_i)}{\Gamma(1/\theta)} \left[1 - \theta \sum_{j=1}^{n_i} \log S_{ij}(t_{ij}) \right]^{-1/\theta - D_i} \tag{5.8}$$

Estimating such model becomes therefore straightforward, which likely contributed to the popularity of Gamma frailty models.

Together with the Gamma distribution, the log-normal distribution is one of the most commonly used frailty distribution, given its strong ties to random effect models. Assuming a log-normal distribution with a single parameter $\theta > 0$ (for comparison with the mathematically tractable Gamma frailty model) with density

$$f(x) = (2\pi\theta)^{-\frac{1}{2}} x^{-1} \exp\left(-\frac{(\log x)^2}{2\theta} \right), \tag{5.9}$$

the resulting model has a frailty whose expectation is finite. Despite that, this frailty distribution cannot be integrated out of the survival function analytically to obtain the population survival function or the likelihood, requiring additional computational complexity (for instance, numerical quadrature or stochastic integration is required). Computational issues in frailty models are discussed in more detail in Section 5.4.2.

As above-mentioned, a shared frailty model assuming a log-normal distribution for the frailty term has strong ties to random-effects models. A log-normal frailty model is formulated as

$$h_{ij}(t|\alpha_i) = \alpha_i h(t|x_{ij}) = \alpha_i h_0(t) \exp(x_{ij}\beta), \tag{5.10}$$

with α_i following a log-normal distribution. On the log-hazard scale:

$$h_{ij}(t|\alpha_i) = h_0(t) \exp(x_{ij}\beta + \eta_i), \tag{5.11}$$

with $\eta_i = \log \alpha_i$. η_i is therefore normally distributed with parameters μ and σ^2 related to those of the log-normal distribution by the relationship

$$E(\alpha_i) = \exp(\mu + \sigma^2/2) \tag{5.12}$$

and

$$Var(\alpha_i) = \exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1) \tag{5.13}$$

By formulating the model on the log-hazard scale, the frailty term has a direct interpretation as a random intercept in the model. It is possible to further extend this model by allowing multiple random covariates effects, potentially ranging over multiple levels of clustering, as described by Crowther *et al.* [108]. Borrowing the usual mixed-effects model notation from Section 2.3, and assuming a single level of clustering:

$$h_{ij}(t|b_i) = h_0(t) \exp(x_{ij}\beta + z_i b_i), \tag{5.14}$$

with x_{ij} the fixed effects and z_i the cluster-specific random effects, with associated coefficients β and b_i, respectively. Under this more general formulation, a survival model can include not only a random intercept but also random effects of other covariates included in the model, potentially over multiple levels of clustering (e.g. patients nested into hospitals nested into Countries).

5.4.2 *Computational Challenges*

I mentioned in Section 5.4.1 that frailty survival models with a log-Normal frailty (and consequently survival models with normally-distributed random effects) require additional computational complexity during the estimation procedure.

For instance, the distribution of the frailty has to be integrated out of the likelihood function (Equation (5.4)) to obtain the marginal likelihood to be maximised, or to obtain the marginal hazard and survival function.

In this Section I will describe Gaussian quadrature, a numerical method that can be used to approximate intractable integrals (such as those mentioned above). Simple Gauss-Hermite quadrature [113, 114] can be used to evaluate analytically intractable integrals of the form

$$\int_{-\infty}^{+\infty} e^{-x^2} f(x)\ dx \approx \sum_{q=1}^{m} w_1 f(x_q) \tag{5.15}$$

with x_q and w_q quadrature nodes and weights, respectively. The quadrature nodes x_q are the q^{th} root of the Hermite polynomial $H_m(x)$, while the quadrature weights w_q can be calculated as

$$w_q = \frac{2^{m-1} m! \sqrt{\pi}}{m^2 [H_{m-1}(x_q)]^2} \tag{5.16}$$

This approximation is exact for polynomials of degree $2m - 1$.

Following Naylor and Smith [113] and Tuerlinckx *et al.* [115], it is possible to replace the weighting function e^{-x^2} in Equation (5.15) with a normal density $\phi(\cdot)$ with mean μ and standard deviation σ:

$$\int_{-\infty}^{+\infty} f(x)\phi(x|\mu, \sigma^2)\ dx = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} f(x) \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]\ dx \tag{5.17}$$

Then, undertaking a change of variable and setting $x = \mu + \sigma\sqrt{2}r$ with $r = (x-\mu)/(\sqrt{2}\sigma)$, Equation (5.17) becomes:

$$\begin{aligned} \int_{-\infty}^{+\infty} f(x)\phi(x|\mu, \sigma^2)\ dx &= \frac{\sqrt{2}\sigma}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} f(\mu + \sigma\sqrt{2}r) e^{-r^2}\ dr \\ &\approx \sum_{q=1}^{m} f(\mu + \sigma\sqrt{2}r) \frac{w_q}{\sqrt{\pi}} \end{aligned} \tag{5.18}$$

This is a Gauss-Hermite quadrature evaluation based on the normal kernel, with nodes $d_q = \mu + \sigma\sqrt{2}x_q$ and weights $v_q = w_q/\sqrt{\pi}$.

The Gauss-Hermite integration introduced so far only applies for a univariate function, e.g. for a single random effect or a frailty term. With a multivariate function (such as in

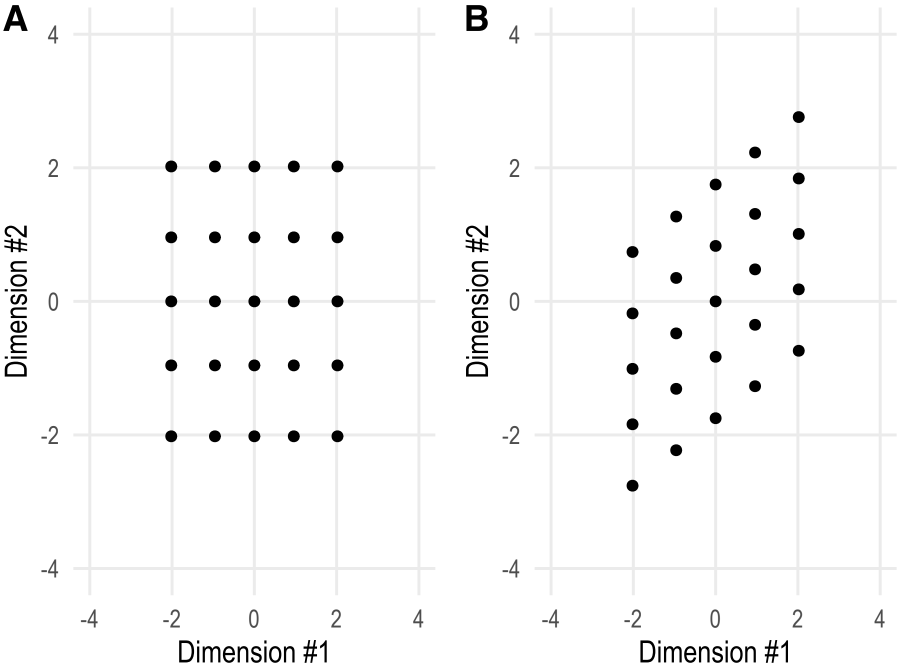


FIGURE 5.4: Gauss-Hermite quadrature nodes for a bi-dimensional Normal kernel. Panel A depicts quadrature nodes when the two dimensions are independent, while panel B depicts quadrature nodes for correlated dimensions.

the settings of multiple random effects, e.g. a random intercept and slope) the integration procedure can be easily extended to the Q-dimensional case [116]; this will be most relevant in Chapters 6 and 7. For instance, in the multivariate case I have a vector of quadrature nodes $\mathbf{d}_{q_1,\dots,q_Q} = (d_{q_1}, \dots, d_{q_Q})$. The difference with the univariate case is that the vector of nodes has to be pre-multiplied by $\Omega^{1/2}$, the Cholesky decomposition (or the spectral decomposition) of the variance-covariance matrix of the multivariate Normal distribution used as kernel to account for correlation.

The grid of nodes for multivariate Gauss-Hermite quadrature is illustrated in Figure 5.4, assuming $Q = 2$ dimensions for simplicity and $m = 9$ nodes. Panel A depicts the grid of quadrature nodes without accounting for the correlation with the two dimensions; conversely, panel B depicts the grid of quadrature nodes assuming the following variance-covariance matrix:

$$\begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

By accounting for the correlation between the two dimensions the grid of quadrature nodes is therefore rotated to better approximate the bivariate distribution.

The accuracy of the approximation of the integral in Equation (5.18) depends on the number of quadrature points m; a standard practice when utilising Gauss-Hermite quadrature to approximate a likelihood function consists of fitting models with an increasing number of quadrature nodes, until the difference between estimates is small enough. However, the computational burden grows considerably as m increases, a burden that further grows exponentially with an increasing number of dimensions Q.

Using the same grid of nodes for each cluster e.g. in a shared frailty model with a random effect is, however, inefficient: when the heterogeneity between clusters is large, the nodes will likely not be placed in the optimal position hence failing to capture key information from the likelihood function. To improve the performance of standard Gauss-Hermite quadrature in the settings that I just described, Pinheiro and Bates proposed an adaptive quadrature method that allows cluster-specific centring and scaling of the quadrature nodes [117]. This procedure, named *adaptive* Gauss-Hermite quadrature, achieves the goal of placing the nodes at the most appropriate position of the integral function for each cluster reducing the computational burden as fewer nodes are required to obtain the same accuracy level of standard Gauss-Hermite quadrature. Adaptive Gauss-Hermite quadrature uses an alternative Normal kernel density with nodes appropriately transformed using the formula $\mathbf{r}_i = \hat{\mathbf{b}}_i + \hat{\mathbf{\Omega}}_i^{1/2} \mathbf{d}_{b_1,\ldots,b_Q}$: rather than centring the quadrature nodes on zero and scaling using the overall Choleski decomposition of the variance-covariance matrix, cluster-specific means and variance-covariance matrices are used for centring and scaling. An example of adaptive Gauss-Hermite quadrature is illustrated in Figure 5.5.

5.5 IMPACT OF MODEL MISSPECIFICATION IN SHARED FRAILTY SURVIVAL MODELS

Survival models as described in Section 2.2.4 require a modelling choice regarding the shape of the baseline hazard function $h_0(t)$.

The standard, most common approach consists of leaving the baseline hazard unspecified (e.g. in a Cox model), especially when relative effect estimates are of interest. Nonetheless, the aim of the analysis often includes obtaining and reporting absolute measures of risk: in that context, modelling the baseline hazard has favourable properties and it can be

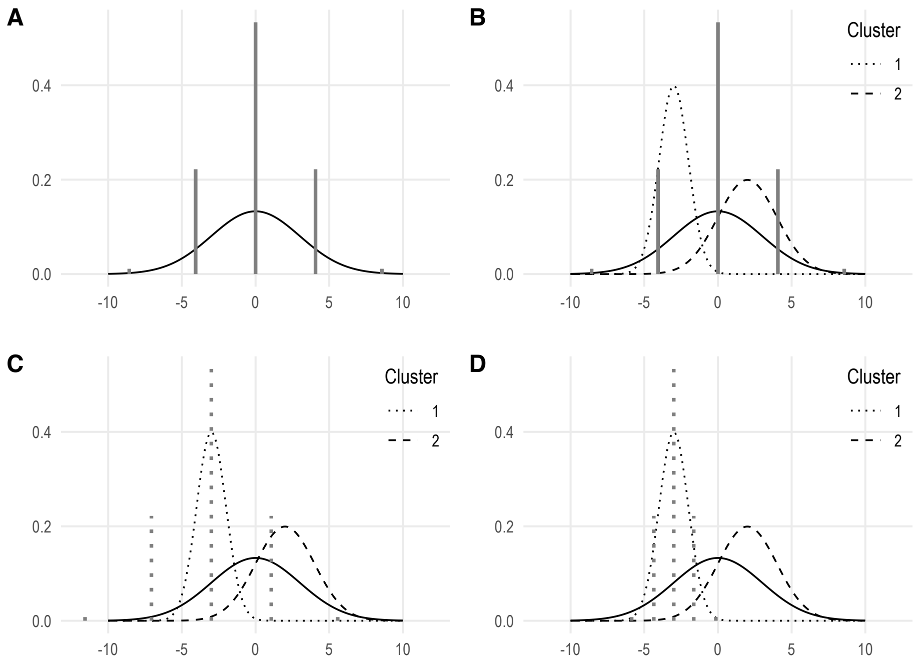


Figure 5.5: Example of Adaptive Gauss-Hermite quadrature. In panel A I depict standard Gauss-Hermite quadrature, with the overall distribution; however, as depicted in panel B, each cluster deviates from the overall distribution. Therefore, adaptive Gauss-Hermite quadrature centres the quadrature nodes on each cluster's estimated mean (panel C) and re-scales the nodes using the estimated variance (panel D). The centring and scaling is depicted only for cluster 1 to reduce clutter

achieved by using standard parametric distributions for the baseline hazard (as described in Section 2.2.4.2). Alternatively, one could use the flexible parametric modelling approach of Royston and Parmar [24] to better capture the shape of complex hazard functions. The latter approach requires choosing the number of degrees of freedom for the spline term used to approximate the baseline hazard: in practice, sensitivity analyses and information criteria (AIC, BIC) have been used to select the best model. Rutherford et al. showed via simulation that, assuming a sufficient number of degrees of freedom is used, the approximated hazard function given by restricted cubic splines fits well for several complex hazard shapes and the hazard ratios estimation is insensitive to the correct specification of the baseline hazard [28]. Using a sufficient number of degrees of freedom, the spline-based approach is able to capture the underlying shape of the hazard function with minimal bias; AIC and BIC can guide the choice of the best fitting model, but they tend to agree to within 1 or 2 degrees of freedom in practice.

The modelling choice regarding the shape of the baseline hazard is required in the settings of shared frailty survival models as well. On top of that, shared frailty survival models require choosing an appropriate distribution for the frailty term: as described above, assuming a Gamma distribution has favourable mathematical properties and, despite that, the log-Normal distribution is commonly assumed as well given its strong ties with random-effects models.

The choice of a particular parametric frailty distribution is known to have a marginal impact on the estimation and testing of regression coefficients. Pickles and Crouchley [118] showed how the estimated values and the distribution of the likelihood ratio test statistic do not differ much comparing a variety of models such as the Weibull survival model with a Gamma or log-Normal frailty. They conclude by arguing that the convenience and generality of the baseline hazard would seem more important than generality of the frailty distribution when fitting a frailty model. Glidden and Vittinghoff reached the same conclusions, highlighting how different frailty distributions can lead to appreciably different association structures despite not greatly affecting the estimation of regression coefficients [119]. Lee and Thompson [120] showed how violations of the normality assumption for random effects in hierarchical models do not affect fixed effects substantially while having a substantial impact on inference regarding the random effects. Thus, they advocate the use of more flexible distributions such as the t or the skewed t for the random effects when the distribution of the random effects is of interest - e.g. in the context of meta-analysis - despite the increased complexity. Liu *et al.* [121] showed that flexible parametric survival models with frailties perform well, both in terms of estimating the regression coefficients and the variance of the frailty; in comparison, semiparametric frailty models with a log-Normal frailty underestimated the variance of the frailty. They also showed that model misspecification could lead to an inflated estimated variance of the frailty and a biased estimated survival function. Duchateau *et al.* [122] showed via simulation that the number of centres and the number of patients per centre influence the quality of the estimates, and they argue (in the context of multi-centre clinical trials) the importance of making sure that a trial is sufficiently large for the estimated heterogeneity parameter to actually describe the true heterogeneity between centres and not just random variability. Finally, Ha *et al.* showed (in the h-likelihood framework) that misspecifying the baseline hazards results in larger bias than assuming the wrong frailty distribution [123]. In conclusion, the small impact of misspecifying the frailty distribution on regression coefficients seems to be well established in the literature, with some evidence pointing towards biased absolute measures of risk. Nevertheless, the structure of the frailty can be as important as the choice of the baseline hazard given that it gets easier to distinguish between unobserved heterogeneity and non-proportional hazards when more information on the correlation structure is available [124, 125].

However, little is known about the impact of misspecifying the baseline hazard in survival models with frailty terms, and the impact of model misspecification on model predictions. Throughout this Section, I will investigate the impact of misspecifying the baseline hazard, the distribution of the frailty, or both on measures of relative (regression coefficients) and absolute (loss in life expectancy [126]) risk, and on heterogeneity measures such as the estimated variance of the frailty component. Absolute measures of risk are particularly important when communicating the results of a study, as it provides additional information that is especially useful to patients and policy-makers. It has been argued that both relative and absolute measures of risk should be reported, as together they provide a complete picture of the effect and its implications [127]; therefore, assessing the impact of model misspecification on such measures is crucial as well. Absolute risk predictions are unfortunately often not implemented in statistical software packages, hence this simulation study will require ad-hoc coding to compute the loss in life expectancy (and its standard error). I will compare a large set of models under different data-generating mechanisms: semiparametric and fully parametric survival models with frailties, models with flexible baseline hazard, and models with flexible baseline hazard and a penalty for the complexity of the spline. The settings of this Monte Carlo simulation study are described in Sections 5.5.1 to 5.5.6, using the ADEMP structure introduced in Section 4.3. Finally, the results are described in Section 5.5.7.

5.5.1 Aims

The impact of misspecifying the baseline hazard, the frailty distribution, or both in survival models with shared frailty is not fully understood. Therefore, the primary aim of this simulation study consists in assessing the consequences of such misspecification on estimates of risk, both relative and absolute. This is particularly relevant as parametric survival models are being increasingly used in applied settings, with flexible frameworks and software readily available [108].

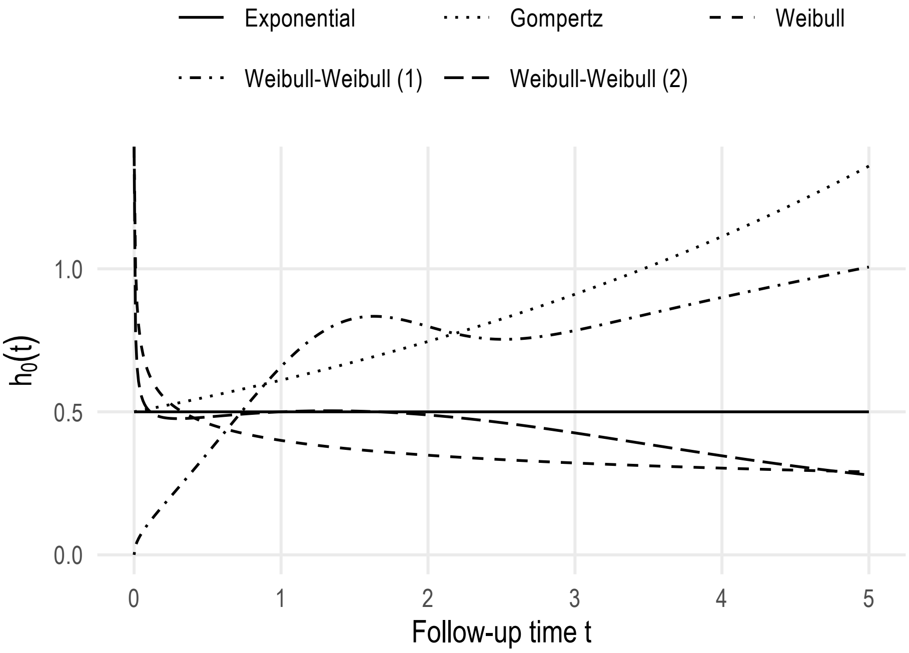


FIGURE 5.6: Simulated baseline hazard functions

I will simulate clinically plausible clustered survival data, aiming to mimic real data scenarios with each data-generating mechanism: clustered studies such as multi-centre clinical trials, individual patient data meta-analysis, paired organ studies, twin studies, and so on.

5.5.2 *Data-Generating Mechanisms*

I simulate data from the following data-generating model:

$$h_{ij}(t) = \theta_i h_0(t) \exp(x_{ij}\beta),$$

where x_{ij} is a binary treatment variable simulated from a Bernoulli random variable with probability $p = 0.50$ and an associated log-hazard ratio $\log(\beta) = -0.50$ and cluster-specific frailty terms following either a Gamma or a log-Normal distribution with variance θ, $\theta \in \{0.25, 0.75, 1.25\}$. I simulate survival times under five different baseline hazard functions using the methods described in Section 4.4; specifically, I chose the exponential, Weibull, Gompertz hazard functions, and two different two-components Weibull-Weibull mixture distribution (Figure 5.6, Table 5.1). Administrative censoring is applied at 5 years.

Then, for each baseline hazard function, I simulated clustered data for 750 clusters of 2 individuals each and 20 clusters of 150 individuals each. I also simulated a mixture

TABLE 5.1: Parameters of data-generating baseline hazard functions

Baseline hazard function	Parameters
Exponential	$\lambda = 0.5$
Weibull	$\lambda = 0.5, \gamma = 0.8$
Gompertz	$\lambda = 0.5, \gamma = 0.2$
Weibull-Weibull (1)	$\lambda_1 = 0.3, \lambda_2 = 0.5, \gamma_1 = 1.5, \gamma_2 = 2.5, \pi = 0.7$
Weibull-Weibull (2)	$\lambda_1 = 0.5, \lambda_2 = 0.5, \gamma_1 = 1.3, \gamma_2 = 0.7, \pi = 0.5$

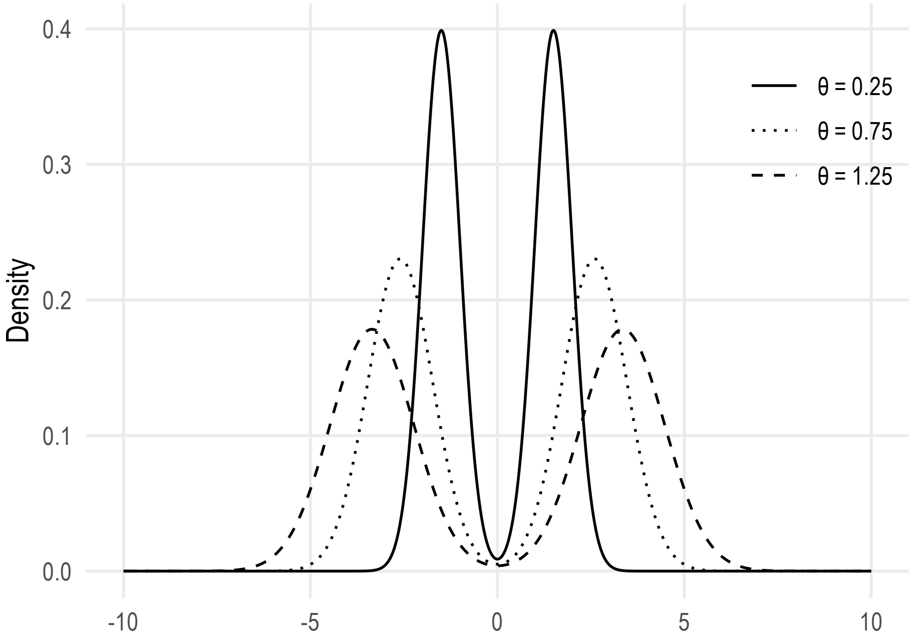


FIGURE 5.7: Simulated settings for the scenarios with a frailty term following a mixture Normal distribution with $\theta \in \{0.25, 0.75, 1.25\}$ frailty

Normal frailty distribution: as a motivation for this distribution, assume the presence of $G = 2$ hidden groups in each cluster (e.g. an unmeasured binary covariate). Formally, let $g = \{1, 2\}$ be an index over the groups, with π_g being the proportion in the groups and $\sum_g \pi_g = 1$. Let the hazard for the i$^{\text{th}}$ individual in the j$^{\text{th}}$ cluster (and g$^{\text{th}}$ hidden group) be:

$$h_{gij}(t) = h_0(t) \exp(x_{gij}\beta + \sum_g \pi_g H_g),$$

where $\sum_g \pi_g H_g$ follows a mixture Normal distribution with mixing probabilities π_g and $H_g \sim N(\mu_g, \sigma_g^2)$. I assume $\pi_1 = \pi_2 = 0.5$, $\mu = \{-3\sqrt{\theta}, +3\sqrt{\theta}\}$, and $\sigma_1^2 = \sigma_2^2 = \theta$ for the purposes of these simulations: this yields very distinct hidden groups, with group-specific means that are 6-standard deviations apart (as illustrated in Figure 5.7).

In conclusion, I simulated clustered survival data for 2 different sample sizes (number of individuals and clusters), 3 possible distribution of the frailty component, 3 frailty variances, and 5 baseline hazard functions. Given that I am using a fully factorial design, this adds up to 90 distinct data-generating mechanisms.

5.5.3 *Estimands*

The estimands of interest are estimates of relative risk, absolute risk, and heterogeneity. Besides, I will monitor and report on convergence rates of each model as well.

Relative Risk

The relative risk estimate of interest is the regression coefficient β associated with the binary treatment; this coefficient can be interpreted as the log-treatment effect, conditional on the unobserved value of the frailty term. Notably, the hazard ratio in a frailty model carries the usual interpretation only when comparing two hazards conditional on a given frailty; unconditionally, at a population level, the proportionality of hazards is not guaranteed to hold even under the proportional hazards parametrisation. For most frailty distributions (including the Gamma and log-Normal) the conditional hazard ratio is a true hazard ratio only at time $t = 0$, as the effect of the covariates on the hazard varies over time depending on the actual distribution of the frailty [100, 112, 125].

Absolute Risk

The absolute risk estimate of interest is the 5-years loss in life expectancy (LLE) (associated with the treatment of interest), defined as the difference in life expectancy between exposed and non-exposed individuals. The marginal 5-years life expectancy (LE) for exposed individuals ($X = 1$) is defined as

$$\begin{aligned} \text{LE}(X = 1) &= \int_0^5 S(u|X = 1) \ du \\ &= \int_0^5 \int_A S(u|X = 1, \alpha) \ p(\alpha) \ d\alpha \ du, \end{aligned} \tag{5.19}$$

where A is the domain of the frailty α and $p(\alpha)$ its density function. The marginal 5-years LE as defined in Equation (5.19) is also known as *restricted mean survival time* [128].

Consequently, the LLE for exposed versus non-exposed individuals is defined as

$$\text{LLE} = \int_0^5 S(u|X = 1)\, du - \int_0^5 S(u|X = 0)\, du. \tag{5.20}$$

Analogously as before, LLE as defined in Equation (5.20) is also known as *difference in restricted mean survival times* [128].

The inner integral in Equation (5.19) has a closed-form when the frailty follows a Gamma distribution; with a log-Normal frailty (and with a mixture Normal frailty), numerical integration is required.

I use the `quadinf` function from the `pracma` package in R to perform numerical integration, which implements the double exponential method for fast numerical integration of smooth real functions on finite intervals [129, 130]. For infinite intervals, the tanh-sinh quadrature scheme is applied [131]. The outer integral in Equation (5.19), however, is approximated using spline-based integration as follows. First, I estimate LE over 1,000 values of t between the minimum and the maximum observed survival times; then, I fit an interpolating natural spline function over the 1,000 LE estimates from step (1), which I finally integrate between 0 and 5 (years) using the double exponential method of `quadinf`. LLE follows by computing the difference between the two integrals. Finally, the standard error of the estimated LLE is computed using the numerical delta method (as implemented in the `predictnl` function from the `rstpm2` R package [132]).

Heterogeneity

With this simulation study I mainly focus on estimates of risk; despite that, measures of heterogeneity are sometimes used to quantify dependence between clustered observations. Therefore, I will report the results of the simulation study regarding the frailty variance as well; as the frailty variance estimated by models assuming either a Gamma or a log-Normal distribution are not directly comparable (being modelled on different scales, hazard versus log-hazard), I will not include summary statistics for the frailty variance where the frailty distribution is misspecified.

5.5.4 *Methods and Software*

Methods included in this comparison are shared frailty models, defined as in Section 5.4.1; as mentioned above, assumptions regarding the shape of the baseline hazard and the distribution of the frailty are required. Therefore, I will compare a variety of shared frailty models, described in the following paragraphs and summarised in Table 5.2.

First, I fit semiparametric shared frailty models by leaving the baseline hazard function unspecified and assuming either a Gamma or a log-Normal frailty (Model 1–2, denoted by *Cox* in plots). The Cox shared frailty model with a Gamma frailty can be fit using the `frailtyEM` R package, which uses maximum likelihood estimation via the Expectation-Maximization algorithm [40] and uses exact formulæ; the Cox shared frailty model with a log-Normal frailty can be fit using the `coxme` R package, and relies on penalised likelihood estimation [109] and the Laplace approximation to integrate out the distribution of the frailty. Hirsch and Wienke compared several R packages for fitting semiparametric frailty models, and `coxme` emerged as the most robust package [133]; I chose to use `frailtyEM` over competing packages for semiparametric shared frailty models with a Gamma frailty as it implements predictions for the marginal survival function, and it is being actively maintained. Other packages that support semiparametric shared frailty models (using different estimation algorithms) are for instance `frailtySurv` [134] and `frailtyHL` [135].

Second, I fit fully parametric survival models by assuming that the baseline hazard function follows an exponential, Weibull, or Gompertz distribution; each of the three models is fit assuming both a Gamma and a log-Normal frailty distribution (Model 3–8, denoted by *Exp*, *Wei*, and *Gom*). These models are fit using the `parfm` R package [136]. As a comparison, I also fit the Weibull model with both Gamma and log-Normal frailties using the `frailtypack` R package (Model 17–18, denoted by *FP(W)*). `parfm` uses the Laplace approximation when fitting models with a log-Normal frailty, while `frailtypack` uses Gaussian quadrature.

Third, I fit Royston-Parmar flexible parametric survival models generalised by Liu *et al.* to account for clustered and correlated survival data [24, 121, 137], as implemented in the `rstpm2` R package [132]. Using the generalised survival model formulation, the model is formulated as

$$S(t_{ij}|x_{ij}, \alpha_i) = \{G[\eta(t_{ij}, x_{ij}; \beta)]\}^{\alpha_i} \tag{5.21}$$

with $G(\cdot) = g^{-1}(\cdot)$ an inverse link function and $\eta(\cdot)$ a linear predictor function of time and covariates.

When choosing the log-log link function, the model is a proportional hazards model; modelling the log of time with natural splines, the resulting model is a Royston-Parmar model whose parameters can be fit using fully parametric maximum likelihood. I fit models with 3, 5, or 9 degrees of freedom for the natural spline of time, and I assume both Gamma and log-Normal shared frailties (Model 9-14, denoted by *RP(df)* where *df* is the number of degrees of freedom).

Liu *et al.* also developed a penalised likelihood estimation procedure that does not require choosing the number of degrees of freedom for the spline term [121]. The penalty term accounts for the complexity of the smoother of time to avoid overfitting the data; however, additional computational complexity is required to select the smoothing parameter (or parameters). The estimation procedure for penalised models is described in detail in Liu *et al.* [121]. Hence, I fit the same Royston-Parmar model with shared frailties using penalised likelihood and either a Gamma or log-Normal frailty (Model 15-16, denoted by *RP(P)*). `rstpm2` uses adaptive Gaussian quadrature to approximate the intractable integrals in shared frailty models with a log-Normal frailty.

Finally, I fit shared frailty models where the baseline hazard function is approximated by cubic M-splines on the hazard scale, as implemented in the `frailtypack` R package [138]. Such models are fitted using a penalised likelihood estimation procedure, and it requires either fixing the smoothing parameter κ or maximizing a likelihood cross-validation criterion to find the optimal value [139]. I choose the former approach and fix the value of κ to the arbitrary values 10 and 10,000, as in the simulations of Liu *et al.* [121] (Model 19-22, denoted by *FP(k=kappa)*, with *kappa* the smoothing parameter κ). As mentioned above, `frailtypack` uses Gaussian quadrature to approximate intractable integrals.

All models included in this simulation study are fitted without tweaking any of the convergence parameters utilised for declaring convergence of the estimation algorithm or any precision setting. All packages but `coxme` returned a standard error for the estimated

Table 5.2: Models included in the Monte Carlo stimulation study on model misspecification in shared frailty survival models

Model	Baseline hazard function	Frailty distribution	R package	Version
1	Unspecified	Gamma	`frailtyEM`	0.8.8
2	Unspecified	Log-Normal	`coxme`	2.2-10
3	Exponential	Gamma	`parfm`	2.7-6
4	Exponential	Log-Normal	`parfm`	2.7.6
5	Weibull	Gamma	`parfm`	2.7.6
6	Weibull	Log-Normal	`parfm`	2.7.6
7	Gompertz	Gamma	`parfm`	2.7.6
8	Gompertz	Log-Normal	`parfm`	2.7.6
9	Royston-Parmar, 3 df	Gamma	`rstpm2`	1.4.5
10	Royston-Parmar, 3 df	Log-Normal	`rstpm2`	1.4.5
11	Royston-Parmar, 5 df	Gamma	`rstpm2`	1.4.5
12	Royston-Parmar, 5 df	Log-Normal	`rstpm2`	1.4.5
13	Royston-Parmar, 9 df	Gamma	`rstpm2`	1.4.5
14	Royston-Parmar, 9 df	Log-Normal	`rstpm2`	1.4.5
15	Royston-Parmar, penalised	Gamma	`rstpm2`	1.4.5
16	Royston-Parmar, penalised	Log-Normal	`rstpm2`	1.4.5
17	Weibull	Gamma	`frailtypack`	3.0.2.1
18	Weibull	Log-Normal	`frailtypack`	3.0.2.1
19	M-splines, $\kappa = 10$	Gamma	`frailtypack`	3.0.2.1
20	M-splines, $\kappa = 10$	Log-Normal	`frailtypack`	3.0.2.1
21	M-splines, $\kappa = 10{,}000$	Gamma	`frailtypack`	3.0.2.1
22	M-splines, $\kappa = 10{,}000$	Log-Normal	`frailtypack`	3.0.2.1

variance of the frailty term; therefore, when fitting a semiparametric shared frailty model with a log-Normal frailty, I used non-parametric bootstrap with 1,000 replications (resampling at the cluster level to preserve the within-cluster correlation) to estimate the variance of the frailty [140]. Lastly, only `frailtyEM`, `rstpm2`, and `frailtypack` implemented a function to predict marginal survival; for `coxme` and `parfm`, I manually wrote ad-hoc R functions to estimate marginal survival, using numerical integration when required (`quadinf` function from the `pracma` package [130]). All the code required to re-run this simulation study is openly available on my GitHub page: `https://github.com/ellessenne/frailtymcsim`.

5.5.5 *Performance Measures*

The first performance measure of interest is bias, quantifying whether an estimator targets the true value on average. Formally, it is defined as $E(\hat{\beta}) - \beta$, with $\hat{\beta}$ estimates of the parameter β. Second, I am interested in coverage, i.e. the proportion of times the $100 \times (1 - \alpha)\%$ confidence interval $\hat{\beta} \pm Z_{1-\alpha/2} \times SE(\hat{\beta})$ includes the true value β. This allows assessing whether the empirical coverage rate approaches the nominal coverage rate ($100 \times (1 - \alpha)\%$). Finally, I am interested in mean squared error (MSE); MSE is the sum of the squared bias and variance of $\hat{\beta}$ and represents a natural way to integrate both performance measures into one. However, the relative influence of bias and variance of $\hat{\beta}$ varies with the number of simulations making generalising results difficult. Further details on each performance measure are given in Chapter 4 and elsewhere [58, 81].

I also report on convergence rates for each model, and I will include Monte Carlo standard errors for bias, coverage, and MSE to quantify the uncertainty in estimating such performance measures [58, 70].

Finally, to avoid the inflation of summary statistics caused by software packages spuriously declaring convergence, I manually declared as non-converged all the model fits that returned standardised point estimates or standardised standard errors larger than 10 in absolute value. I standardised values using median and interquartile range for robustness.

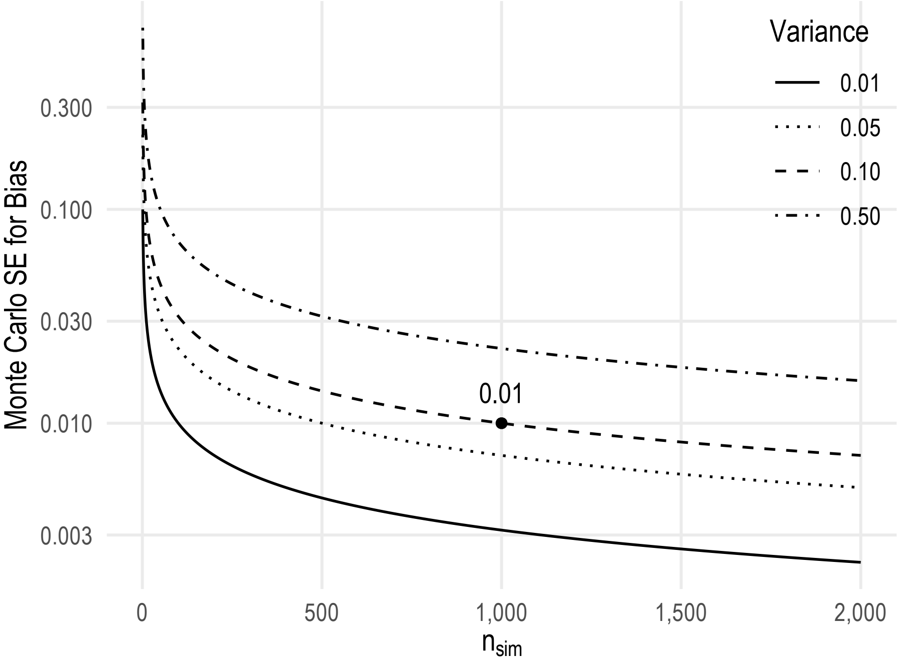


Figure 5.8: Monte Carlo standard error for bias under various combinations of expected bias variance and number of replications

5.5.6 *Number of Simulations*

The choice of the number of replications for this simulations is related to the required degree of precision (and hence the required Monte Carlo standard error).

The key performance measures are bias and coverage; starting with bias, its Monte Carlo standard error can be written as

$$\text{MCSE} = \sqrt{\frac{\text{Var}}{n_{\text{sim}}}} \tag{5.22}$$

Re-arranging the equation, I obtain

$$n_{\text{sim}} = \frac{\text{Var}}{\text{MCSE}^2} \tag{5.23}$$

The expected Monte Carlo standard error for bias as a function of n_{sim} and expected variance is illustrated in Figure 5.8. Assuming a variance of 0.1 for the estimated bias, the expected Monte Carlo standard error with 1,000 replications would be 0.01: I deem this to be acceptable.

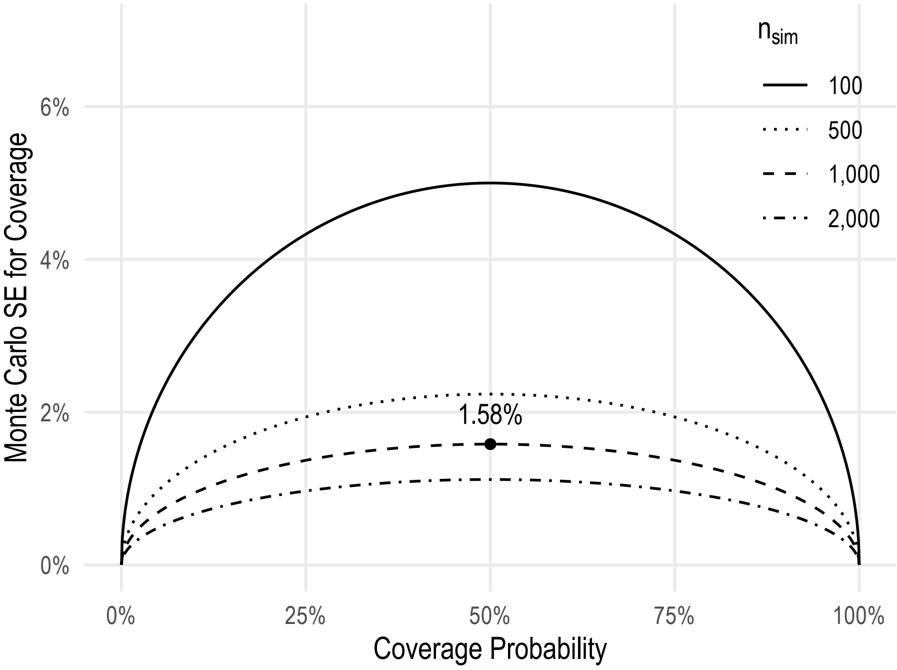


FIGURE 5.9: Monte Carlo standard error for coverage probability under various combinations of expected coverage and number of replications

Analogously for coverage, the Monte Carlo standard error can be written as

$$\text{MCSE} = \sqrt{\frac{\text{Coverage} \times (1 - \text{Coverage})}{n_{\text{sim}}}} \tag{5.24}$$

Re-arranging the equation, I obtain

$$n_{\text{sim}} = \frac{\text{Coverage} \times (1 - \text{Coverage})}{\text{MCSE}^2} \tag{5.25}$$

The expected Monte Carlo standard error for coverage as a function of n_{sim} and expected coverage probability is illustrated in Figure 5.9. The Monte Carlo standard error for coverage is maximised when coverage is 50%; with $1,000$ replications, the expected Monte Carlo standard error for coverage in the worst-case scenario would be 1.58%. Should coverage be optimal at 95%, the expected Monte Carlo standard error would be 0.68%: I deem this expected Monte Carlo standard error acceptable.

In conclusion, I run 1,000 replications for each scenario of this simulation study; the expected Monte Carlo standard error for bias is 0.01, while the expected Monte Carlo standard error for coverage probability is $\leq$ 1.58%.

5.5.7 *Results*

Convergence Rates

Convergence rates were good for most models and scenarios, with 75% of model–scenarios combinations showing a convergence rate of 98% or above (Figure 5.10). However, some exceptions could be found (Figure 5.11):

1. Parametric models with a Gompertz baseline hazard had the worst convergence rates, with a median convergence rate of 43.20% (inter-quartile range: 29.25% – 55.85%);
2. Parametric models with a Weibull baseline hazard and a log-Normal frailty fitted using the `frailtypack` package caused R to hang indefinitely in several scenarios, yielding a median convergence rate of 81.00% (inter-quartile range: 65.30% – 97.30%);
3. `frailtypack` models with a smooth baseline hazard modelled using M-splines showed low convergence rates for other scenarios as well, especially when simulating heterogeneity from a mixture Normal frailty distribution;
4. Some `frailtypack` models and some parametric models with a Gompertz baseline hazard did not converge at all in some scenarios with simulated data assuming a mixture Normal frailty.

Furthermore, I include in Figure 5.12 convergence rate ratios for each data-generating mechanism and model included in this simulation study. Convergence rate ratios were estimated using a Poisson model, fitted using quasi-likelihood to account for over-dispersion, and using robust standard errors [36]; I set as reference levels all levels of each DGM with the highest average convergence rate (sample size of 750 clusters of 2 individuals each, frailty variance of 0.25, Gamma frailty distribution, Gompertz baseline hazard function, and Weibull parametric model with a Gamma frailty). Using a type-II Anova based on the F-test (as suggested by Hastie and Pregibon [141] for settings where a dispersion parameter is estimated from data), the following factors result associated with the convergence rate: sample size (p-value of 0.04), frailty variance (p-value of < 0.01), frailty distribution (p-value of < 0.01), and model (p-value of < 0.01). The Gompertz models showed the worst convergence rate ratios (< 0.50), with the M-splines model

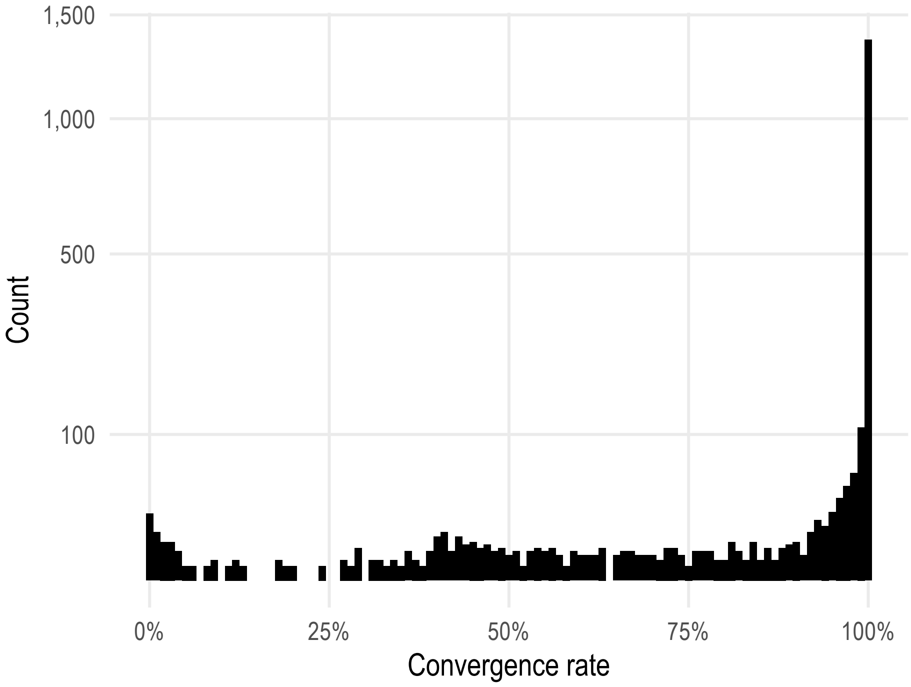


Figure 5.10: Histogram of convergence rates per each model–scenario combination, Monte Carlo simulation on impact of model misspecification in shared frailty survival models

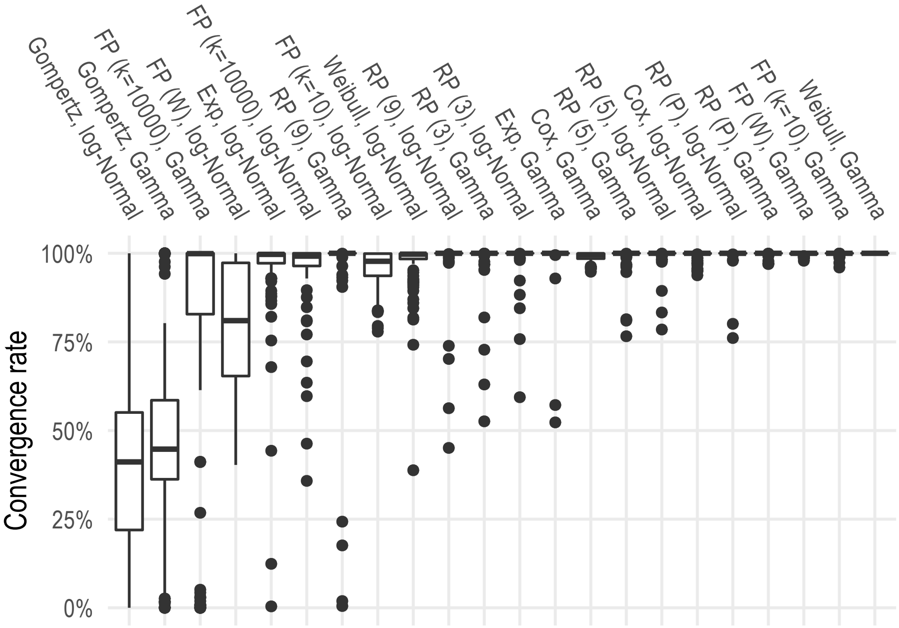


Figure 5.11: Distribution of convergence rates by model, Monte Carlo simulation on impact of model misspecification in shared frailty survival models. Boxplots are sorted by median convergence rate

with a Gamma frailty and κ = 10,000 and the Weibull model with a log-Normal frailty fit using the `frailtypack` being second and third worst, respectively. The Royston-Parmar model with 9 degrees of freedom and a Gamma frailty, the M-splines model with either κ = 10, 10,000 and a log-Normal frailty, and the exponential model with a log-Normal frailty showed reduced convergence rate ratios compared to the reference as well. The rate ratios of convergence also decreased as the variance of the frailty increased, and when simulating from a mixture Normal distribution. Finally, simulating from a Weibull or Weibull-Weibull (2) baseline hazard function and simulating 20 clusters of 150 individuals each also yielded slightly reduced convergence rate ratios.

Finally, Figure 5.13 depicts convergence rates for every scenario against the average proportion of observed events. Models with worse convergence rates showed an association between non-convergence and the average proportion of events, with stronger right censoring being associated with worse convergence rates.

Ultimately, the factors that seemed to be associated with convergence rates were the censoring proportion, the variance of the frailty, and the distribution of the frailty. Nevertheless, the software implementation and the algorithms used for fitting each model seem to play an important role, with some software implementations being more robust than others to variations in the factors outlined before.

Results for the Regression Coefficient

Bias, and coverage probability for scenarios with 20 clusters of 150 individuals each and scenarios with 750 clusters of 2 individuals each are presented in Figures 5.14, 5.16, and 5.15, 5.17, respectively.

With a simple, exponential true baseline hazard all models performed equally well in terms of bias and coverage, with minimal bias in scenarios simulated from a mixture Normal frailty distribution and consequently sub-optimal coverage in the same settings. Conversely, assuming a too simple parametric distribution with a more complex true baseline hazard (or misspecifying the baseline hazard) yielded a biased regression coefficient: significant positive bias up to 0.20 and negative bias up to -0.12 for models assuming an exponential baseline hazard, and positive bias up to 0.21 and negative bias up to -0.12 for models assuming a Gompertz baseline hazard. A positive bias of

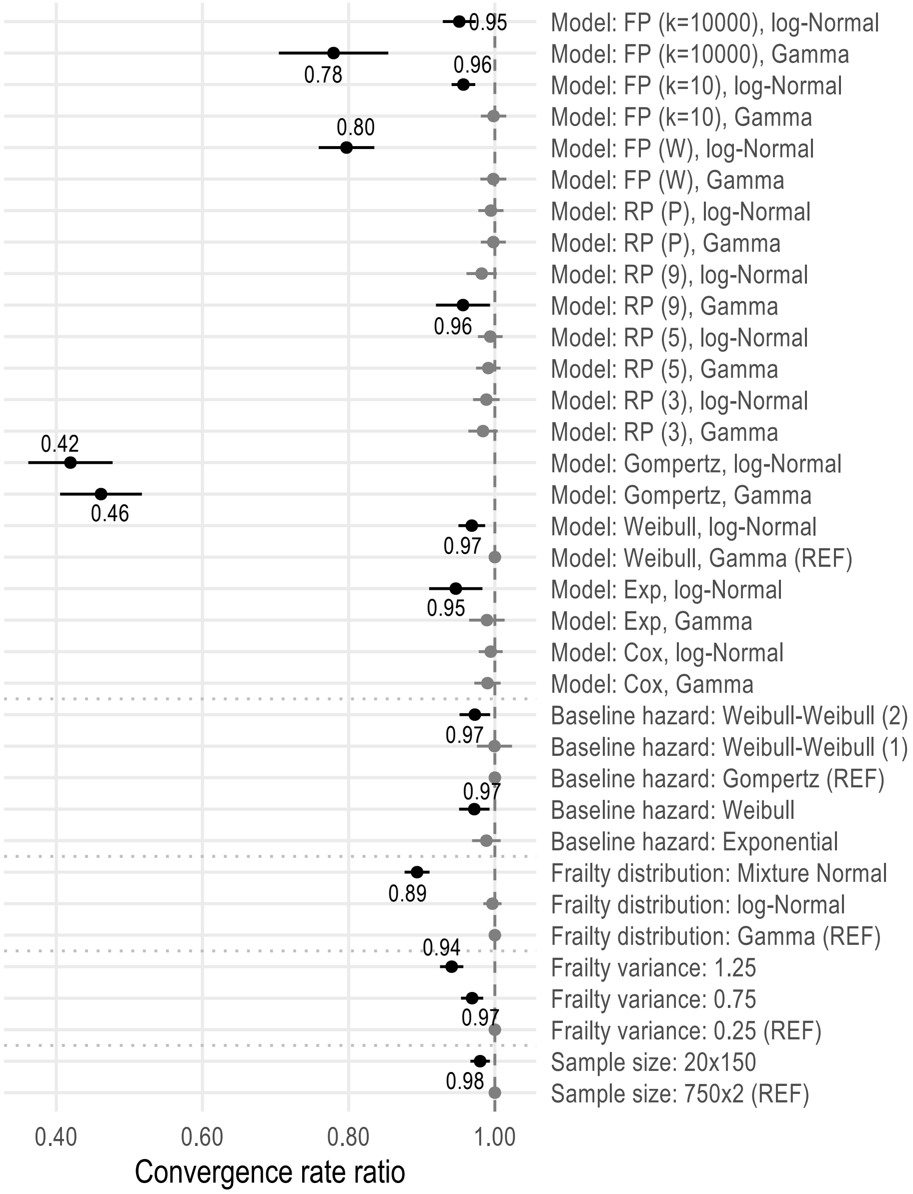


FIGURE 5.12: Convergence rate ratios by data-generating factor, Monte Carlo simulation on impact of model misspecification in shared frailty survival models. Convergence rate ratios were estimated using quasi-Poisson regression; values depicted in black and labelled were statistically significant, assuming $\alpha = 0.05$

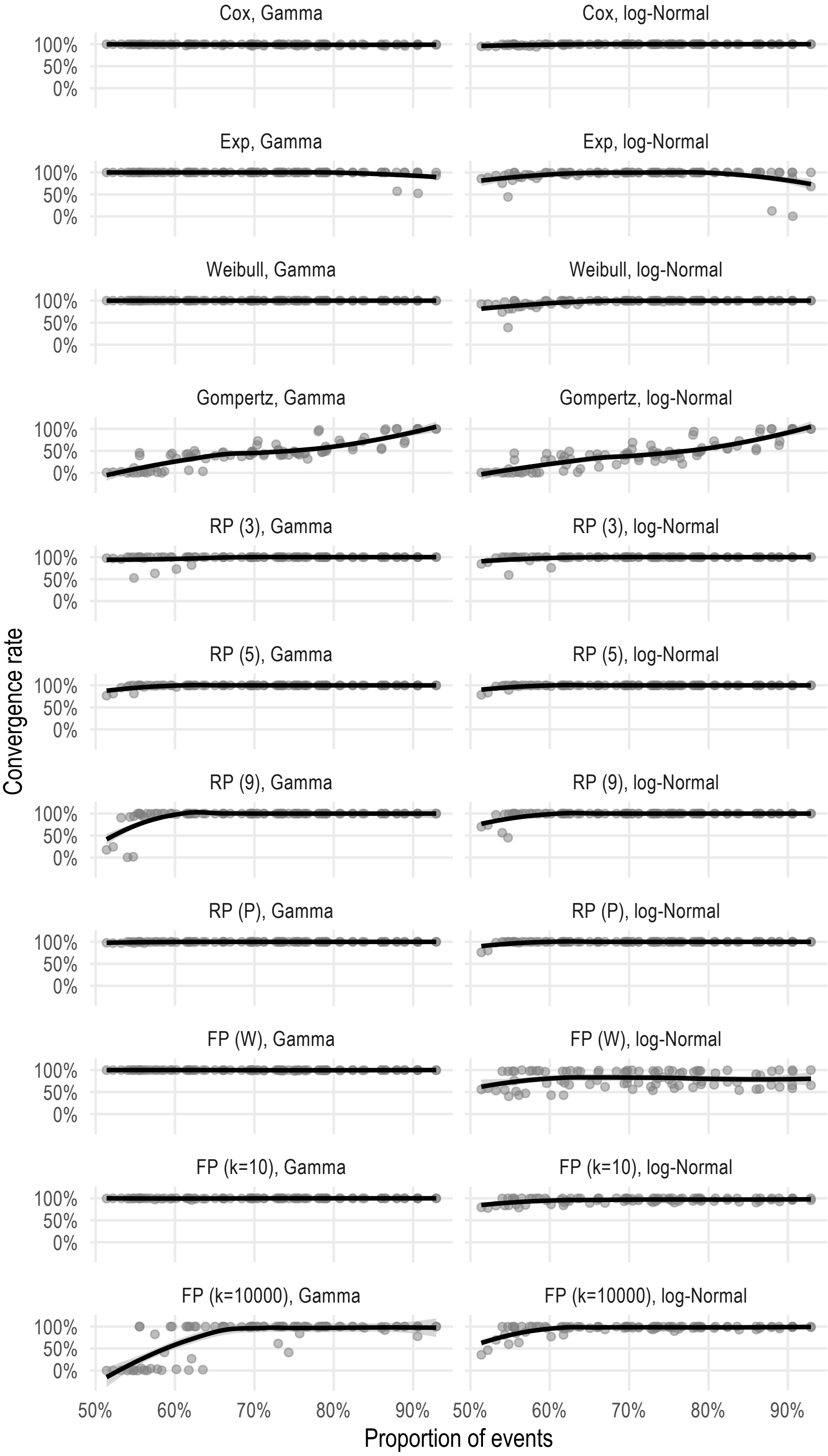


FIGURE 5.13: Convergence rates versus average proportion of events per simulated scenario. Model baseline hazards identify rows, while the model frailty identifies columns

0.21 on the log-hazard ratio scale corresponds to a 23% relative risk overestimation; a negative bias of -0.12 corresponds to an 11% relative risk underestimation. Conversely, models assuming a Weibull baseline hazard performed slightly better in these simulated settings, with positive bias up to 0.05 and negative bias up to -0.06. The semiparametric Cox models and all the flexible parametric models (irrespectively of the number of degrees of freedom employed and of the estimation procedure) yielded unbiased results, with the exception of the model with 9 degrees of freedom and a log-Normal frailty in four scenarios with a frailty simulated from a mixture Normal distribution with component-specific variances of 1.25 and assuming a Weibull and a mixture Weibull (2) true baseline hazard. In those scenarios, the flexible parametric model yielded large biases of 0.18 to 0.40, although this seems to be a somewhat spurious result given the performance of the same method in other scenarios. All models using M-splines on the hazard scale performed similarly to the parametric Weibull, with little to no bias; however, the performance of models using M-splines worsened with a true mixture Normal frailty distribution. Coverage was optimal for all models producing unbiased estimates; conversely, coverage dropped considerably for models that yielded biased estimates with coverage values as low as 5% for models showing the largest bias.

Interestingly, misspecification of the frailty distribution did not affect much the pattern of results; despite that, bias seemed to worsen when the frailty was simulated from a log-Normal or mixture Normal distribution compared to a Gamma distribution, exacerbating the effect of misspecifying the baseline hazard. In addition to that, in scenarios with 750 clusters of 2 individuals each the performance of most methods worsened compared to the settings of 20 clusters of 150 individuals, especially when simulating from a mixture Normal frailty distribution. For instance, the Cox model with a log-Normal frailty yielded positive bias up to 0.12 in scenarios with a frailty simulated from a mixture Normal distribution with component-specific variances of 1.25. Other than that, the patterns of results from scenarios with different sample sizes are pretty much comparable.

Finally, mean squared errors are presented in Appendix E, as Figures E.1 and E.2. The models that showed the lowest MSE were the semiparametric models, the flexible parametric models, and the models using M-splines - irrespectively of the true baseline hazard and distribution of the frailty. The exponential and Gompertz parametric models showed a larger MSE - up to 10-fold larger - when the baseline hazard was misspecified, with the Gompertz model showing an MSE larger than semiparametric and flexible parametric models even when well specified. The Weibull model, as observed before, performed similarly to semiparametric and flexible parametric models.

Results for the 5-years LLE

Bias and coverage probability for the 5-years LLE are presented in Figures 5.18, 5.20 for scenarios with 20 clusters of 150 individuals each, and in Figures 5.19, 5.21 for scenarios with 750 clusters of 2 individuals each.

The pattern of results for LLE mirrors the pattern observed for the regression coefficient: models with a misspecified baseline hazard (or a baseline hazard not flexible enough to capture the underlying shape) yielded biased results, both positive and negative. Negative bias was up to -0.06 and positive bias was up to 0.20: this corresponds, respectively, to a difference of approximately (minus) 1 month and 2 and a half months in the estimated LLE. The Weibull model with a log-Normal frailty fit with `frailtypack` largely underestimated the 5-years LLE in all scenarios with a true Gamma or log-Normal frailty and a sample size of 20 clusters of 150 individuals (negative bias between -0.18 and -0.10); in scenarios with 750 clusters of 20 individuals, the same model performed better with minimal to no bias. In the same aforementioned settings, M-splines model with smoothing parameter κ = 10,000 and a Gamma distribution performed even worse when the true baseline hazard followed a Weibull-Weibull (1) distribution (negative bias between -0.46 and -0.17). The large bias observed for this M-splines model in these scenarios seems to be spurious - analogously as before with the flexible parametric model. Interestingly, models with a well-specified frailty seemed to perform better than models with a misspecified frailty, both for a true Gamma and log-Normal frailty, and especially when the frailty variance was large (1.25). When simulating from a mixture Normal distribution all models performed poorly with positively biased results (up to 0.26, e.g. 3 months) with exceptions being the `frailtypack` models described before, where underestimation of the results still applied. Coverage followed a similar pattern, with optimal coverage for models with small bias and reduced coverage for models that yielded biased results; overall, coverage was better when the frailty distribution was well specified. As a consequence of the large positive bias, coverage in scenarios simulated from a mixture Normal distribution was

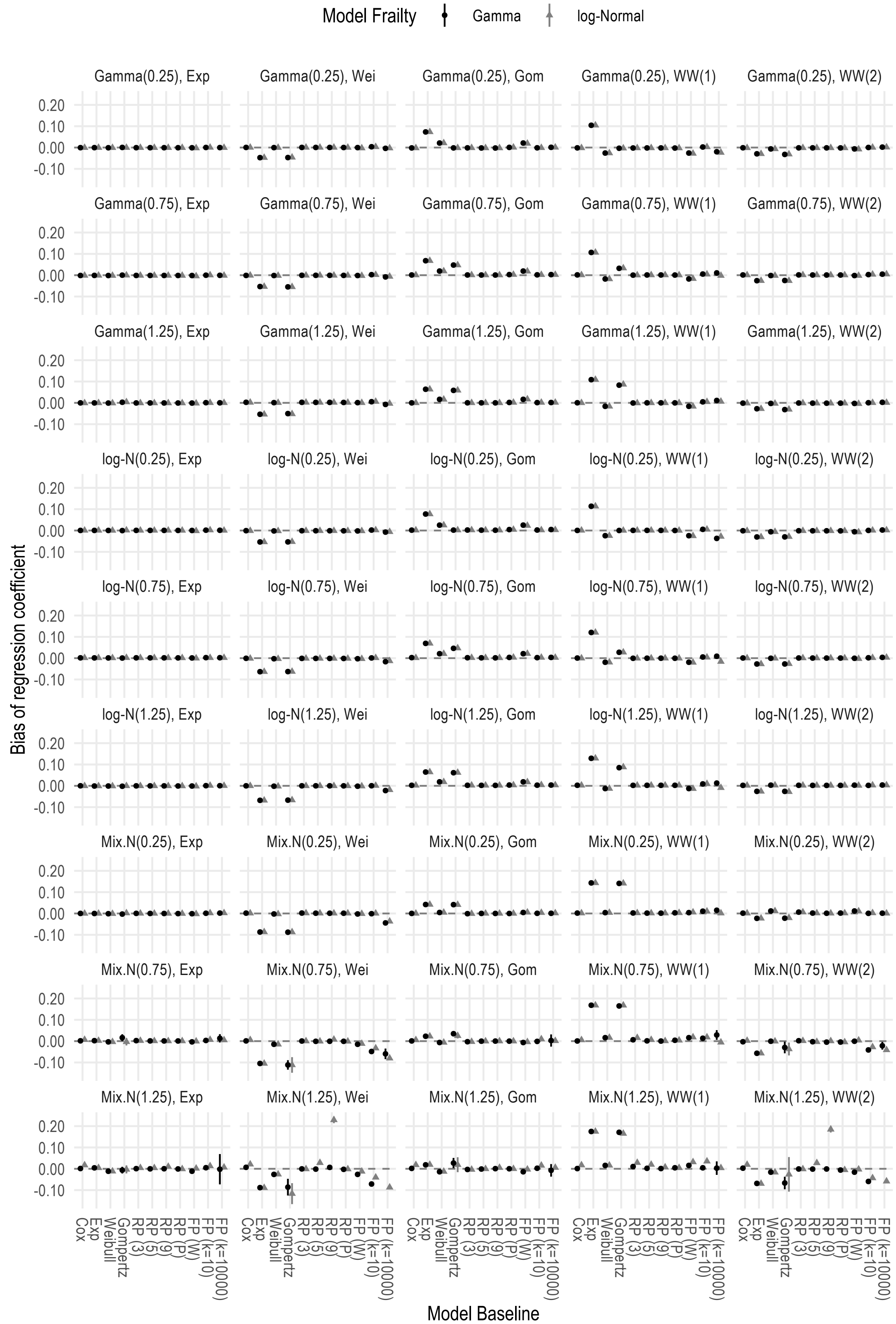


Figure 5.14: Bias of regression coefficient, scenarios with 20 clusters of 150 individuals each. Colours represent the model frailty, and each subplot includes results for a given combination of data-generating baseline hazard and frailty

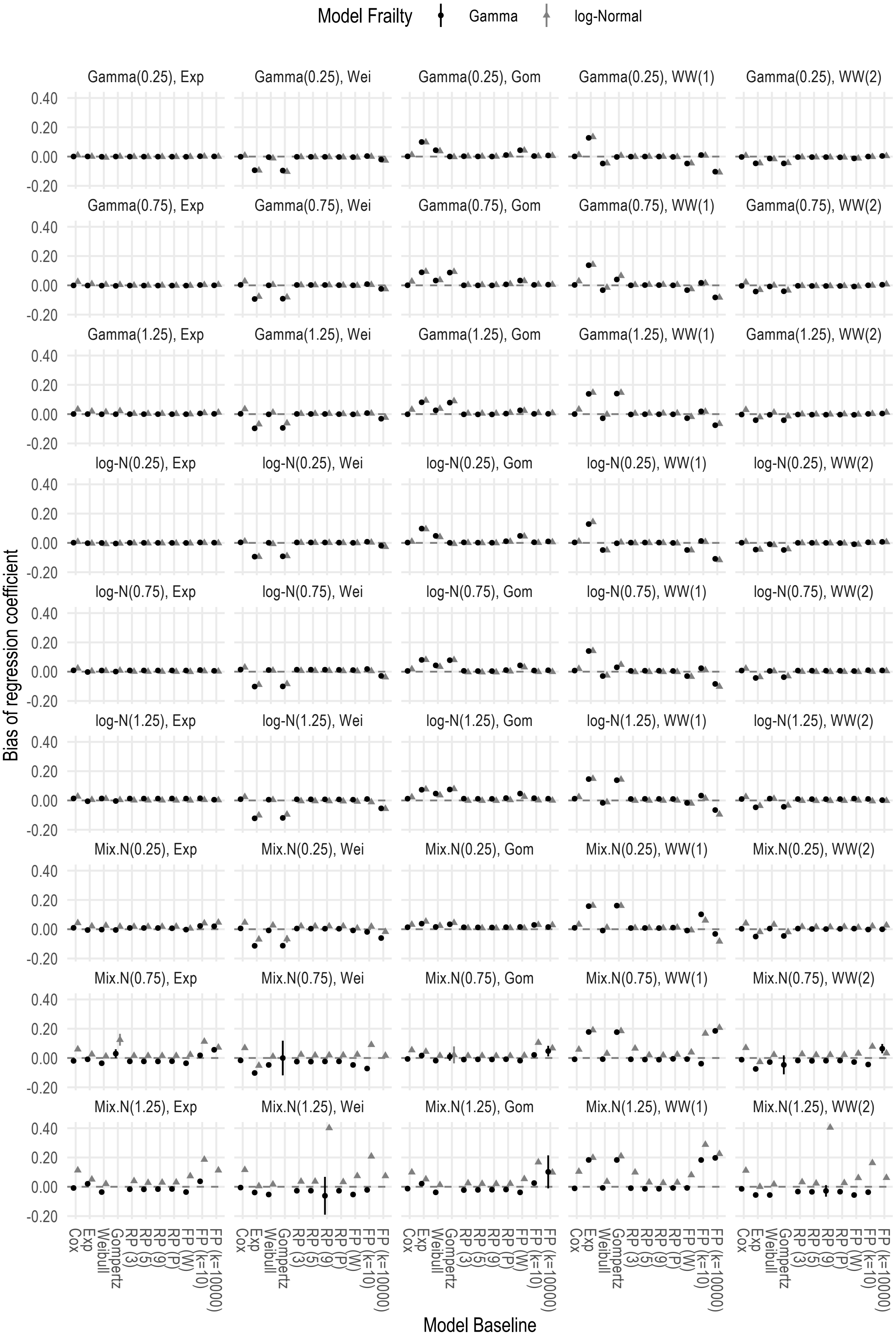


FIGURE 5.15: Bias of regression coefficient, scenarios with 750 clusters of 2 individuals each. Colours represent the model frailty, and each subplot includes results for a given combination of data-generating baseline hazard and frailty

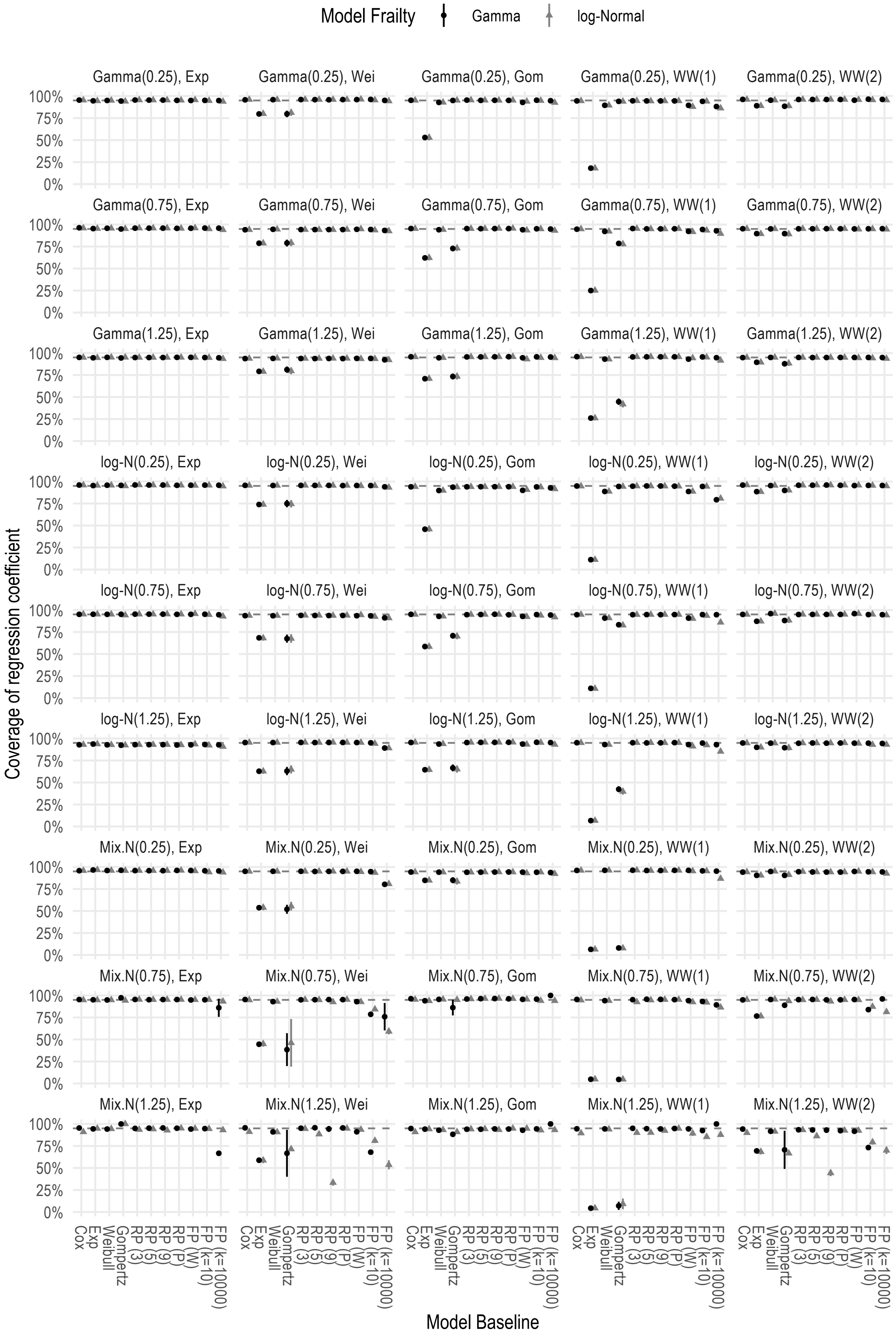


Figure 5.16: Coverage of regression coefficient, scenarios with 20 clusters of 150 individuals each. Colours represent the model frailty, and each subplot includes results for a given combination of data-generating baseline hazard and frailty

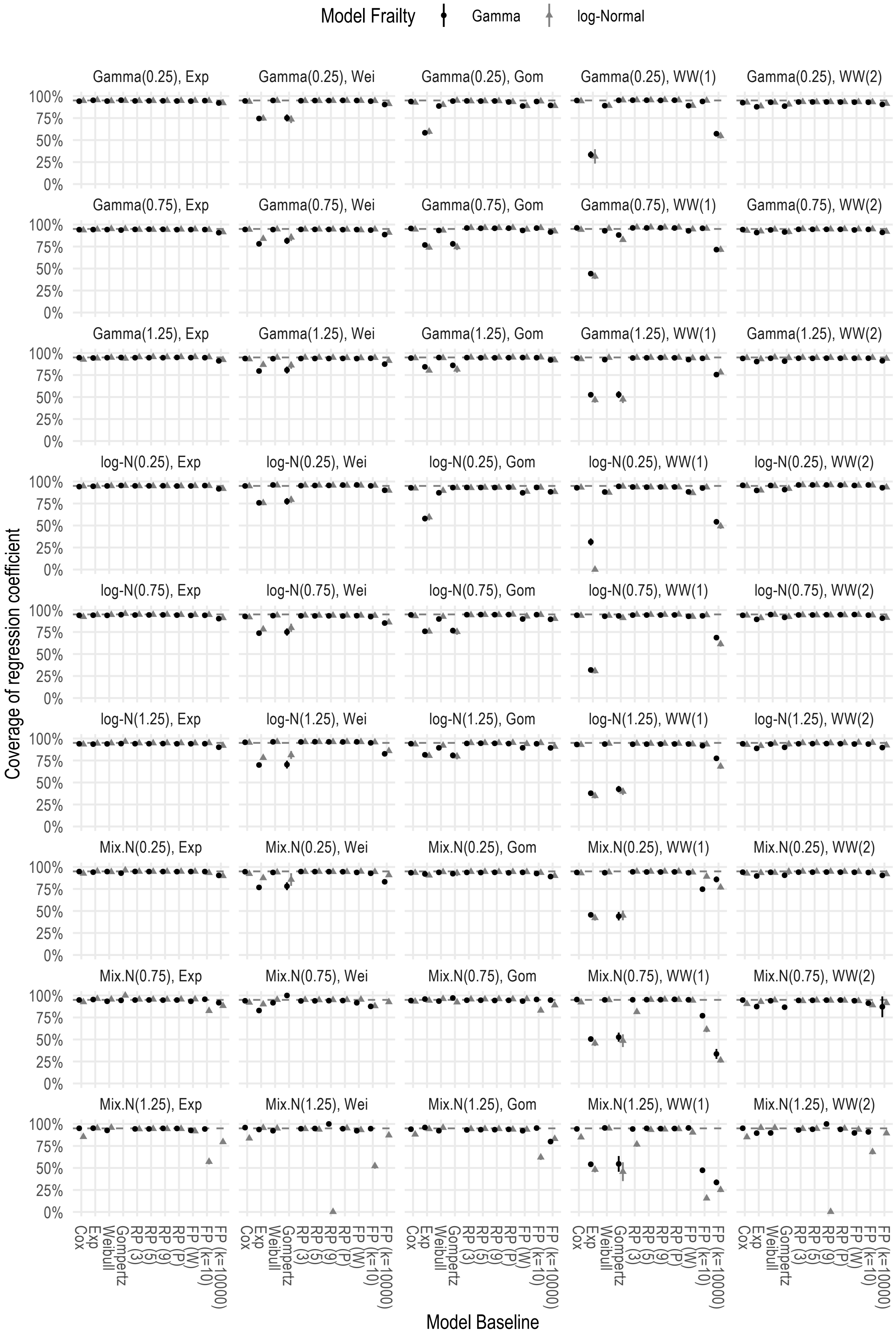


FIGURE 5.17: Coverage of regression coefficient, scenarios with 750 clusters of 2 individuals each. Colours represent the model frailty, and each subplot includes results for a given combination of data-generating baseline hazard and frailty

poor.

Sample size seemed to affect some patterns, as described above; however, misspecification of the baseline hazard and/or the frailty distribution seemed to matter more. Higher degrees of heterogeneity seemed to exacerbate the impact of misspecifying the frailty distribution.

Mean squared errors are included in Appendix E as before, as Figures E.3 and E.4. MSE was similar across all scenarios with variability following the pattern described above for bias - which seemed to be the main component driving the magnitude of MSE. For instance, models fitted using M-splines for the baseline hazard and a log-Normal frailty had a much higher MSE - approximately 10 times larger - in the aforementioned scenarios. Once again, in scenarios simulated from a mixture Normal frailty, MSE was the largest across the board.

Results for the Frailty Variance

Results for the frailty variance are included in Appendix E. In particular, bias and coverage probabilities for scenarios with 20 clusters of 150 individuals each are presented in Figures E.5 and E.7, while bias and coverage probabilities for scenarios with 750 clusters of 2 individuals each are presented in Figures clusters E.6 and E.8. Mean squared errors are included in Figures E.9 and E.10. As I mentioned in Section 5.5.3, results for when the frailty variance is misspecified are not included since the frailty is modelled on different scales.

The variance of Gamma frailties was generally well estimated, with slight bias or no bias at all; the exception was the model with M-splines and κ = 10,000, which yielded largely biased results when the true baseline hazard function followed a Weibull-Weibull (1) distribution. Parametric frailty models, especially the exponential and Gompertz models, were the methods that yielded slight bias when the baseline hazard did not follow an exponential distribution.

A similar pattern of results could be observed for log-Normal frailties, with exponential and Gompertz models yielding slightly biased results. Models with M-splines and κ = 10,000 yielded biases of similar magnitude when the baseline hazard followed a Weibull-Weibull (1) distribution. Interestingly, the Weibull model fitted using `frailtypack`

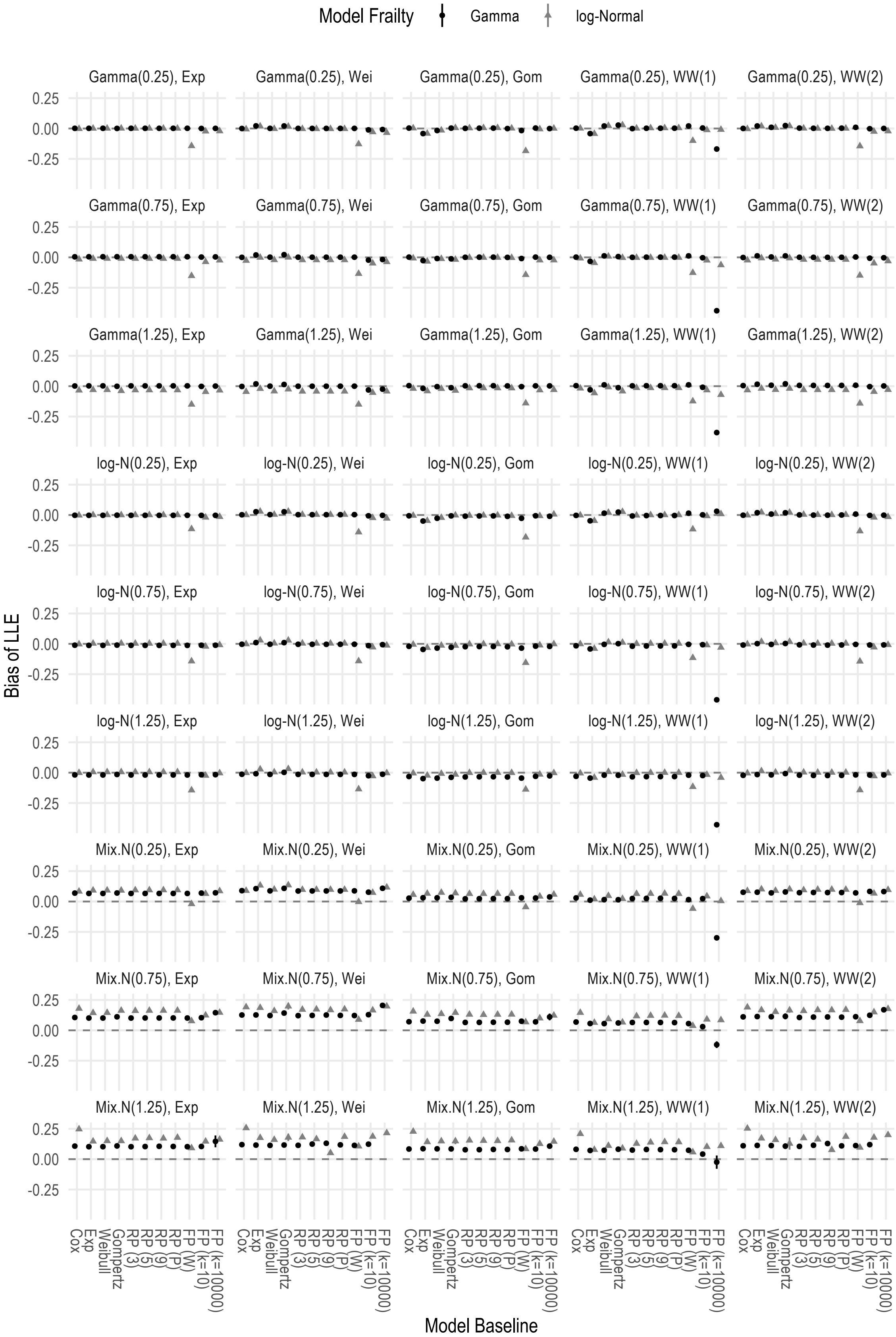


FIGURE 5.18: Bias of LLE, scenarios with 20 clusters of 150 individuals each. Colours represent the model frailty, and each subplot includes results for a given combination of data-generating baseline hazard and frailty

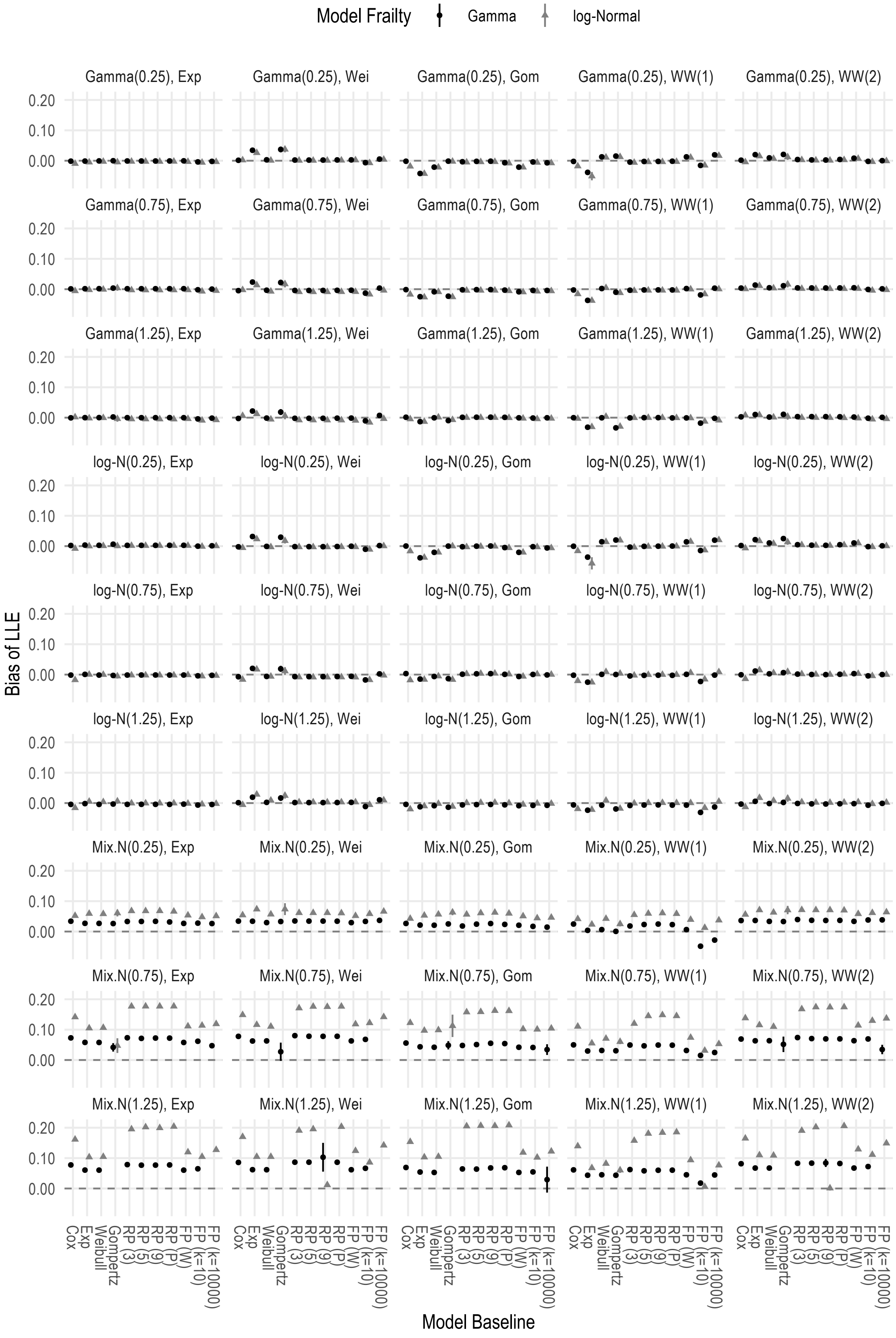


FIGURE 5.19: Bias of LLE, scenarios with 750 clusters of 2 individuals each. Colours represent the model frailty, and each subplot includes results for a given combination of data-generating baseline hazard and frailty

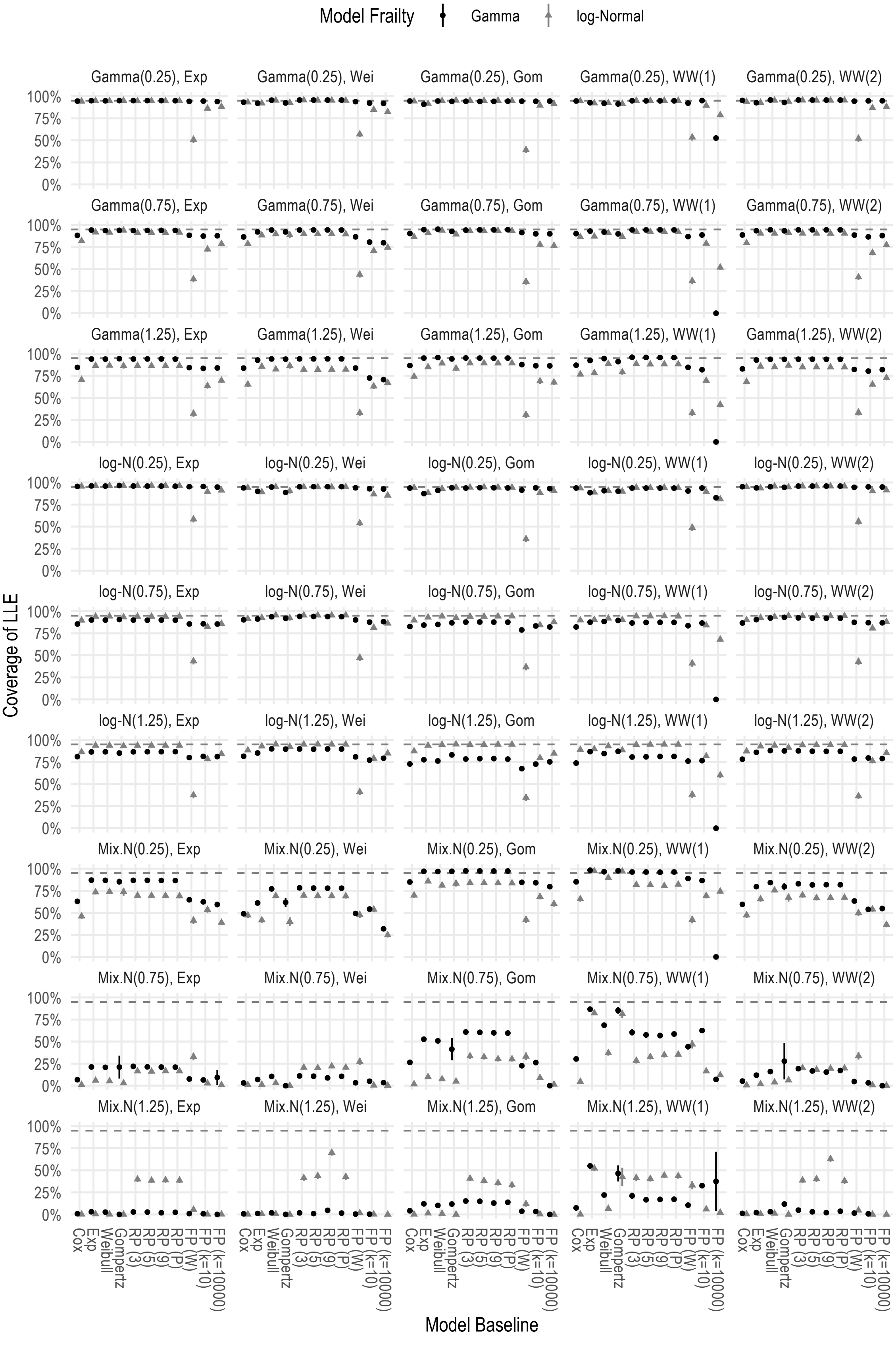


FIGURE 5.20: Coverage of LLE, scenarios with 20 clusters of 150 individuals each. Colours represent the model frailty, and each subplot includes results for a given combination of data-generating baseline hazard and frailty

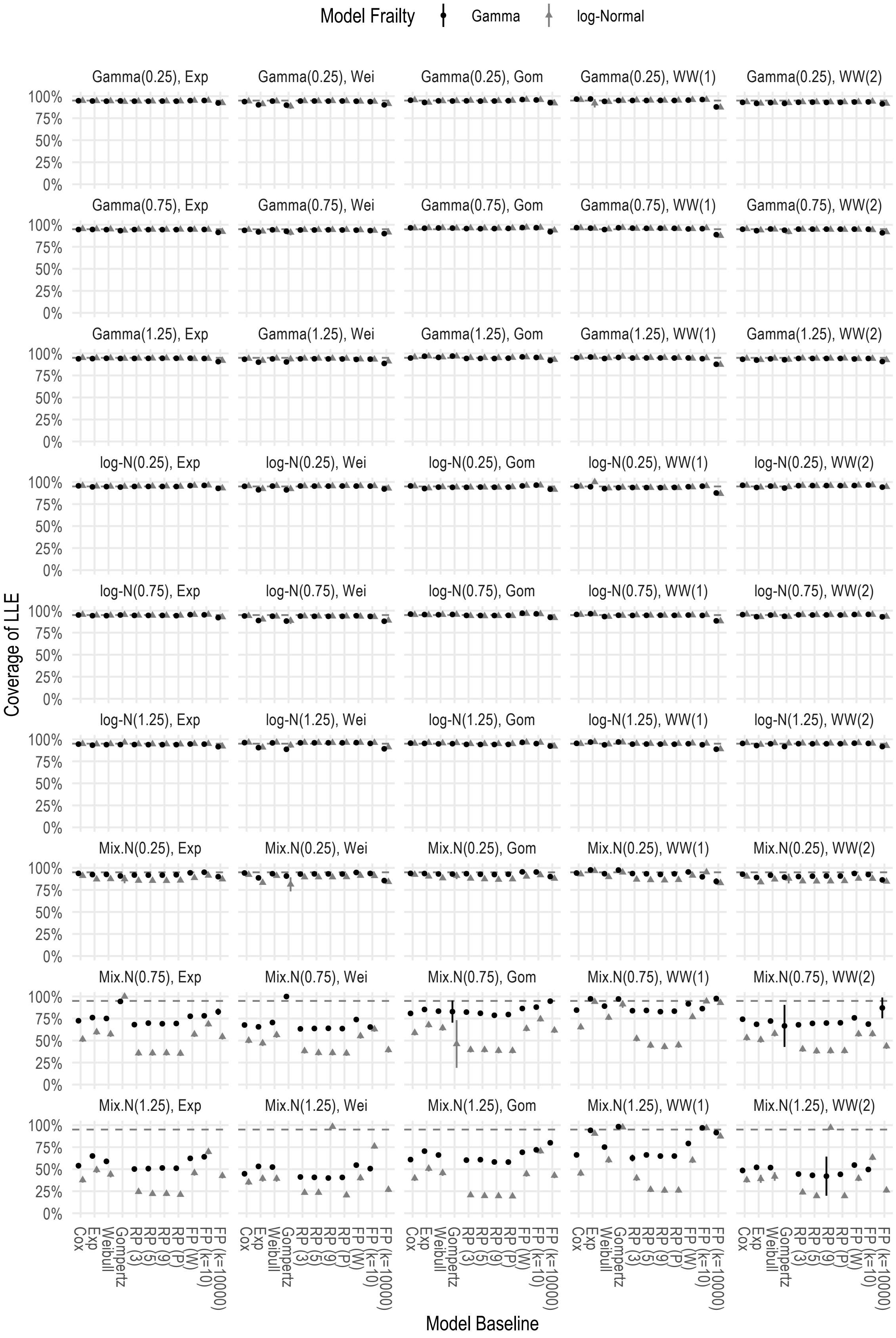


Figure 5.21: Coverage of LLE, scenarios with 750 clusters of 2 individuals each. Colours represent the model frailty, and each subplot includes results for a given combination of data-generating baseline hazard and frailty

yielded large bias in scenarios with 20 clusters of 150 individuals each, up to 2.53.

Coverage probabilities were generally good, except for scenarios with misspecified parametric baseline hazard functions and scenarios with large biases described above. In these settings, coverage was poor and even null (0%) in some scenarios. Overall, coverage was best when the true baseline hazard followed an exponential distribution, and with flexible parametric models (irrespectively of degrees of freedom and estimation method).

Mean squared errors are also included in Appendix E, and depicted in Figures E.9 and E.10; once again, MSEs were mostly driven by the large biases of some models (in some scenarios), with comparable values everywhere else.

5.6 Application to PSP-CKD Data

In this Section I will illustrate the results of the simulation study in practice using data from the PSP-CKD study, a pragmatic trial on chronic kidney disease previously described in Chapter 3.

The outcome of interest here is kidney failure, focussing on the cause-specific hazard for simplicity. The exposure of interest is enhanced CKD care (against standard care), and I will include age at baseline and sex as covariates. Data is clustered within primary care practices, hence I will include a frailty term to account for the correlation of individuals belonging to the same practice.

I will report estimates of treatment effect and three-years LLE; LLE can be interpreted as the difference in expected time to kidney failure between individuals receiving enhanced CKD care and individuals receiving traditional care over three years. I will estimate LLE for individuals with a given covariates pattern: in particular, I estimate LLE for individuals with median age (75 years) and the most frequent gender (females).

I fit a model of the kind:

$$h(t_{ij}) = \alpha_j h_0(t_{ij}) \exp(\beta_1 \times \text{treatment}_j + \beta_2 \times \text{age}_{ij} + \beta_3 \times \text{sex}_{ij}),$$

with i and j identifying individuals and clusters, respectively. treatment_j is the treatment modality that the j^{th} practice was randomised to, age_{ij} and sex_{ij} are age at baseline and sex for the i^{th} individual in the j^{th} practice. α_j is the frailty for the j^{th} practice.

I model the baseline hazard $h_0(\cdot)$ via fully parametric and flexible parametric distributions or by leaving it unspecified, as in the simulations of Section 5.5. The flexible parametric models are modelled on the log-cumulative hazard scale:

$$\log H(t_{ij}) = s(\log(t_{ij})|\gamma, k_0) + \beta_1 \times \text{treatment}_j + \beta_2 \times \text{age}_{ij} + \beta_3 \times \text{sex}_{ij} + \eta_i,$$

with $s(\cdot)$ a spline function of log-time with parameter vector γ and knot vector k_0. Despite being on the log-cumulative hazard scale, the aforementioned model is still a proportional hazards model. I model the frailty distribution assuming either a Gamma or log-Normal distribution.

Finally, I also fit models with the baseline hazard modelled using M-splines and a smoothing parameter κ selected via cross-validation, assuming either a Gamma or log-Normal frailty. This model can be fit using the R package `frailtypack`.

Results from each model fit are included in Table 5.3 and plotted in Figure 5.22. Most models yield comparable estimates for the log-treatment effect (and therefore the corresponding hazard ratio), except models fitted using the `frailtypack` package: models with a log-Normal frailty yielded a lower estimate, while models with a Gamma frailty yielded larger estimates - with the only exception being the Weibull model with a Gamma frailty which yielded comparable results to other methods. Estimates for LLE followed a similar pattern; however, Gompertz and Exponential models yielded larger estimates irrespectively of the frailty distribution.

AIC and BIC for models using full likelihood are included in Table 5.4, with best AIC and BIC in bold. The best model according to AIC is the flexible parametric model with 3 degrees of freedom and a Gamma frailty; conversely, according to the BIC, the best model is the Weibull model with a log-Normal frailty. However, other information criteria are available - especially for models with random effects: for instance, Vaida and Blanchard [142] suggested the use of conditional AIC (cAIC) for model selection in linear mixed models. They demonstrated that a classical AIC (i.e. a marginal AIC) and its small sample correction are inappropriate when the interest is on clusters, see also Liang, Wu, and

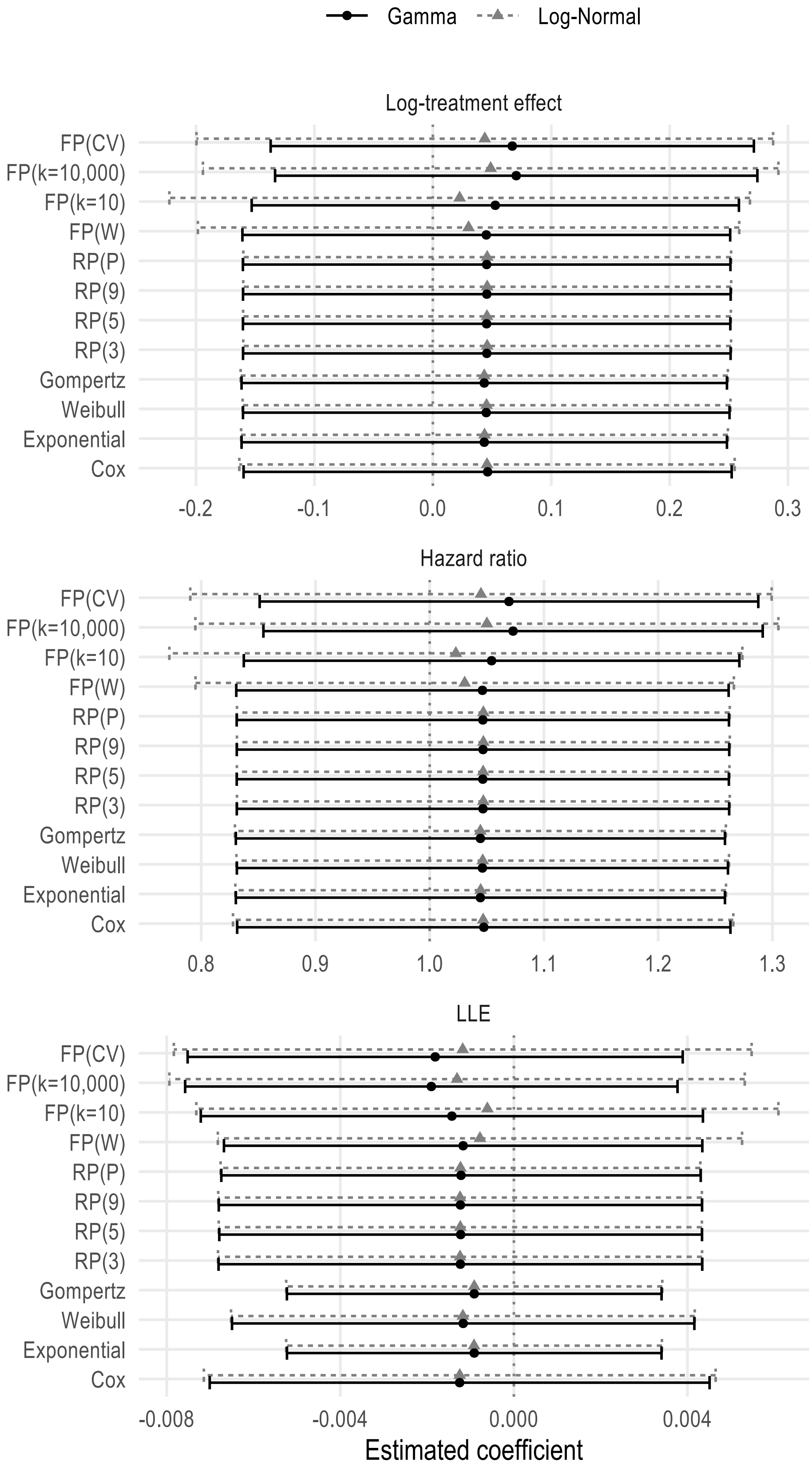


Figure 5.22: Results: application of shared frailty survival models to PSP-CKD data

Table 5.3: Results: application of shared frailty models to PSP-CKD data. Values in brackets are standard errors for each estimate

Baseline	Log-treatment effect	Hazard ratio	LLE
Gamma frailty:			
Cox	0.0463 (0.1052)	1.0473 (0.1101)	-0.0013 (0.0029)
Exponential	0.0434 (0.1046)	1.0444 (0.1092)	-0.0009 (0.0022)
Weibull	0.0451 (0.1048)	1.0461 (0.1097)	-0.0012 (0.0027)
Gompertz	0.0434 (0.1046)	1.0444 (0.1092)	-0.0009 (0.0022)
RP(3)	0.0456 (0.1050)	1.0466 (0.1099)	-0.0012 (0.0028)
RP(5)	0.0454 (0.1050)	1.0465 (0.1099)	-0.0012 (0.0028)
RP(9)	0.0456 (0.1051)	1.0467 (0.1100)	-0.0012 (0.0028)
RP(P)	0.0454 (0.1050)	1.0465 (0.1099)	-0.0012 (0.0028)
FP(W)	0.0451 (0.1051)	1.0462 (0.1099)	-0.0012 (0.0028)
FP(k=10)	0.0527 (0.1050)	1.0542 (0.1107)	-0.0014 (0.0030)
FP(k=10,000)	0.0704 (0.1039)	1.0730 (0.1115)	-0.0019 (0.0029)
FP(CV)	0.0671 (0.1041)	1.0694 (0.1113)	-0.0018 (0.0029)
Log-normal frailty:			
Cox	0.0457 (0.1067)	1.0468 (0.1117)	-0.0012 (0.0030)
Exponential	0.0436 (0.1049)	1.0446 (0.1096)	-0.0009 (0.0022)
Weibull	0.0454 (0.1051)	1.0464 (0.1100)	-0.0012 (0.0027)
Gompertz	0.0434 (0.1050)	1.0444 (0.1096)	-0.0009 (0.0022)
RP(3)	0.0459 (0.1051)	1.0469 (0.1100)	-0.0012 (0.0028)
RP(5)	0.0457 (0.1051)	1.0468 (0.1100)	-0.0012 (0.0028)
RP(9)	0.0459 (0.1051)	1.0470 (0.1101)	-0.0012 (0.0028)
RP(P)	0.0459 (0.1051)	1.0470 (0.1100)	-0.0012 (0.0028)
FP(W)	0.0302 (0.1166)	1.0306 (0.1202)	-0.0008 (0.0031)
FP(k=10)	0.0226 (0.1251)	1.0229 (0.1280)	-0.0006 (0.0034)
FP(k=10,000)	0.0488 (0.1240)	1.0500 (0.1302)	-0.0013 (0.0034)
FP(CV)	0.0439 (0.1242)	1.0449 (0.1298)	-0.0012 (0.0034)

TABLE 5.4: Results: application of shared frailty models to PSP-CKD data, comparison of AIC/BIC. Best AIC/BIC values are in bold

Baseline	AIC	BIC
Gamma frailty:		
Cox	—	—
Exponential	5,654.16	5,694.96
Weibull	5,563.91	5,612.88
Gompertz	5,656.16	5,705.13
RP(3)	**5,554.48**	5,619.77
RP(5)	5,557.33	5,638.94
RP(9)	5,564.54	5,678.80
RP(P)	—	—
FP(W)	5,563.91	5,612.88
FP(k=10)	—	—
FP(k=10,000)	—	—
FP(CV)	—	—
Log-normal frailty:		
Cox	—	—
Exponential	5,654.14	5,694.94
Weibull	5,563.89	**5,612.86**
Gompertz	5,656.14	5,705.11
RP(3)	5,554.48	5,619.77
RP(5)	5,557.34	5,638.95
RP(9)	5,564.55	5,678.81
RP(P)	—	—
FP(W)	5,564.94	5,613.91
FP(k=10)	—	—
FP(k=10,000)	—	—
FP(CV)	—	—

Zou [143]. Unfortunately, the cAIC is not routinely reported by software fitting shared frailty models (except for `frailtyHL` [135]), making the use of cAIC for selecting the best fitting model a much harder task.

To decide which model to select as the model that fits the data best, I present non-parametric smoothed hazard using the method of Rebora *et al.* [144]. Smoothed overall hazard and treatment-specific hazards are included in Figure 5.23. Assuming a Weibull hazard function would fit the data fairly well; however, (1) the hazard seems to almost plateau towards the end of follow-up, (2) the flexible parametric model with 3 degrees of freedom and a Gamma frailty performs second best in terms of BIC (excluding models with a Weibull baseline hazard), (3) flexible parametric models provide additional advantages e.g. in terms of extrapolation, and (4) the `rstpm2` package is much more

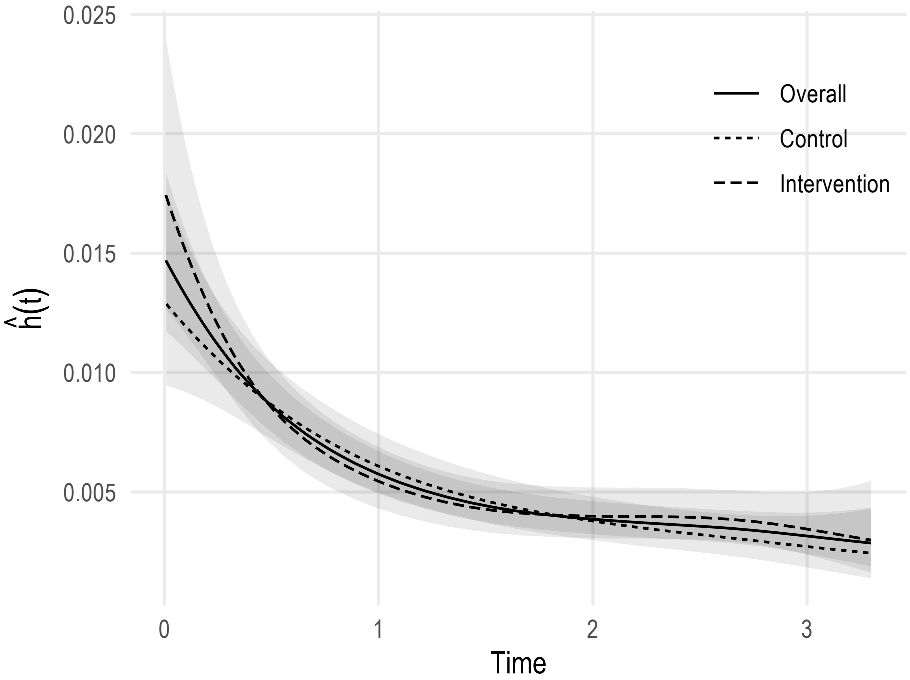


FIGURE 5.23: Application of shared frailty models to PSP-CKD data, smoothed non-parametric hazard estimate overall and by treatment modality

flexible than `parfm` in terms of predictions. For all of the reasons above, I choose the flexible parametric model with 3 degrees of freedom and a Gamma frailty as the best model for this applied example.

The best model yields a hazard ratio for treatment of 1.05 (95% confidence interval: 0.83 – 1.26). The risk of kidney failure seems to not be significantly affected by enhanced CKD care, and in fact the estimated 3-years LLE is -0.0012 with a 95% confidence interval of -0.0068 – 0.0043, assuming years as the unit of time. For instance, the corresponding LLE in days would be -0.45, a number that can be deemed not clinically meaningful.

The flexible parametric model can easily be extended to include time-dependent effects, e.g. to test a time-dependent treatment effect. For instance, I can include an interaction between treatment and the natural logarithm of time, modelled using a natural spline:

$$
\begin{aligned}
\log H(t_{ij}) = & s(\log(t_{ij})|\gamma, k_0) + \beta_1 \times \text{treatment}_j + \beta_2 \times \text{age}_{ij} + \beta_3 \times \text{sex}_{ij} \\
& + \beta_* \times \text{treatment}_j \times s(\log(t_{ij})|\delta, l_0) + \eta_i,
\end{aligned}
$$

where the treatment variable is interacting with a spline function of log-time with associated coefficient vector δ, knots vector l_0, and regression coefficients β_*. Flexible parametric models have been shown to be insensitive to the number of knots utilised

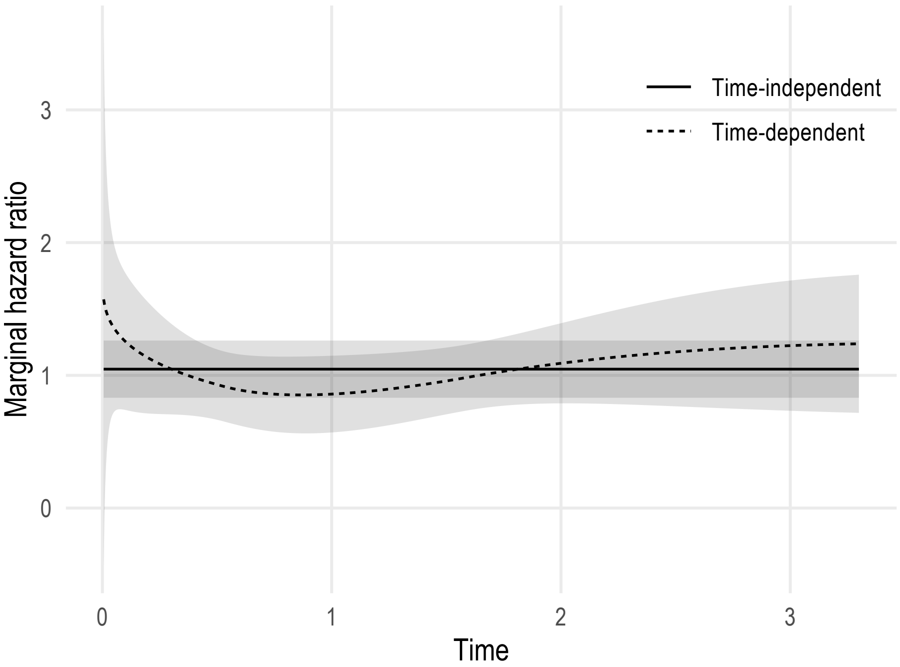


Figure 5.24: Application of shared frailty models to PSP-CKD data, comparison of time-dependent and time-independent marginal hazard ratio for treatment

to model time-varying effects, therefore I am using 3 degrees of freedom for simplicity [145]. The difference between the marginal hazard ratio estimated using the model with a time-dependent treatment effect and the model without is depicted in Figure 5.24, while the difference between the marginal survival difference is included in Figure 5.25. The marginal hazard ratio seems to be higher early on, then decreasing over time until it flattens at approximately 3 years of follow-up; analogously, the marginal survival difference shows a similar pattern.

Despite that, the difference with time-invariant effects of treatment does not seem to be large: I could, therefore, test whether the time-treatment interaction is statistically significant using a likelihood ratio test. The resulting χ^2 test statistic is 2.56 with a p-value of 0.46: this suggest that there is not enough evidence to support the presence of a time-dependent treatment effect.

5.7 Discussion

In observational studies and clinical trials with survival outcomes and an intrinsic hierarchical structure, survival models with shared frailty terms and/or random effects have moved from being a speciality rarely used that requires ad hoc software to being

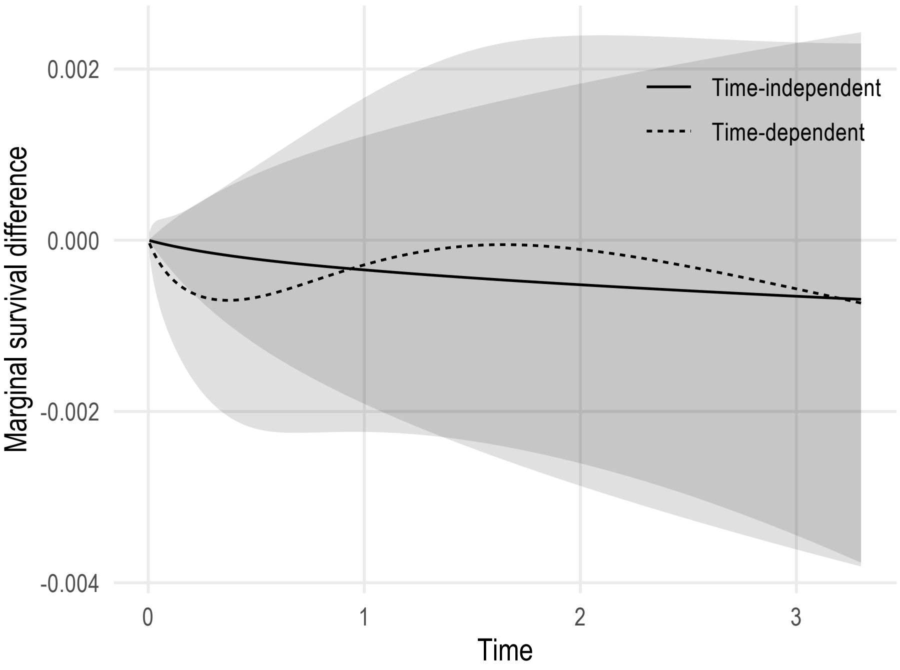


FIGURE 5.25: Application of shared frailty models to PSP-CKD data, comparison of time-dependent and time-independent marginal survival difference for treatment

mainstream methods that can be utilised with any general-purpose statistical software such as R and Stata. This is especially relevant in the settings of electronic health records, where it is common to encounter hierarchical survival data such as individuals clustered within e.g. primary care practice.

Compared to a marginal approach (i.e. accounting for clustering by using a robust estimator of the variance-covariance matrix of the estimated coefficients), the frailty approach allows focussing on inference within the clusters and quantifying the amount of heterogeneity between clusters by directly modelling it. Additionally, the frailty approach can be used to model recurrent events data, assuming that the recurrent event times are independent conditional on the covariates and random effects [146, 147]. Consequently, the adoption and use of such models have been steadily increasing in all fields of application: for instance, psychiatry [148], orthodontia [149], diabetes [150, 151], healthcare research [152], and even animal ecology [153].

Glidden and Vittinghoff [119] showed the benefit of using frailty models instead of models with fixed effects only or stratified approaches in the setting of multi-centre clinical trials, which may have driven adoption and use. Despite the increasing use of such methods, however, there has been little research on the impact of violating modelling assumptions - especially regarding the shape of the baseline hazard. Much research has focussed on misspecification of the frailty distribution, and the consensus is that relative risk estimates are largely unaffected by it [118–121, 123]. However, little was known about e.g. the impact of misspecifying the frailty on measures of absolute risk, or the impact of misspecifying the baseline hazard on estimated measures of heterogeneity. With this simulation study, I aimed to shed further light on the topic and ultimately provide additional guidance to applied researchers.

I simulated clustered survival data under a variety of clinically plausible scenarios, assuming different shapes for the baseline hazard function and different distributions for the shared frailty. I varied the amount of heterogeneity (in terms of variance of the frailty), and I also varied sample size - both in terms of number of clusters and number of individuals per cluster. I then fitted a large variety of survival models with shared frailty terms: assuming standard parametric distributions for the baseline hazard, flexibly modelling the baseline hazard via restricted cubic splines, and also leaving the baseline hazard unspecified. Each model was fit assuming both a Gamma and a log-Normal distribution for the frailty, arguably the most common choices in literature: the Gamma frailty has convenient mathematical features and it is analytically tractable, while the log-Normal frailty has a direct interpretation as a random intercept in a multilevel mixed-effects survival model. To the best of my knowledge, this is the most extensive simulation study on the impact of misspecifying the baseline hazard, the frailty distribution, or both in shared frailty survival models: Rutherford *et al.* studied the robustness of flexible parametric models without considering frailty terms, while Pickles and Crouchley, Glidden and Vittinghoff, and Lee and Thompson only studied misspecification of the random effects distribution [28, 118–120]. Liu *et al.* focussed on generalised survival model [121], and Ha *et al.* studied both misspecification of the baseline hazard and the frailty distribution but included fewer models in their comparison and simulated a small amount of scenarios [123].

The results of this extensive simulation study confirm the robustness of regression coefficients to misspecification of the frailty distribution, irrespectively of sample size and amount of heterogeneity in the data. However, the results also show the importance of properly modelling the baseline hazard. For instance, as shown in Section 5.5.7, the bias induced by assuming a standard parametric distribution with a true complex baseline hazard can be clinically relevant. In practical terms, this means that by failing to model the baseline hazard the effect of interest could be largely over- or under-estimated. I showed that absolute measures of risk such as the loss in expectation of life are affected by misspecification of both the baseline hazard and the frailty distribution: assuming a baseline hazard that is too simple or misspecifying the frailty distribution yields biased estimates and larger mean squared errors compared to well-specified models. Further to that, estimation of the frailty variance is also affected by poorly modelling the baseline hazard, with misspecified parametric models yielding biased estimates of heterogeneity. This highlights once again the necessity of using models that are flexible enough and the importance of assessing model fit regarding the distribution of the frailty by using information criteria (e.g. the AIC, BIC, cAIC) if no previous biological knowledge is available. The loss in life expectancy is a measure that is rarely implemented in statistical software, therefore I produced code in R to estimate such quantity using each model included in the simulation study. This code is openly available on-line for everyone to use at `https://github.com/ellessenne/frailtymcsim`.

The performance of semiparametric Cox models and flexible parametric models in the settings of this simulation study is comparable, as they produce largely unbiased relative risk estimates. However, the necessity of estimating the baseline hazard (e.g. by using the Breslow estimator) heavily affects the usage of semiparametric models when absolute risk measures are of interest. The Cox model is, de facto, the standard model fitted by applied researchers when dealing with time to event data; despite that, Sir David Cox himself argued in favour of parametric models [154], especially when interested in predicting the outcome for a given individual. Parametric models are indeed known to have desirable features in terms of prediction, extrapolation, quantification of absolute risk measures. Flexible parametric models represent an attractive alternative to semiparametric and fully parametric survival models: they retain both the robustness to misspecification of the baseline hazard and the appealing advantages of parametric models for prediction, extrapolation, quantification. Since their introduction by Royston and Parmar [24] in 2002, flexible parametric models have entered the statistical mainstream and have been extended to accommodate (among other) relative survival [155], random effects [108, 121], and generalised link functions [137]. The advantage of using flexible parametric models compared to semiparametric models by modelling the baseline hazard is particularly noteworthy: this allows translating relative risk measures on the absolute scale in a straightforward way, aiding interpretation.

In the applied example of Section 5.6, all models but those fitted using the `frailtypack` package yield comparable estimates in term of treatment effect. Conversely, when estimating the 3-years LLE models with models assuming an exponential or Gompertz baseline hazard (or with the aforementioned models from the `frailtypack` package) the difference was more noticeable, despite still not being clinically relevant. The models assuming a Weibull baseline hazard performed better than models with an exponential or Gompertz baseline hazard, which could be explained by the smooth, monotonic underlying hazard (Figure 5.23) that could fit well a Weibull baseline hazard distribution. Putting the results of the best-fitting model in perspective, they are consistent with the results of the PSP-CKD pragmatic trial [48]: PSP-CKD investigators concluded that after 42 months of follow-up the estimated renal function did not differ significantly between control and intervention groups.

The wide variety of simulated data-generating mechanisms (90) is one of the advantages of this simulation study. I also included the most common frailty distributions (Gamma and log-Normal), and I simulated survival data under many different and clinically plausible baseline hazards. This is particularly important, as if I only simulated data from a Weibull model, I would have been assuming a baseline hazard that increases or decreased monotonically. While such an assumption could be reasonable in some settings, sometimes fully parametric distributions are just not flexible enough to capture complex baseline hazards with turning points that are often observed in clinical datasets [28, 64].

This simulation study has also some limitations. First, I only simulated clusters of equal size and I did not include all the frailty distributions that have been proposed in the literature, e.g. positive stable or inverse Gaussian. Second, I only simulated right-censored survival data; settings with delayed entry or interval censoring require further investigation. Third, all methods use maximum likelihood which returns negatively biased estimates of the variance components; such bias decreases as the number of clusters increases and can be observed (for instance) with the results of the scenarios with 20 clusters of 150 individuals each (Figure E.6). The restricted maximum likelihood method could be used with a small number of clusters to obtain unbiased estimates of the variance components [156]. Fourth, I designed and analysed this simulation study using a fully factorial design; even though I simulated a large number of scenarios, incomplete designs and meta-modelling could be implemented to further increase the external validity and the ability to generalise the results, as described in more detail in Chapter 8. Finally, I heavily rely on the performances and R implementation of the models included in this comparison. Hirsch and Wienke [133] compared several implementations of the semiparametric Cox model with frailty terms and found `coxme` (the R package I chose to fit semiparametric log-Normal frailty models) to be among the most robust. Regardless, all the packages I chose are well established and utilised in practice, and I mimicked applied research by applying these methods as they are intended to be used, i.e. without modifying convergence criteria and/or starting values of the estimation procedure.

The work presented in this Chapter has been published in Statistics in Medicine [93], and can also be found in Appendix D.

6 *Joint Modelling of Longitudinal and Time to Event Data*

6.1 OUTLINE

In this Chapter, I introduce the topic of joint modelling of longitudinal and survival data. I will introduce the rationale behind joint modelling in Section 6.2, and the formulation of a standard joint model for longitudinal and survival data (*joint model* in brief) in Section 6.3. Then, I will describe association structures between the longitudinal and the survival components of the joint model that have been introduced in the literature in Section 6.4, and the estimation process in Section 6.5. I will continue by describing in Section 6.6 the issue of informative drop-out in longitudinal studies, introducing methods that have been proposed in the literature to account for it in the analysis, with an illustrative applied example using data from VASST in Chapter 6.7. Finally, I will conclude the Chapter with a discussion in Section 6.8.

6.2 INTRODUCTION TO JOINT MODELLING OF LONGITUDINAL AND SURVIVAL DATA

Routinely collected EHRs are being used more and more for research purposes, as outlined in Chapter 1. One of the defining characteristics of EHRs is the presence of repeated measures recorded over time as individuals access health care: when they attend a visit and have e.g. a blood test, new measurements are recorded and added to the system. This is increasingly common in observational studies and clinical trials as well, as participants are followed over time and abundant data on clinical features is recorded throughout the study. One of the main challenges when analysing longitudinal data is the clustering structure, e.g. the correlation between observations from a given individual needs to be taken into account; mixed-effects models achieve so by including latent terms (the random effects), as described in Section 2.3.

Alongside longitudinal data, time to the occurrence of an event is often the outcome of interest - either when analysing EHRs or even data from clinical studies and trials. As a consequence, researchers often encounter longitudinally recorded covariates that they need to account for when studying the clinical outcome of interest. Researchers then face two options: (1) select only one of the multiple values per individual and analyse as such (e.g. values recorded at baseline), ignoring much of the available data, or (2) take into account the potential dependency and association between the repeatedly measured covariates and the clinical outcome. The latter is usually the most sensible choice, as the longitudinal data can contain important predictors or surrogates of the time to event outcome. A powerful tool to achieve so is given by joint models for longitudinal and survival data, in which the longitudinal and survival processes are modelled jointly into a single model allowing to infer their association. Previous attempts to tackle this problem consisted in (1) fitting a time-dependent Cox model [22] by splitting individual rows every time a new observation from the longitudinal covariate becomes available, and (2) by using two-stage methods in which the longitudinal and survival data are modelled separately [157]. These two methods could be easily applied using standard statistical software; nevertheless, it has been shown that joint modelling provides several advantages in terms of increased efficiency, bias reduction, and improved predictions at the same time [158, 159].

Conversely, this problem could also be observed from a different angle. Longitudinal studies are often affected by drop-out, and general methods for the analysis of such data (as those described in Section 2.3) assume that the longitudinal outcome and the drop-out process are independent. This assumption is often unreasonable: it is not hard to imagine settings where drop-out is associated with the underlying profile of a biomarker, e.g. when abnormal values are associated with an increased risk of mortality. It can be shown that ignoring the drop-out process can affect the results of the longitudinal analysis [160, 161]; more details are included in Section 6.6.

Seminal work on joint modelling of longitudinal and survival data was motivated by clinical trials on zidovudine for the treatment of human immunodeficiency virus (HIV) [157, 162–164]. In those settings, CD4 lymphocyte counts were recorded throughout the study: CD4 counts are known to be associated with clinical outcomes, and the aim of the analysis consisted of understanding CD4 trajectories and the degree of the association with survival from a prognostic point of view. Another seminal paper by Henderson *et al.* [165] was motivated by a clinical trial on drug therapy for schizophrenia patients [166]. Specifically, the outcome of interest was mean scores for each of three treatment groups on a particular measure of psychiatric disorder; not all patients completed the trial, and in fact, the analysis of survival curves showed that a substantial proportion of each treatment group withdrew before completing the measurement schedule. It was therefore not clear whether the apparent decrease in scores profiles reflected a genuine change over time, or was an artefact caused by differential drop-out.

More recent discussions on the topic are presented in Ibrahim *et al.* [167], Rizopoulos [168], and Gould *et al.* [169]. Applications of joint models for longitudinal and survival data to answer complex study questions using complex clinical data are also increasingly common in medical literature, in a variety of settings: among others, cardiology [170], nephrology [171], and intensive care medicine [172].

Several extensions of the standard joint model with a single longitudinal outcome and a single survival outcome have recently appeared in the literature. The standard joint model formulations required a proportional hazards model for the time to event component [163, 165]; Tseng *et al.* developed an alternative joint model where the time to event sub-model is an accelerated failure time (AFT) model, as described in Section 2.2.4 [173]. The standard joint model also left the baseline hazard of the survival sub-model unspecified; as described in Section 2.2.4, this has advantages and disadvantages. Further to that, Hsieh *et al.* [174] showed that in the settings of joint modelling, leaving the baseline hazard unspecified results in under-estimation of the parameter standard errors, requiring bootstrapping to obtain appropriate standard errors. Crowther *et al.* showed that it is possible to port flexible parametric models to the joint modelling framework, providing the advantages of fully modelling the baseline hazard and removing the need for bootstrapping [175]. Proust-Lima *et al.* showed that it is also possible to formalise a joint longitudinal-survival model using the joint latent class approach, which consists in assuming that a latent class structure entirely captures the correlation between the longitudinal marker trajectory and the risk of the event [176]. Finally, joint models have been extended to accommodate competing risks and multiple longitudinal trajectories, with more details provided elsewhere [177–179].

6.3 MODEL FORMULATION

Throughout this Chapter I will focus on the standard formulation of the joint model. As mentioned in Section 6.2, this formulation consists of a joint model with two components: a sub-model for the longitudinal trajectory, and a sub-model for the time to event outcome. These two components will then share one or more parameters: the most common formulation in the literature assumes the components share a latent structure that will describe the association between the two processes, therefore linking them.

Building on the notation from Chapter 2 and from Rizopoulos [168], let $t_i = \min(t_i^*, c_i)$ be the observed survival time with t_i^* the true survival time and c_i the censoring time. Let d_i be an event indicator variable, which takes the value 1 if $t_i^* < c_i$ and 0 otherwise. Let $y_{ij} = \{y_{ij}(t_{ij}) \; \forall \, j = 1, \dots, n_i\}$ be the observed longitudinal response for the i^{th} subject, with $y_{ij}(t_{ij})$ the observed response at time t_{ij} and n_i the number of longitudinal observations. Let U_i be a vector of time-independent baseline covariates.

The longitudinal component of the joint model is modelled within the mixed-effects framework [31], as longitudinal data is generally measured intermittently and with error. Therefore:

$$y_i(t_i) = m_i(t_i) + \epsilon_i(t_i), \;\; \epsilon_i(t_i) \sim N(0, \sigma_\epsilon^2) \tag{6.1}$$

with

$$m_i(t_i) = X_i(t_i)\beta + Z_i(t_i)b_i, \;\; b_i \sim N(0, \Sigma) \tag{6.2}$$

with $X_i(t_i)$ and $Z_i(t_i)$ the (possibly) time-dependent design matrices for the fixed and random effects β and b_i, respectively. $y_i(t_i)$ represents the observed longitudinal trajectory at time t, which could be decomposed into the true longitudinal trajectory $m_i(t_i)$ plus the measurement error $\epsilon_i(t_i)$. I also assume that the measurement error $\epsilon_i(t_i)$ is normally distributed with variance σ_ϵ^2, independent of the random effects, and that $\text{Cov}(\epsilon_i(t_i), \epsilon_i(u_i)) = 0 \; \forall t_i \neq u_i$. Flexibility in the longitudinal submodel can be incorporated by modelling the effect of time e.g. using fractional polynomials, B-splines, or restricted cubic splines [180–182].

The survival component of the joint model is modelled using a proportional hazards time to event model, given the true unobserved longitudinal trajectory up to time t_i, i.e. $M_i(t_i) = \{m_i(s_i) \; \forall \; 0 \leq s_i \leq t_i\}$:

$$h(t_i | M_i(t_i), W_i) = h_0(t_i) \exp(W_i \psi + \alpha m_i(t_i)), \tag{6.3}$$

where $h_0(t_i)$ is the baseline hazard function and $W_i \in U_i$ is a vector of time-fixed covariates with regression parameters ψ. α is the association parameter that links the true unobserved trajectory function $m_i(t_i)$ and the survival submodel. The association parameter α can be interpreted as the log-hazard ratio for a unit increase in the true longitudinal trajectory $m_i(t_i)$, at time t_i; in this setting, the association is based on the current value of the longitudinal response at time t_i. Additional association structures are available and further described in Section 6.4.

The survival function follows as

$$S(t_i | M_i(t_i), W_i) = \exp \left(- \int_0^{t_i} h_0(u) \exp(W_i \psi + \alpha m_i(u)) \, du \right), \tag{6.4}$$

and it is clear that the survival function (according to this definition) depends on the entire history of the longitudinal trajectory up to time t_i.

Finally, the choice of the baseline hazard $h_0(t_i)$ follows the usual rationale (as mentioned before). Traditionally, in the joint modelling settings the baseline hazard function has been left unspecified [163, 165]; however, as mentioned in Section 6.2, it has been shown that leaving the baseline hazard unspecified yields standard errors for the regression parameters that are underestimated [174]. Bootstrapping is therefore required to obtain appropriate standard errors, with additional computational complexity. Alternatively, it is possible to assume a parametric distribution for the baseline hazard, or even use flexible parametric formulations as described in Crowther *et al.* [175].

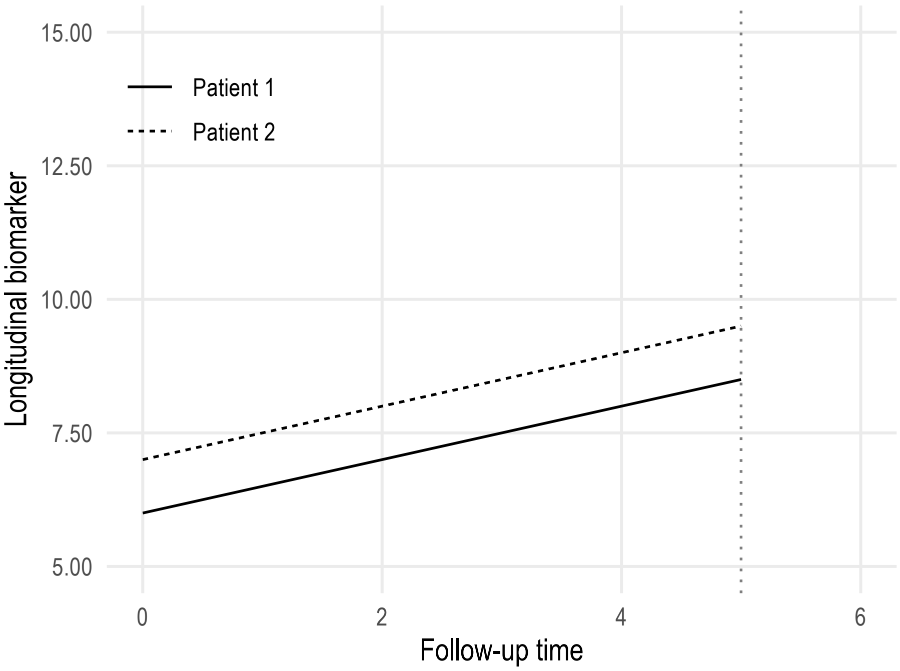


FIGURE 6.1: Longitudinal profile of a biomarker for two distinct individuals with different current value and the same rate of change

6.4 ASSOCIATION STRUCTURES

Several alternative, clinically meaningful association structures are available in the joint modelling framework and described throughout this Section.

The association structure that links the longitudinal and time to event components of the joint model described in the previous Section relates the true unobserved value of the longitudinal trajectory at time t_i directly to the risk of event at time t_i: this is often referred to as the *current value* association structure. In practice, this association structure states that only the current level of a biomarker is predictive of future outcomes; individuals with a different current value of a biomarker e.g. after 5 years of follow-up - as in Figure 6.1 - will have a different predicted survival.

First, it is possible to allow for different values of the association parameter for different sub-groups of patients by including interaction terms with the true unobserved longitudinal trajectory as follows:

$$h(t_i|M_i(t_i), W_{i1}, W_{i2}) = h_0(t_i) \exp(W_{i1} \psi + (W_{i2} m_i(t_i)) \alpha), \tag{6.5}$$

with $W_{i1}, W_{i2} \in U_i$. This yields a vector of association parameters α, providing different association parameters for different covariate patterns; for instance, the association between e.g. the current value of the biomarker and survival could differ between e.g. treated and non-treated individuals.

Second, the association structures described so far link the current value of the longitudinal trajectory to the survival submodel. It is possible to define an association structure that links the rate of change (or slope) of the longitudinal trajectory:

$$h(t_i|M_i(t_i), W_i) = h_0(t_i)\exp(W_i\psi + \alpha m_i'(t_i)), \tag{6.6}$$

with

$$m_i'(t_i) = \frac{\partial m_i(t_i)}{\partial t_i} = \frac{\partial(X_i(t_i)\beta + Z_i(t_i)b_i)}{\partial t_i}$$

This association structure is often referred to as *slope* association structure, and states that the rate at which the biomarker change is predictive of future outcomes. For instance, individuals whose longitudinal trajectory rises sharply may have worse predicted survival than individuals with a stable biomarker (Figure 6.2). The slope association structure can also be combined with the current value association structure to yield a *current value and slope* association structure:

$$h(t_i|M_i(t_i), W_i) = h_0(t_i)\exp(W_i\psi + \alpha_1 m_i(t_i) + \alpha_2 m_i'(t_i)) \tag{6.7}$$

In this setting, individuals with different current value of a biomarker and a different rate of change will have a different predicted survival (Figure 6.3).

Next, the *cumulative effect* association structure links the risk of event with the cumulative effect of a longitudinal trajectory, calculated as the area under the curve:

$$h(t_i|M_i(t_i), W_i) = h_0(t_i)\exp(W_i\psi + \alpha\int_0^{t_i} m_i(u)\ du) \tag{6.8}$$

In practice, the cumulative exposure to the biomarker is assumed to be predictive of future outcomes; for instance, the cumulative effect of inflammatory response biomarkers (e.g. C-reactive protein) could be a risk factor for cardiovascular disease.

Interestingly, the association structures described so far link the current (at time t_i) value (or slope, cumulative effect) of the longitudinal trajectory to the risk of event at time t_i:

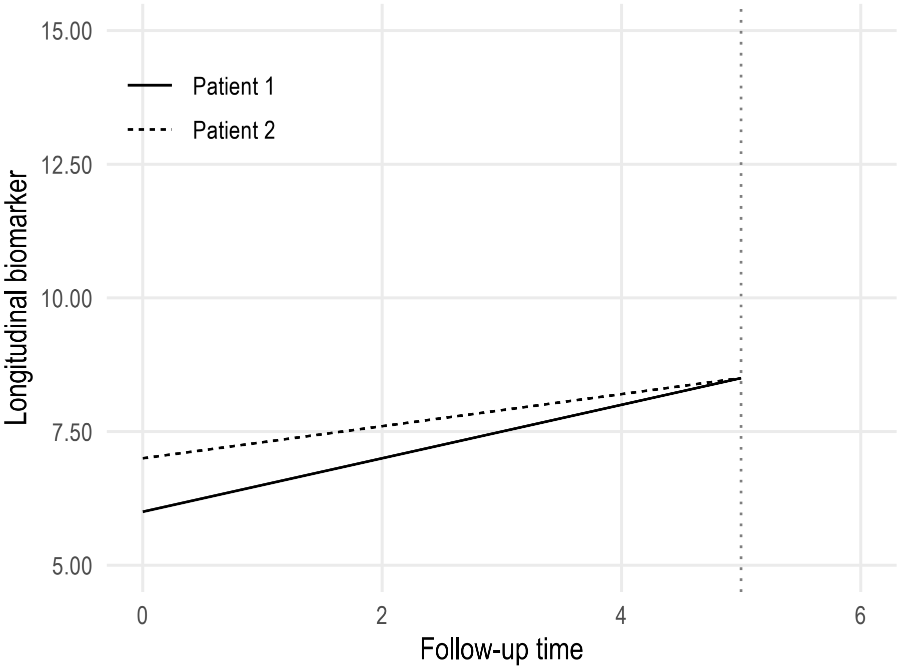


Figure 6.2: Longitudinal profile of a biomarker for two distinct individuals with the same current value and a different rate of change

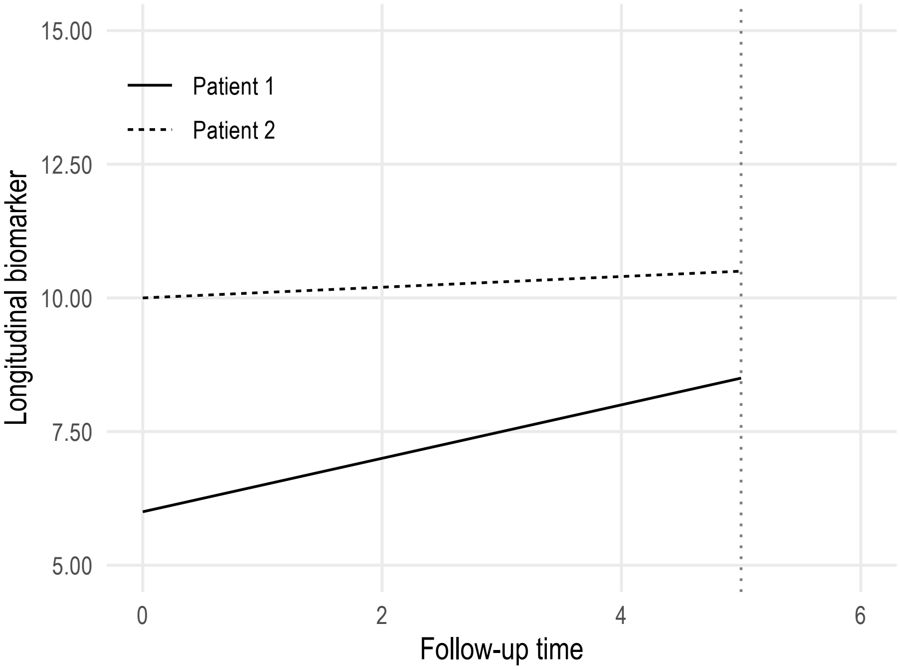


Figure 6.3: Longitudinal profile of a biomarker for two distinct individuals with different current value and rate of change

this can be further generalised by allowing lagged effects, e.g. linking the current value (or slope, cumulative effect, etc.) at time $u_i < t_i$ of the longitudinal trajectory to the risk of event at time t_i.

Finally, the *random effects* association structure is a time-independent association structure that includes only the random effects of the longitudinal trajectory in the linear predictor of the survival sub-model:

$$h(t_i|M_i(t_i), W_i) = h_0(t_i) \exp(W_i \psi + (\beta + b_i)\alpha) \tag{6.9}$$

Equation (6.9) includes both the population-level mean of the random effect (β) and the subject-specific deviation b_i. Alternatively, it is possible to include only the subject-specific deviation:

$$h(t_i|M_i(t_i), W_i) = h_0(t_i) \exp(W_i \psi + b_i \alpha) \tag{6.10}$$

Interpretation of the association parameter will differ depending on whether the population-level mean is included or not. For instance, if the population-level mean is included then the association parameter represents the change in risk for a unit increase in the overall trajectory; conversely, the association parameter represents the change in risk for a unit increase in the deviation from the population mean.

6.5 Model Estimation

Estimation of a joint model for longitudinal and survival data is a non-trivial task. The complexity of jointly modelling the longitudinal component and the survival component motivated the use of two-stages procedures as mentioned in Section 6.2: with that approach, the longitudinal component is modelled and estimated separately; consequently, subject-specific predictions from the longitudinal model are produced and plugged into the survival model as time-varying covariates. Despite the simplicity of this approach, it has been shown that it produces substantial bias and poor coverage [170, 183]; therefore, an approach that models both processes jointly is required. In particular, two approaches are predominant: a full likelihood approach, and a Bayesian approach; both have appealing characteristics, but they share the feature of being computationally intensive.

Focusing on the full likelihood approach, it is possible to formulate the joint likelihood for the overall parameter vector $\theta = \{\theta_t, \theta_y, \theta_b\}$, formed by the parameters of the survival component, the parameters of the longitudinal component, and the elements of the variance-covariance matrix of the random effects, respectively [168]. The joint distribution of the observed survival time t_i, the event indicator d_i, and the longitudinal response y_i, conditional on the random effects b_i, can be expressed as:

$$f(t_i, d_i, y_i | b_i, \theta) = f(t_i, d_i | b_i, \theta) f(y_i | b_i, \theta), \tag{6.11}$$

with

$$f(y_i | b_i, \theta) = \prod_{j=1}^{n_i} f(y_i(t_{ij}) | b_i, \theta). \tag{6.12}$$

It is important to note that the survival process and the longitudinal process are assumed to be independent, conditionally on the random effects b_i. It follows that the contribution to the log-likelihood for the i^{th} patient is

$$\begin{aligned} \log L_i(\theta) &= \log \int_{-\infty}^{+\infty} f(t_i, d_i, y_i, b_i; \theta)\, db_i \\ &= \log \int_{-\infty}^{+\infty} f(t_i, d_i | b_i, \theta_t) \left[\prod_{j=1}^{n_i} f(y_i(t_{ij}) | b_i, \theta_y) \right] f(b_i | \theta_b)\, db_i \end{aligned} \tag{6.13}$$

with $f(t_i, d_i | b_i, \theta_t)$ the contribution to the likelihood relative to the survival component of the model (assuming the current value association structure):

$$\begin{aligned} f(t_i, d_i | b_i, \theta_t) &= h(t_i | M_i(t_i), W_i, \theta_t)^{d_i} S(t_i | M_i(t_i), W_i, \theta_t) \\ &= [h_0(t_i) \exp(W_i \psi + \alpha m_i(t_i))]^{d_i} \exp\left[- \int_0^{t_i} h_0(u) \exp(W_i \psi + \alpha m_i(u))\, du \right], \end{aligned} \tag{6.14}$$

$f(y_i(t_{ij}) | b_i, \theta_y)$ the contribution to the likelihood of the longitudinal process:

$$f(y_i(t_{ij}) | b_i, \theta_y) = (2\pi\sigma_\epsilon^2)^{-1/2} \exp\left[- \frac{(y_i(t_{ij}) - m_i(t_{ij}))^2}{2\sigma_\epsilon^2} \right], \tag{6.15}$$

and $f(b_i|\theta_b)$ the density of the random effects:

$$f(b_i|\theta_b) = (2\pi)^{-q_b/2}|\Sigma|^{-1/2} \exp\left[-\frac{b_i^T\Sigma^{-1}b_i}{2}\right], \tag{6.16}$$

with q_b being the dimension of the random effects.

Historically, the predominant method for maximising the full joint likelihood has been the Expectation-Maximisation algorithm, where the random effects are treated as missing values [40]. Alternatively, within the maximum likelihood framework, it is possible to directly maximise the full joint likelihood using any general-purpose optimiser and standard maximisation algorithms such as the Newton-Raphson algorithm. This approach is computationally intensive, as the integral in Equation (6.13) does not have a closed-form and therefore requires a numerical approximation to be computed. The computational effort for joint models with a single random effect (e.g. a random intercept only) is similar to the computational effort required to fit shared frailty models, as described in Section 5.4.2. Methods such as standard Gaussian quadrature and adaptive Gaussian quadrature are routinely used in the joint modelling settings, with the latter method vastly preferred. However, as the number of random effects included in the model increases the computational burden required to fit a joint model grows exponentially: for instance, the multi-dimensional integral in Equation (6.13) requires k^q function evaluations, where k is the number of quadrature nodes and q is the number of random effects. Finally, under a parametric survival submodel, the integral in Equation (6.14) requires numerical integration to be evaluated as well when using a time-dependent association structure (e.g. the current value association structure); Gauss-Legendre quadrature can be used for that purpose [184]. Interestingly, by choosing a time-independent association structure (e.g. the random effects association structure) the requirement for numerical integration can be avoided as the cumulative hazard function has a closed-form, providing direct computational benefits.

Overall, it is clear that given the requirement of numerical integration to calculate the survival function which is nested within (possibly multi-dimensional) numerical integration to integrate over the random effects, estimation of a joint model is a computationally demanding and challenging task.

6.6 MODELLING THE DROP-OUT PROCESS

Most of the current work on joint models for longitudinal and time to event data has traditionally focussed on the survival outcome, adjusting for a time-varying covariate measured with error (the longitudinal outcome of the joint model). Despite that, as mentioned in Section 6.2, the interest of the analysis often lays in the longitudinal outcome: for instance, one may be interested in studying the evolution over time of a given biomarker and how it may be affected by a given treatment. Methods for the analysis of longitudinal data over time (as described in Section 2.3) rely on the assumption that factors that affect drop-out from the study (and truncation of the longitudinal trajectory) are not related to the study outcome. This assumption is often unreasonable, as drop-out may be affected by e.g. adverse reactions to treatment, lack of effectiveness, or concurrent health status. Hence, drop-out is often informative or non-ignorable [185, 186], being a potential source of bias in the analysis of longitudinal data.

A recent review in the settings of clinical trials concluded that 36% of studies did not account for potentially informative drop-out and carried out just a complete-case analysis [187]: the authors speculate that the under-utilisation of methods that account for informative drop-out could be due to lack of awareness or lack of research demonstrating the methods in practice.

I will start by defining the characteristics of the drop-out process. Following the terminology of Diggle and Kenward [160], the drop-out mechanism can be classified as:

1. Drop-out *completely at random*, when the drop-out process and longitudinal process are independent;
2. Drop-out *at random*, when the drop-out process depends on the observed longitudinal process;
3. *Informative* drop-out (or drop-out *not at random*), when the drop-out depends on unobserved characteristics of the longitudinal process.

This definition of the drop-out process is analogous to missingness mechanisms described in Rubin [185], and can be formalised as in Rizopoulos [168].

First, a *missing data indicator* can be defined as

$$r_{ij} = \begin{cases} 1 & \text{if } y_{ij} \text{ is observed} \\ 0 & \text{otherwise} \end{cases} \tag{6.17}$$

Then, the *complete* response vector for the i^{th} subject y_i can be partitioned into its *observed* part y_i^o (where $r_{ij} = 1$) and *unobserved* part y_i^m. When missingness is restricted to drop-out, the missing data indicator has always the form $(1, \dots, 1, 0, \dots, 0)$ and can hence be replaced by a scalar variable r_i^d:

$$r_i^d = 1 + \sum_{j=1}^{n_i} r_{ij} \tag{6.18}$$

r_i^d represents the occasion at which drop-out occurred in incomplete trajectories, $n_i + 1$ otherwise.

Consequently, the drop-out mechanisms previously described can be formalised by defining the probability model with parameters vector θ_r that describes the relationship between the drop-out process r_i^d and the complete response vector y_i:

1. Drop-out completely at random assumes that the probability of drop-out is unrelated to both observed and unobserved data:

$$P(r_i^d | y_i^o, y_i^m; \theta_r) = P(r_i^d; \theta_r) \tag{6.19}$$

2. Drop-out at random assumes that the probability of drop-out is related to observed data only:

$$P(r_i^d | y_i^o, y_i^m; \theta_r) = P(r_i^d | y_i^o; \theta_r) \tag{6.20}$$

3. Finally, informative drop-out assumes that the probability of drop-out depends on unobserved data, even if conditioning on the observed data:

$$P(r_i^d | y_i^m; \theta_r) \text{ or } P(r_i^d | y_i^o, y_i^m; \theta_r) \tag{6.21}$$

Kolamunnage-Dona *et al.* [188] illustrate how joint models for longitudinal and survival can be utilised to account for the drop-out process in the analysis of longitudinal data, assuming the drop-out process is *at least* at random. In particular, the drop-out process can be modelled via the time to event component of the joint model, and competing causes of drop-out can be accommodated as well (if required). Ibrahim *et al.* [167] show that in their applied example on quality of life in cancer patients the model coefficients for the longitudinal trajectory resulting from two-stage methods differ from the model coefficients obtained from the joint model that includes the time-to-event component. Li and Su [189] showed similar results in their analysis of longitudinal CD4 cell count profiles from the HIV Epidemiology Research Study, with the linear mixed model that disregards the drop-out process underestimating the CD4 cell time slope as patients who stayed in the study tended to have a less rapid decline of CD4 cell count.

Other methods have been suggested to account for drop-out in longitudinal data analysis. The method of generalised estimating equations (GEE) assumes that drop-out is completely at random, but Robins *et al.* showed that it is possible to restore the unbiasedness of the GEE method by properly weighting the analysis by the probability of dropping out of the study [190]. The estimator they propose is commonly referred to as *inverse probability of censoring weighted* estimator, and it can accommodate drop-out at random but not drop-out not at random. Diggle and Kenward [160] suggested an approach based on modelling the longitudinal outcome and the drop-out probability that yields a likelihood-based inference on the coefficients of the longitudinal model that can ignore the drop-out process; this approach is also referred to as *selection models*. Next, *pattern mixture models* factorise the joint distribution of the longitudinal and drop-out process into the marginal distribution of the drop-out process and the conditional distribution of the longitudinal outcome, given the drop-out process [191]. However, pattern mixture models yield identifiability issues in the settings of informative drop-out, as described in Molenberghs *et al.* [192]. Finally, *random effects models* arise from the intuitive idea that drop-out is likely related to subjects' characteristics, including some which are unobserved. A subject's propensity to drop-out can then be modelled by including a random effect which accounts for the above-mentioned unobserved characteristics [193]. The joint distribution of the longitudinal outcome and the drop-out process can be formulated, and it can be shown that the two components are conditionally independent given the random effects. All of the methods just mentioned are described in more detail in Chapter 13 of Diggle *et al.* [31].

Joint models for longitudinal and survival data lay in the class of random effects models described above; I will focus on the use of joint models for longitudinal and survival data to account for drop-out, and in particular I will illustrate the use of joint modelling with a posthoc analysis of data from VASST in the next Section.

6.7 Application to VASST Data

The application in this Section is an unplanned posthoc analysis of VASST, a multi-centre, double-blind randomised controlled trial that assigned patients who had septic shock and were receiving a minimum of 5 μg/min of norepinephrine to receive either low-dose vasopressin (0.01 to 0.03 U/min.) or norepinephrine (5 to 15 μg/min) plus open-label vasopressors. The primary end-point was mortality rate 28 days after treatment initiation, and VASST investigators found that there was no significant difference between treatments arms (35.4% and 39.3%, respectively; p-value: 0.26) [55].

Nevertheless, SOFA score is considered to be a relevant clinical outcome given that treatment-associated changes in SOFA score from baseline have been shown to be reliably and consistently associated with observed mortality [57]. A challenge in using non-mortal outcomes, however, is that bias may occur when such end-points are informatively missing (or censored) due to mortality: the observed slope of change in the outcome may be associated with censoring, which is generally worse for those who die. If the censoring process is associated with the intervention, then the truncated observation of the outcome will be informative, as described in Section 6.6.

Therefore, the aim of this analysis is two-fold. First, I aim to compare the effect of vasopressin versus norepinephrine on the SOFA score over 28 days of follow-up using a joint longitudinal-survival model to account for potentially informative drop-out. Second, I aim to compare the results obtained from the joint model with results obtained from a standard linear mixed model that disregards the drop-out process completely.

I extracted from VASST all individuals with at least one SOFA score measurement, which yielded 763 study subjects: 389 in the vasopressin arm and 374 in the norepinephrine arm. The SOFA score was calculated daily using the six organ subscales using data recorded in the VASST case report form.

Possible causes of drop-out from the study were death within 28 days, or discharge

Table 6.1: Summary of drop-out by treatment, application to VASST data

Status	Vasopressin	Norepinephrine	Overall
N. of study subjects	389 (50.98%)	374 (49.02%)	763
Completed follow-up	151 (38.82%)	140 (37.43%)	291 (38.14%)
Overall drop-out	238 (61.18%)	234 (62.57%)	472 (61.86%)
Case 1: Drop-out due to death	137 (35.22%)	147 (39.30%)	284 (37.22%)
Case 2: Drop-out due to discharge	101 (25.96%)	87 (23.26%)	188 (24.64%)

from the intensive care unit (ICU); the drop-out process is summarised in Table 6.1. It is possible to see that more individuals completed follow-up in the vasopressin group, 38.82% versus 37.43% in the norepinephrine arm. The main reason for drop-out was death, with 37.22% of individuals dying during follow-up; as a comparison, 24.64% of individuals dropped out because of discharge from the ICU. Comparing the two treatment arms, the norepinephrine arm had a higher percentage of study subject dropping out because of death compared to the vasopressin arm (39.30% vs 35.22%), while the opposite was true for discharge (23.26% in the norepinephrine arm compared to 25.96% in the vasopressin arm).

Raw, unadjusted SOFA score trajectories by completion of follow-up (or not) are depicted in Figures 6.4 and 6.5. Figure 6.4 depicts trajectories for each possible cause of drop-out, while in Figure 6.5 all causes of drop-out are pooled together for simplicity. The trajectories are clearly different between individuals that completed follow-up and individuals that did not: for instance, individuals dropping out because of death showed a sharp rise in SOFA score right before their death, while individuals dropping out because of discharge showed a steady, linear decrease in SOFA score values. As a comparison, individuals that completed follow-up showed a sharp decrease in SOFA score early on, with a flat trajectory thereafter. Finally, differences were small when comparing treatment arms.

The sharp difference in raw trajectories between individuals completing follow-up and individuals dropping out, together with (arguably small) differences in drop-out proportions between treatment arms motivate the use of joint longitudinal-survival modelling to account for potentially informative drop-out. In particular, I include treatment effect in the time to event sub-model, and the effect of time plus a time-treatment interaction in the longitudinal sub-model; by not including the main

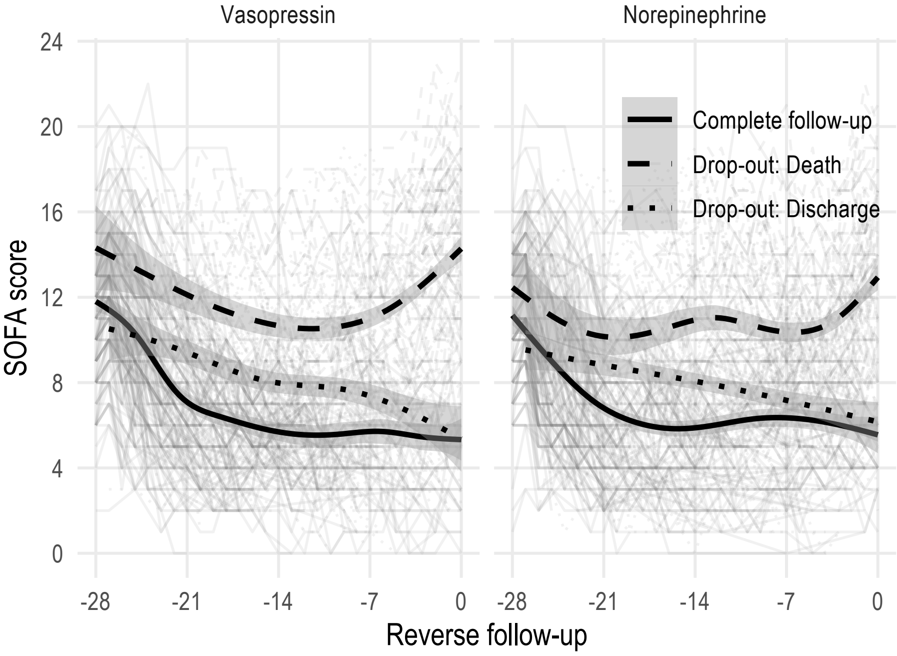


Figure 6.4: Subject-specific and smoothed trajectories by completion of follow-up or cause-specific drop-out, application to VASST data. Smoothed trajectories were obtained by fitting a generalised additive model with a penalised cubic spline smoother, as implemented in `mgcv::gam` with `bs = "cs"`.

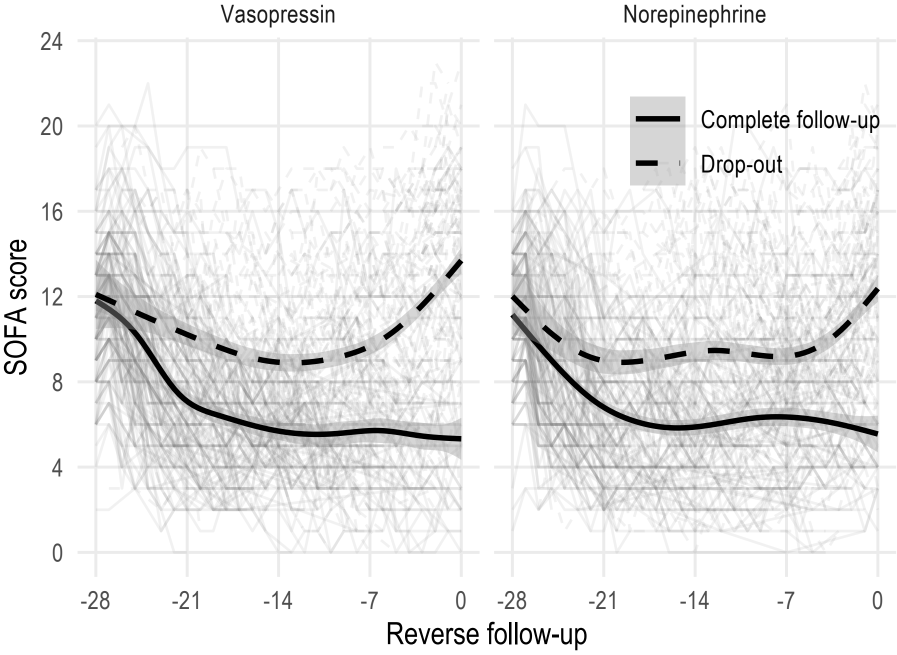


Figure 6.5: Subject-specific and smoothed trajectories by completion of follow-up or overall drop-out, application to VASST data. Smoothed trajectories were obtained by fitting a generalised additive model with a penalised cubic spline smoother, as implemented in `mgcv::gam` with `bs = "cs"`.

effect of treatment in the longitudinal sub-model, I am assuming that randomisation yielded a balanced distribution of SOFA scores at baseline. I model overall drop-out using the time to event submodel, pooling together all possible causes of drop-out for simplicity; individual causes of drop-out could be studied as well by fitting cause-specific survival sub-models [177, 188].

The joint model I use follows the formulation of Henderson *et al.* [165], as implemented in the `joineRML` R package [194, 195]. Under this formulation, the longitudinal outcome is assumed to follow the model

$$y_i(t_i) = X_i(t_i)\beta + \Lambda_{1i}(t_i) + \epsilon_i(t_i), \tag{6.22}$$

where $X_i(t_i)\beta$ is the mean response and $\epsilon_i(t_i)$ is an error term assumed to be independent and identically distributed normal with zero-mean and variance σ_ϵ^2. The covariates $X_i(t_i)$ are possibly time-varying, and have associated regression coefficients β. $\Lambda_{1i}(t_i)$ is a latent term describing the random effects included in the model:

$$\Lambda_{1i}(t_i) = Z_i(t_i)b_i, \tag{6.23}$$

with b_i assumed to follow a zero-mean multivariate normal distribution with variance-covariance matrix Σ. $Z_i(t_i)$ represents covariates with random effects b_i, and it is assumed that $Z_i(t_i) \subseteq X_i(t_i)$. Note that $X_i(t_i)\beta + \Lambda_{1i}(t_i) = m_i(t_i)$ from Equations (6.1) and (6.2). Correlation between random effects is allowed, and it is also assumed that the random effects b_i and the error term $\epsilon_i(t_i)$ are independent.

The time to event sub-model is given by the hazard function

$$h(t_i) = h_0(t_i) \exp(W_i(t_i)\gamma + \Lambda_{2i}(t_i)), \tag{6.24}$$

with $h_0(t_i)$ an unspecified baseline hazard function and $W_i(t_i)$ possibly time-varying covariates with associated regression coefficients γ. By defining $\Lambda_{2i}(t_i)$ as a linear combination of $\Lambda_{1i}(t_i)$ a latent association is established; in particular, the joint model assumes that

$$\Lambda_{2i}(t_i) = \alpha\Lambda_{1i}(t_i), \tag{6.25}$$

with α representing the association between the longitudinal and the survival sub-models. The joint model is fitted using the EM algorithm as described by Lin *et al.* for the settings of multivariate joint modelling [196]; the intractable integrals in the likelihood computation are approximated via Monte Carlo integration with antithetic simulation for variance reduction, as originally suggested by Henderson *et al.* [165].

To identify the model formulation that best fits the observed data, I fit several joint models with different model formulations in terms of fixed and random effect of time. In particular, I investigate:

1. A linear, quadratic, cubic, or splined (with 2 to 7 degrees of freedom) fixed effect of time;
2. A random intercept only, or a random intercept plus a linear, quadratic, or splined (2 degrees of freedom) random effect of time.

I fit every possible combination of fixed and random effects, for a total of 36 models; the fit of each model in terms of AIC and BIC is presented in Table 6.2. Zhang *et al.* [197] showed how it is possible to decompose the AIC and BIC of a joint model into additive components that allow assessing the fit of each component of the joint model (e.g. longitudinal and survival components, separately). Unfortunately, this decomposition is currently not implemented in `joineRML`.

The model that fits the data best is the model with a fixed effect of time modelled via a restricted cubic spline with 7 degrees of freedom, a random intercept, and a random effect of time modelled via a restricted cubic spline with 2 degrees of freedom. The resulting longitudinal sub-model is therefore:

$$y_i(t_i) = \beta_0 + \beta_{1-7}^T s(\text{Time}, 7) + \beta_{8-14}^T s(\text{Time}, 7) \times \text{Treat} + b_{0,i} + b_{1-2,i}^T s(\text{Time}, 2) + \epsilon_i(t), \quad (6.26)$$

where $s(\text{Time}, k)$ represents a restricted cubic spline expansion of time with k degrees of freedom, for conciseness. Analogously, β_{m-l} represents the m^{th} to l^{th} regression coefficients associated with the spline terms and $b_{m-l,i}$ the m^{th} to l^{th} random effect. The survival submodel is

$$h(t_i) = h_0(t_i) \exp\{\gamma \text{Treat} + \alpha[b_{0i} + b_{1-2,i}^T s(\text{Time}, 2)]\} \quad (6.27)$$

Table 6.2: Comparison of model fit (AIC/BIC), application to VASST data

Fixed effect	Random effect	AIC	BIC
Linear	Random intercept	38,058.72	38,091.18
Linear	Random intercept + slope	36,296.23	36,337.97
Linear	Random intercept + quadratic slope	35,243.55	35,299.20
Linear	Random intercept + splined slope (2df)	35,123.16	35,178.81
Squared	Random intercept	37,114.27	37,156.00
Squared	Random intercept + slope	35,581.99	35,633.00
Squared	Random intercept + quadratic slope	35,092.27	35,157.19
Squared	Random intercept + splined slope (2df)	35,079.04	35,143.96
Cubic	Random intercept	36,963.58	37,014.59
Cubic	Random intercept + slope	35,426.43	35,486.72
Cubic	Random intercept + quadratic slope	34,989.19	35,063.39
Cubic	Random intercept + splined slope (2df)	34,953.99	35,028.19
Spline (2df)	Random intercept	37,009.70	37,051.43
Spline (2df)	Random intercept + slope	35,464.20	35,515.21
Spline (2df)	Random intercept + quadratic slope	34,968.61	35,033.53
Spline (2df)	Random intercept + splined slope (2df)	34,963.26	35,028.18
Spline (3df)	Random intercept	36,965.11	37,016.12
Spline (3df)	Random intercept + slope	35,415.10	35,475.38
Spline (3df)	Random intercept + quadratic slope	34,939.57	35,013.77
Spline (3df)	Random intercept + splined slope (2df)	34,930.19	35,004.39
Spline (4df)	Random intercept	36,810.21	36,870.49
Spline (4df)	Random intercept + slope	35,193.12	35,262.68
Spline (4df)	Random intercept + quadratic slope	34,724.26	34,807.73
Spline (4df)	Random intercept + splined slope (2df)	34,684.71	34,768.18
Spline (5df)	Random intercept	36,779.19	36,848.75
Spline (5df)	Random intercept + slope	35,143.63	35,222.47
Spline (5df)	Random intercept + quadratic slope	34,670.60	34,763.35
Spline (5df)	Random intercept + splined slope (2df)	34,633.04	34,725.78
Spline (6df)	Random intercept	36,681.60	36,760.44
Spline (6df)	Random intercept + slope	34,998.78	35,086.89
Spline (6df)	Random intercept + quadratic slope	34,505.46	34,607.48
Spline (6df)	Random intercept + splined slope (2df)	34,461.01	34,563.03
Spline (7df)	Random intercept	36,636.48	36,724.59
Spline (7df)	Random intercept + slope	34,927.75	35,025.14
Spline (7df)	Random intercept + quadratic slope	34,420.46	34,531.76
Spline (7df)	Random intercept + splined slope (2df)	**34,374.32**	**34,485.61**

Given that $h_0(t)$ in Equation (6.27) is left unspecified, I use 1,000 non-parametric bootstrap replications (resampling at the cluster level) to estimate standard errors for all regression coefficients. Finally, I fit the equivalent linear mixed model (Equation (6.26)) disregarding the drop-out process using the `nlme` R package [198] for comparison purposes.

Focussing first on the survival sub-model, the estimated hazard ratio of drop-out

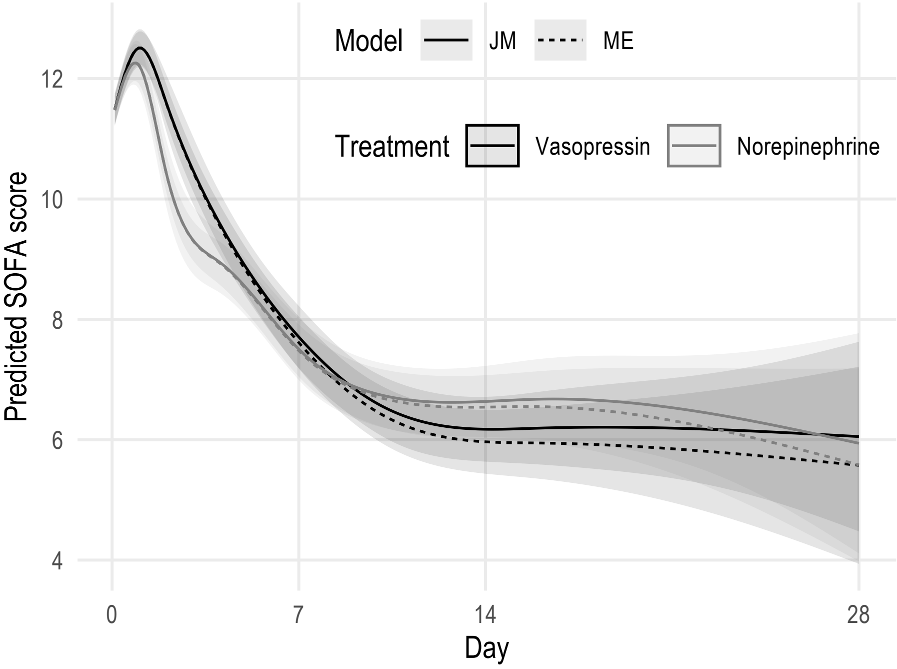


FIGURE 6.6: Predicted longitudinal trajectories by treatment arm, application to VASST data. Solid lines represent trajectories fitted using a joint model, while dashed lines represent trajectories fitted using a linear mixed model

for norepinephrine versus vasopressin is 1.04 (95% C.I.: (0.85–1.23)), showing a non-significant difference in the risk of drop-out between the two treatment arms. The association parameter α takes a positive value (0.14, with 95% C.I.: 0.10–0.17) showing a significant positive association between the longitudinal and the time to event outcomes. In particular, the association parameter can be interpreted as follows: larger subject-specific deviations from the average longitudinal trajectory are associated with an increased risk of drop-out. This result is consistent with the pattern observed in the raw trajectories of Figure 6.5.

Coefficients for the longitudinal sub-model are omitted, given their difficult interpretation due to the interaction with spline terms; instead, I include predicted trajectories for each treatment arm in Figure 6.6. Testing the joint significance of the regression coefficients associated with the main fixed effect of time using a Wald test, I obtain a χ^2 test statistic value of 901.11 with a p-value of < 0.01. Analogously for the interaction between time and treatment, the χ^2 test statistic value of 1,683.08 yields a p-value of < 0.01. This shows that the longitudinal SOFA score trajectory differs significantly between treatment arms, as depicted in Figure 6.6.

Despite the time–treatment interaction being statistically significant, the difference

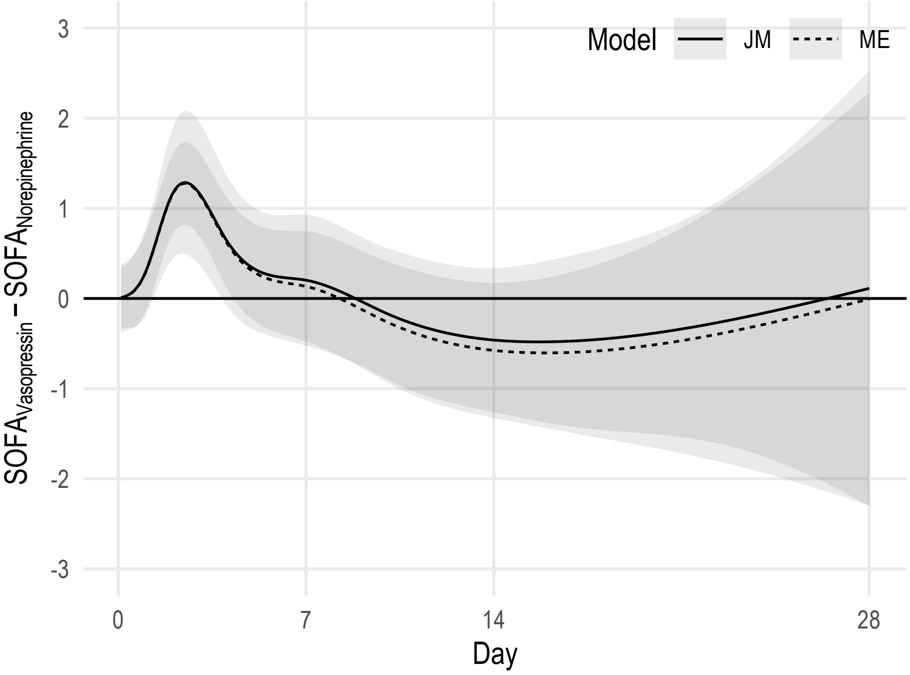


FIGURE 6.7: Difference in predicted SOFA score between treatment arms, application to VASST data. The solid line is obtained from a joint model, while the dashed line is obtained from a linear mixed model

between treatment arms does not seem to be clinically relevant. For instance, there is a advantage of vasopressin early on with a maximum difference of 1.17 at day 3. Nevertheless, the advantage of vasopressin fades quickly, with no significant difference between arms from day 4 onwards. The difference in SOFA score between treatment arms over time is depicted in Figure 6.7.

Finally, in Figures 6.6 and 6.7 I also include trajectories and difference between trajectories estimated using a plain linear mixed model. Longitudinal trajectories estimated with either a joint model or a mixed model are almost identical early on; however, as time goes by, more and more individuals drop-out of the study and the difference between the trajectories increases. In particular, individuals dropping out have (on average) higher SOFA score values (as depicted in Figure 6.5), hence the longitudinal trajectories are under-estimated by the plain mixed model. Interestingly, the estimated difference between treatment arms was very similar between the joint model and the plain mixed model, with the difference between methods being not relevant in clinical terms.

In conclusion, the results from the joint model show that there is a statistically significant difference in the evolution of SOFA score over time between subjects receiving vasopressin and individuals receiving norepinephrine. Despite that, the difference between treatment arms is perhaps not clinically relevant, with a maximum difference of approximately 1 SOFA score point early on during follow-up and no difference later on; thus, the more rapid decline in the SOFA score in norepinephrine compared to the vasopressin group would not likely change current practice. Furthermore, I showed that by using a joint model the estimated trajectories differ compared to a plain linear mixed model that disregards the drop-out process: in the settings of VASST, the mixed model under-estimated the longitudinal trajectories late during follow-up as more individuals dropped out of the study - even though the difference between the joint models and the plain linear mixed model was not clinically meaningful.

6.8 Discussion

The topic of joint modelling of longitudinal and time to event data has received considerable interest in the past years since the seminal papers by Wulfsohn and Tsiatis and Henderson *et al.* [163, 165]. This methodology provides an attractive framework for assessing the association between a longitudinal and a survival outcome (or vice-versa), given the opportunity to study and model their relationship in a variety of clinically meaningful formulations.

Despite that, the application in practice is scarce - possibly because of the complex technicalities and computational requirements. The availability of user-friendly software and excellent review papers and books on the topic [167–169] has surely contributed to the adoption of this methodology; regardless, the heavy computational requirements still stand.

Throughout this Chapter I introduced and formalised joint models for longitudinal and survival data, focussing on the standard joint model formulation with a single longitudinal outcome and a single time to event outcome. I illustrated commonly used association structures and the estimation process, highlighting the heavy computational requirements of this methodology.

I approached the topic of joint modelling from the (arguably) less common point of view of focussing on the longitudinal outcome as the primary outcome of interest. In the settings of longitudinal data analysis, it has been shown that truncation of the longitudinal trajectory (i.e. because of drop-out from the study) leads to biased results if the truncation process and the outcome of interest are not independent. This issue is commonly referred to as informative drop-out, and I described it in more detail in Section 6.6; in particular, I defined characteristics of the drop-out process and formalised the problem within a missing data framework. Finally, in Sections 6.6 and 6.7 I described how the joint modelling approach introduced in this Chapter can be used to jointly model the longitudinal outcome of interest and the drop-out process: by doing so, the longitudinal analysis accounts for the truncation of the longitudinal trajectory avoiding the biases that would otherwise arise.

The applied example of Section 6.7 using data from VASST illustrates the joint modelling approach in practice, including a comparison with a standard mixed-effects model that ignores the drop-out process. The results of the application motivate the use of joint modelling in these settings, as the predicted trajectories that are obtained from the two approaches (Figure 6.6) diverge more and more as time goes by and more individuals drop-out of the study. A manuscript based on these results is currently under review for publication in Critical Care Medicine, including a tutorial on the use of joint modelling in the settings of intensive care medicine.

The joint modelling framework will be used further in Chapter 7, where I will describe how a joint longitudinal-survival model can be used to model longitudinal data when the assumption of independence between the outcome and the timing between observations does not hold. More details on the issue of informative observation times are also included in Chapter 7, complementing the issue of informative drop-out discussed within this Chapter.

7 *Modelling the Observation Process*

7.1 OUTLINE

The analysis of longitudinal data is essential to understand the evolution of a disease and the effect of interventions over time. As outlined in Chapter 1, longitudinal data is prevalent in the settings of EHRs; however, methodological challenges arise when applying traditional analysis methods in these settings. First and foremost, observation times are likely to be correlated with the underlying disease severity in healthcare consumption datasets: individuals tend to have irregular observation times as patients with more severe conditions (or showing early symptoms of a disease) tend to seek medical care more often than those with milder conditions (and no symptoms). Their worse disease status is also likely to be reflected in worse biomarkers being recorded at such visits, causing abnormal values of such biomarkers to be overrepresented and normal values to be under-represented.

Traditional methods used to analyse longitudinal data rely - among others - on the following assumptions:

1. Study drop-out is independent of disease severity, as discussed in Section 6.6;
2. The underlying mechanism that controls the observation time must be independent of disease severity.

Unfortunately, these assumptions are unlikely to hold in the settings of EHRs. In the previous Chapter, I discussed how failing to account for informative drop-out in a longitudinal study could yield biased estimates of the model parameters [193]; in this Chapter I will focus on the problem of informative observation times, as it can be shown that bias ensues if naively applying traditional methods when the follow-up is irregular and associated with the outcome [199]. Despite the potential for bias, there is some evidence pointing towards a lack of awareness in longitudinal studies with healthcare data irregularly collected over time: a recent literature review on the topic showed that 86% of the included studies did not report enough information to evaluate whether the visiting process was informative or not, and only one study used a method capable of dealing with an informative observation process [200]. This is especially concerning when the aim of a research project is aetiology.

As mentioned above, in this Chapter I focus on the problem of informative observation times and the biases that may arise when data on covariates and outcomes is collected at irregular, subject-specific intervals: in fact, when analysing data originating from electronic health records, data is collected only when study subjects consume health care (e.g. by visiting their doctor or going to the hospital). As a consequence, visit times are likely to be informative and to depend on the clinical history and/or health status of an individual. Characteristics of the observation process and biases that arise when the observation process can be deemed informative are discussed in Section 7.2.

Methods that have been developed to deal with this problem and published in the current literature are discussed in Section 7.3. In particular, I will focus on two broad families of methods, methods based on inverse probability weighting and methods based on joint modelling. Further to that, I will describe a more general joint modelling approach to the issue of informative observation times in Section 7.4, and compare the performance of some of the methods in simulated settings in Section 7.5. The Monte Carlo simulation study has been published in Statistica Neerlandica, featuring in a Special Issue on the 2018 Survival Analysis for Junior Researchers conference, with more details included in Appendix F [201]. The results of the simulation study are also illustrated in practice using data from the PSP-CKD study in Section 7.6.

Finally, I will conclude the Chapter with a discussion in Section 7.7.

7.2 CHARACTERISTICS OF THE OBSERVATION PROCESS

An observation process can have regular or irregular visits. With regular visits, the j^{th} visit time for the i^{th} individual T_{ij} is the same for all individuals: $T_{ij} = t_j \; \forall \; i, \; j$, with $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, n_i$. Conversely, with irregular visits that is no longer true.

I start by defining the observation process $N_i(t)$ using counting process notation. A counting process $N_i(t)$ is a stochastic process with values that are non-negative, integer, and non-decreasing:

$$\begin{cases} N_i(t) \geq 0 \\ N_i(t) \text{ is an integer} \\ N_i(s) \leq N_i(t) \text{ if } s \leq t \end{cases} \tag{7.1}$$

Practically speaking, the observation process is a counting process that increments every time a new observation is recorded.

When the visiting pattern is irregular, $N_i(t)$ can be defined to be completely at random when visit times and outcome(s) are independent [199]:

$$E[\Delta N_i(t) | \bar{Y}_i(\infty), \bar{X}_i(\infty)] = E[\Delta N_i(t)], \tag{7.2}$$

where $\Delta N_i(t) = N_i(t) - N_i(t^-)$, with t^- being the instant of time right before t. $\bar{Y}_i(\infty)$ and $\bar{X}_i(\infty)$ denote the values of outcome and covariates for any $t > 0$.

The observation process can be deemed *informative* when it is not completely at random, i.e. when the condition above is not verified. In that case, it is possible to identify the following two scenarios:

- Observation process at random, when visiting at time t is independent of the outcome at time t given data recorded up to time t:

  $$E[\Delta N_i(t) | \bar{X}_i(t), \bar{N}_i(t^-), \bar{Y}_i^{\text{obs}}(t^-), Y_i(t)] = E[\Delta N_i(t) | \bar{X}_i^{\text{obs}}(t), \bar{N}_i(t^-), \bar{Y}_i^{\text{obs}}(t^-)], \tag{7.3}$$

  where $\bar{X}_i(t)$ and $\bar{X}_i^{\text{obs}}(t)$ denote the covariates history up to time t and its observed values, $\bar{N}_i(t^-)$ denotes the history of the observation process up to time t^-, and $\bar{Y}_i^{\text{obs}}(t^-)$ the observed values of the outcome up to time t^-;

- Observation process not at random, where the definition of missing at random does not hold. That is, the scenario where visiting at time t is not independent of the outcome at time t, even after conditioning on data recorded up to time t:

  $$E[\Delta N_i(t) | \bar{X}_i(t), \bar{N}_i(t^-), \bar{Y}_i^{\text{obs}}(t^-), Y_i(t)] \neq E[\Delta N_i(t) | \bar{X}_i^{\text{obs}}(t), \bar{N}_i(t^-), \bar{Y}_i^{\text{obs}}(t^-)] \tag{7.4}$$

Gruger *et al.* [202] illustrate four possible models that could be linked to the above-mentioned scenarios:

1. The *examination at regular intervals* model, consisting of observation times that are pre-defined and equal for all patients (as in clinical trials). This scenario yields so-called *balanced panel data*;
2. The *random sampling* model, consisting of a sampling scheme (e.g. an observation process) that is not pre-defined, but still independent of the disease history of the study subjects;
3. The *doctor's care* model, consisting of an observation process that depends on the characteristics of the patient at the moment of the current doctor's examination. For instance, a doctor could require stricter monitoring for subjects with more advanced disease status, or with abnormal values of a biomarker;
4. The *patient self-selection* model, yielding observations that are triggered by the patients themselves. According to this model, patients may choose to visit their doctor when they feel unwell, or they may choose to skip a visit that was pre-planned when they feel the treatment they are receiving is not beneficial to their health status. Unfortunately, the factors that cause patients to self-select themselves are generally unknown or not recorded.

Models (1) and (2) could be characterised as *observation completely at random*; model (3) could be characterised as *observation at random*; finally, model (4) could be characterised as *observation not at random.*

The bias that one may encounter when the observation process is informative can be classified into two types: selection bias and confounding [203]. Selection bias arises because of the inclusion of only observed individuals in the analysis. This bias is the same bias induced by informative censoring due to loss to follow-up [204]: censoring is the extreme case of an observation process where an individual is not observed ever again. Conversely, confounding arises when there are common causes of both the exposure and the outcome, where I consider the exposure to be the observation process. For instance, when visit times are decided by a physician or by the patient itself based on e.g. current health status, which itself is associated with the observed longitudinal outcome, then ignoring the observation process in the analysis yields confounding. The settings of dynamic observation processes are discussed in more detail elsewhere, including directed acyclic graphs (DAGs) that illustrate the underlying causal mechanisms [203].

7.3 METHODS TO ACCOUNT FOR AN INFORMATIVE OBSERVATION PROCESS

Standard methods for the analysis of longitudinal data (such as those described in Section 2.3) can be used when the visiting process can be deemed completely at random, according to the definition introduced in the previous Section. Conversely, the analysis method needs to explicitly account for the visiting process; a general overview of methods that can be used in the settings of informative visiting processes is presented in Pullenayegum and Lim [199].

In particular, methods can be broadly classified into methods based on inverse intensity of visiting weighting, and methods based on joint modelling the longitudinal outcome and the visiting process. These two approaches are described in more detail in Sections 7.3.1 and 7.3.2, respectively.

7.3.1 Inverse Intensity of Visiting Weighting

Inverse intensity of visiting weighting (IIVW) was first proposed by Lin *et al.* and Robins *et al.* and further extended by Buzkova and Lumley [190, 205, 206]. The IIVW approach accommodates an informative observation process in a marginal regression model by weighting each observation by the inverse of the probability of each measurement to be recorded; consequently, this approach creates a pseudo-population in which the observation process is static (e.g. completely at random) and can, therefore, be ignored. The weights can be estimated by fitting a regression model including all covariates that inform the observation process and further stabilised to increase efficiency [207].

This method assumes a general link function g for the marginal model for the longitudinal outcome, generally a GEE model, and possibly time varying covariates $X(t)$:

$$g[\mu_i(t)] = X_i(t)\beta \tag{7.5}$$

Then, it is assumed that there are auxiliary variables $Z_i(t)$ (which may include functions of $N_i(s)$, $X_i(s)$, $\bar{Y}^{\text{obs}}(s)$ for any $s < t$) such that the visiting process is independent of the current outcome given the observed auxiliary variables at time t:

$$\begin{aligned}
&\lim_{\delta \to 0} \frac{E[N_i(t) - N_i(t - \delta)|X_i(t), Z_i(t), Y_i(t)]}{\delta} \\
&= \lim_{\delta \to 0} \frac{E[N_i(t) - N_i(t - \delta)|Z_i(t)]}{\delta} \\
&= h(t|Z_i(t))
\end{aligned} \tag{7.6}$$

$h(t, Z_i(t))$ represents then the intensity of visiting, which is assumed to follow a proportional hazards model (such as those discussed in Section 2.2):

$$h(t, Z_i(t)) = h_0(t) \exp(Z_i(t)\gamma) \tag{7.7}$$

This proportional hazards model is used to estimate the following weights:

$$w_i(t) = \frac{s(t)}{h_0(t) \exp(Z_i(t)\gamma)}, \tag{7.8}$$

where $s(t)$ is a stabilising function. By choosing the baseline hazard function $h_0(t)$ as stabilising function, the weights become $1/\exp(Z_i(t)\gamma)$ therefore removing the need to estimate the baseline hazard. These stabilised weights are then used to weight the GEE analysis and can account for the settings of observation processes at random (according to the definition introduced in the previous section).

In addition to weighting a GEE analysis, this method could be used to weight the Lin and Ying estimating equations for irregular longitudinal data [208, 209]. In brief, the Lin and Ying estimating equations have been proposed to accommodate irregular longitudinal data and extend the traditional GEE approach by allowing for a non-parametric intercept in the semi-parametric model; more details on this method are presented elsewhere [199].

Finally, doubly robust estimating equations have been proposed for the settings of irregularly recorded longitudinal data [210]. The benefit of doubly robust inference is that it gives consistent estimates of the regression coefficients if either the model for the visiting process or the model for the outcome is correctly specified; that is, this method is robust to misspecification of one (but not both) of the models.

7.3.2 Joint Modelling

Semiparametric and parametric joint models for the longitudinal outcome and the visiting process are discussed in Pullenayegum and Lim [199].

Several semiparametric joint models have been proposed in the literature [211–214], with each formulation differing in terms of modelling assumptions:

1. Whether time-independent or time-dependent covariates can be included in the model;
2. Whether covariates included in the longitudinal and visiting process models are constrained to be the same;
3. The parametrisation of the random effects.

All semiparametric models aim to capture the correlation between the outcome and visit processes via shared or correlated random effects, without requiring the explicit specification of a parametric model for either process or the random effects; all models use nonparametric intercepts in the longitudinal model. Nevertheless, this class of models offers considerable flexibility in modelling the visit and outcome models.

Another assumption of semiparametric joint models is that the random effects are time-invariant; some authors show that this assumption could be relaxed in some settings, e.g. when assuming that the conditional mean of the random effects given the covariates is zero and that the covariance of the random effects at time t does not depend on the values of the observed covariates at time t.

If willing to specify a parametric model for the outcomes (including the distribution of the random effects), a more general dependence structure between outcomes can be specified; this yields the possibility to do inference via maximum likelihood. An example of a parametric joint model is given in Liu *et al.* [215]: they illustrate a trivariate joint model that accommodates a longitudinal outcome, the observation process, and the drop-out process as well. Their model assumes that the random effects follow a multivariate normal distribution and that the baseline intensity functions (for the visiting and drop-out processes) follow a piece-wise constant function.

The major drawback of the joint models introduced in this section is that they are not readily usable in practice: they are not implemented in standard statistical software packages, and the code required to fit the models is often non-trivial and/or not openly available at all. In the next Section, I will introduce a joint modelling approach to modelling the longitudinal outcome and the observation process based on the joint modelling formulation of Chapter 6; most interestingly, the joint models under this approach can be fit using readily available statistical software and can be easily extended to accommodate additional complexity.

7.4 AN EXTENDED JOINT MODELLING APPROACH

The joint modelling approach to account for the observation process in the analysis of longitudinal data can be formulated in the joint modelling framework described in Chapter 6. Let $D_{ij}(t) = I(T_{ij} = t)$ denote the presence of an observation at time t for the i^{th} individual: at each $D_{ij}(t) = 1$ a new observation of the longitudinal outcome Y_{ij} is recorded. Let $\tilde{t}_{ij}$ be the gap time between the j^{th} and $(j+1)^{\text{th}}$ measurement for the i^{th} individual. Let $\tilde{d}_{ij}$ be the binary indicator variable that denotes whether the gap-time $\tilde{t}_{ij}$ is observed (or not). In practice, gap-time are always observed except when the observation process is censored at the end of follow-up, e.g. the date when the data extraction occurs. Let z_{ij} be the covariate vector for the longitudinal outcome, and w_i the covariate vector for the observation process; z_{ij} and w_i do not necessarily overlap, and it is assumed that both could be extended to include time-dependent exogenous covariates (e.g. w_{ij}). The observation process and the repeated measures process are modelled jointly using a joint longitudinal-survival model as described in Chapter 6.

Conditional on the random effect u_i, the submodel for the time to each observation is a proportional hazards model with hazard for gap time $\tilde{t}_{ij}$:

$$r(\tilde{t}_{ij} | w_{ij}, u_i, \theta_t) = r_0(\tilde{t}_{ij}) \exp(w_{ij}\beta + u_i), \tag{1}$$

where $\theta_t = \beta$. This is a recurrent events model where the within-individual correlation is accounted for via the frailty, as described in Section 5.4. $r_0(\tilde{t}_{ij})$ can be any parametric or flexible parametric [24] baseline hazard function (also referred to as baseline intensity, I will use the terms hazard and intensity interchangeably).

The submodel for the j^{th} longitudinal observation of the i^{th} individual is

$$(y_{ij}|D_{ij}(t) = 1, z_{ij}, u_i, v_i, \theta_y) = m_{ij} + \epsilon_{ij} = z_{ij}\alpha + \gamma u_i + v_i + \epsilon_{ij}, \tag{2}$$

where $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ and $\theta_y = \{\alpha, \gamma, \sigma_\epsilon^2\}$. This is a mixed-effects model that can be used to model longitudinal data, as described in Section 2.3, with a random intercept v_i.

The two processes are linked together via the shared, individual-specific, random effect u_i. Including the γ parameter in the longitudinal model allows for an association between the two equations that will be estimated from data. When $\gamma = 0$, the two processes are independent of each other. Finally, the random effects are assumed to follow a multivariate normal distribution with null mean vector and variance-covariance matrix $\Sigma_{u,v}$.

The model is fitted using maximum likelihood; the individual-specific contribution to the likelihood can be written as:

$$\begin{aligned} L_i(\theta) &= \int p(\tilde{t}_{ij}, \tilde{d}_{ij}, y_{ij}, b_i; \theta)\ db_i \\ &= \int \prod_{j=1}^{n_i} \left[p(\tilde{t}_{ij}, \tilde{d}_{ij}|b_i, \theta_t) p(y_{ij}|b_i, \theta_y) \right] p(b_i|\theta_b)\ db_i \end{aligned} \tag{7.9}$$

where $\theta = \{\theta_t, \theta_y, \theta_b\}$ is the overall parameters vector, $b_i = \{u_i, v_i\}$ is the vector of random effects,

$$p(\tilde{t}_{ij}, \tilde{d}_{ij}|b_i, \theta_t) = r(\tilde{t}_{ij}|w_{ij}, u_i, \theta_t)^{\tilde{d}_{ij}} \exp\left(- \int_0^{\tilde{t}_{ij}} r(s|w_{ij}, u_i, \theta_t)\ ds \right) \tag{7.10}$$

is the contribution to the likelihood of the time to the j^{th} observation in individual i,

$$p(y_{ij}|b_i, \theta_y) = (2\pi\sigma_\epsilon^2)^{-1/2} \exp\left(-\frac{(y_{ij} - m_{ij})^2}{2\sigma_\epsilon^2} \right) \tag{7.11}$$

is the contribution of the j^{th} longitudinal observation, and $p(b_i|\theta_b)$ is the density of the random effects. The likelihood does not have a closed-form, as outlined in Section 6.5; therefore, it is necessary to integrate out the distribution of the random effects using methods outlined in Section 5.4.2.

A simplified DAG that illustrates how the joint model accounts for the correlation between a longitudinal outcome Y and its observation process R is included as Figure

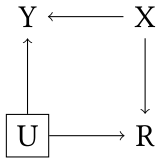


FIGURE 7.1: Simplified DAG depicting a joint model for a longitudinal outcome and its observation process.

7.1 [137]; X represents covariates included in the model, and U represents the shared random effects. After adjusting for all covariates X, the longitudinal outcome and the observation process are associated only through the shared U. However, when estimating the joint model the random effects U are assumed to follow a given distribution (e.g. Gaussian) and are then integrated out of the marginal likelihood, blocking the path between Y and R. Therefore, for the joint model to be valid the observation process has to be at least *at random*, according to the definition of Section 7.2.

I will focus on the aforementioned model for (1) simplicity and (2) as it is comparable with the trivariate joint model suggested by Liu *et al.* and mentioned in the previous Section [215]. However, this model is nested within a wide family of multivariate generalised linear and non-linear mixed-effects models [216] and can be easily extended. For instance:

1. Multiple random effects, with potentially different levels of nesting, can be included;
2. Different distributions for the baseline hazard can be assumed, ranging from standard parametric distributions (such as exponential, Weibull, Gompertz) to spline-based formulations on the log-cumulative hazard or log-hazard scale;
3. Additional outcomes can be accommodated, such as multiple longitudinal outcomes or a drop-out process;
4. The association structure between sub-models can be extended to any of the association structures described in Section 6.4.

Most interestingly, the joint model I will focus on (and several extensions) can easily be fitted using the user-written command `merlin` in Stata and R [217, 218].

7.5 A Methods Comparison Using Monte Carlo Simulation

As described in the previous Section, several methods have been proposed in the current literature to account for the observation process in the analysis of longitudinal data. Despite that, Farzanfar *et al.* showed that such methods are not routinely used in applied research [200]: they described a low proportion of studies reporting on the potential informativeness of visit times, and they concluded by arguing that there is a need for guidance to researchers on the potential for bias and the reporting of longitudinal studies subject to irregular follow-up.

The lack of awareness of methods to account for informative visiting times in the analysis could also stem from the lack of methods comparisons in the literature: in order to choose the appropriate analysis method, researchers need to understand assumptions, benefits, and potential pitfalls underlying each method. A qualitative comparison of methods was published by Pullenayegum and Lim [199], but to the best of my knowledge, there is only one quantitative comparison existing in the current literature. Moreover, that comparison yielded negative results: Neuhaus *et al.* concluded that fitting ordinary linear mixed models that disregard the observation process completely yielded the smallest bias and showed that adding regular visits to the observation schedule (if possible) reduced that bias even more [219].

Therefore, I set out to design and run a Monte Carlo simulation study aimed at assessing the impact of ignoring the observation process in the analysis of longitudinal data. The simulation study is described in Sections 7.5.1 to 7.5.6 using the ADEMP structure; its results are summarised in Section 7.5.7.

7.5.1 Aims

As mentioned before, the aim of this simulation study consists of assessing the impact of ignoring the observation process in longitudinal mixed-effects models when the observation process is informative, while at the same time comparing methods that have been proposed to account for the observation process in the analysis.

7.5.2 *Data-Generating Mechanisms*

I simulate data from the following joint model:

$$
\begin{aligned}
r(\tilde{t}) &= r_0(\tilde{t}) \exp(Z_i \beta + u_i) \\
y_{ij} | (D_{ij}(t) = 1) &= \alpha_0 + Z_i \alpha_1 + t_{ij} \alpha_2 + \gamma u_i + v_i + \epsilon_{ij}
\end{aligned}
\tag{7.12}
$$

Z_i is a time-invariant covariate (for simplicity) representing a binary treatment, simulated from a Bernoulli random variable with probability 0.5: $Z_i \sim \text{Bern}(1, 0.5)$. The coefficient associated to the treatment variable is $\beta = 1$ for the observation process, $\alpha_1 = 1$ for the longitudinal process. The fixed intercept of the longitudinal model is $\alpha_0 = 0$, and the fixed effect of time is $\alpha_2 = 0.2$. The random effects u_i and v_i are simulated from a Normal random variable with null mean and variance $\sigma_u^2 = 1$ and $\sigma_v^2 = 0.5$, respectively. The residual error of the longitudinal model is assumed to follow a Normal distribution with null mean and variance $\sigma_\epsilon^2 = 1$.

I assume independence between the random effects and the residual variance, and between random effects (i.e. $\Sigma_{u,v}$ is a diagonal matrix with $\text{diag}(\Sigma_{u,v}) = \{\sigma_u^2, \sigma_v^2\}$); further to that, I assume independent random effects for simplicity - however, it would be possible to accommodate correlated random effects within the data-generating model. The joint model with correlated random effects can be thought of as a reparameterization of the joint model with independent random effects, where the association parameter γ is related to the correlation between the two random effects in the bivariate version.

The baseline hazard from the recurrent visit process is assumed to follow a Weibull distribution with shape parameter $p = 1.05$; I vary the scale parameter λ and therefore the baseline intensity of the visiting process, with $\lambda = \{0.10, 0.30, 1.00\}$. This baseline intensities along with the value of β correspond to an expected median gap time between observations of 5.83 and 2.25 years for unexposed and exposed individuals if $\lambda = 0.10$, 2.05 and 0.79 years if $\lambda = 0.30$, and 0.65 and 0.25 years if $\lambda = 1.00$, respectively. Each observation time is simulated using the inversion method described in Section 4.4, assuming a gap time scale (where the time index is reset to zero after the occurrence of each observation; the resulting recurrent events model is then a semi-Markov model).

I vary the association parameter γ between the two sub-models, with $\gamma = \{0.00, 1.50\}$;

I expect all methods to perform similarly when $\gamma = 0$, that is, when the longitudinal process is independent of the observation process.

In addition to simulating data from the joint model above, I also generate the observation process by drawing from a Gamma distribution. Specifically, I draw the observation times from a Gamma distribution with shape = 2.00 and scale:

$$\exp(-\psi\beta Z_i + \xi_i), \tag{7.13}$$

where ξ_i is simulated from a Normal distribution with null mean and variance $\sigma_\xi^2 = 0.1$. Z_i is the same binary treatment covariate as before, with the same associated parameter $\beta = 1$. The value of ψ defines the association between the observation, e.g. when $\psi = 0$ the observation process is not informative; I vary and set $\psi = \{0.00, 2.00\}$. I also simulate a scenario where the observation process depends on the treatment and previous values of the longitudinal outcome Y. In this setting, observation times are drawn from a Gamma distribution with shape = 2.00 and scale:

$$\exp(-\psi\beta Z_i + \omega y_{i,j-1} + \xi_i) \tag{7.14}$$

for the j^{th} observation time of the i^{th} individual, with $\psi = 2.00$ and $\omega = 0.20$.

Finally, I simulate a scenario from a joint model to which regular (i.e. planned) visits are added every year, as suggested by Neuhaus *et al.* [219]. This scenario is simulated from the above-mentioned joint model, with $\gamma = 3.00$ and $\lambda = 0.05$ to obtain an observation process that is sparse and strongly associated with the longitudinal outcome.

200 study individuals are simulated under each data-generating mechanism, and the recurrent observation process continues for each individual until the occurrence of administrative censoring - which is simulated from a $\text{Unif}(5, 10)$ random variable. The last gap time for each individual is defined as the difference between the last observation and the censoring time.

7.5.3 Estimands

The main estimand of interest is the vector of regression coefficients $\alpha = \{\alpha_0, \alpha_1, \alpha_2\}$, with specific focus on the treatment effect α_1. I will also report on the estimated association parameter γ and on the estimated variance of the random effects and the residual errors: σ_u^2, σ_v^2, and σ_ϵ^2.

7.5.4 *Methods and Software*

I fit five competing models to each simulated dataset:

1. Model A, the joint model described above (at the beginning of the "Data-generating mechanisms" Section) and corresponding to the true data-generating mechanisms when simulating data from a joint model;
2. Model B, a linear mixed model including the number of visits (centred on the mean value) as a fixed effect in the model;
3. Model C, a linear mixed model including the cumulative number of visits as a fixed effect in the model;
4. Model D, a linear mixed model that disregards the observation process completely;
5. Model E, a marginal model fitted using generalised estimating equations and inverse intensity of visiting weights.

Model A is described in more detail in Section 7.4 and fit using `merlin` [217] and `gsem` in Stata.

Model B follows from previous work by Goldstein *et al.* [220], where they demonstrate that conditioning on the number of health-care encounters it is possible to remove bias due to an informative observation process (they denote this bias as *informed presence bias*). I therefore include the number of observations per individual, centred on the mean value, in a mixed-effects model for the longitudinal outcome:

$$y_{ij} = \alpha_0 + Z_i\alpha_1 + t_{ij}\alpha_2 + n_i^c\alpha_3 + v_i + \epsilon_{ij}, \tag{7.15}$$

with v_i a random intercept and n_i^c the number of observations for the i^{th} individual.

Model C is analogous to model B, adjusting for the cumulative number of measurements up to time j instead, denoted as $\bar{n}_{it_j}$:

$$y_{ij} = \alpha_0 + Z_i\alpha_1 + t_{ij}\alpha_2 + \bar{n}_{it_j}\alpha_3 + v_i + \epsilon_{ij} \tag{7.16}$$

Sun *et al.* [221] suggest that a straightforward approach would be including variable that records the number of visits prior to the current observation time in the model for the longitudinal response of interest. Intuitively, one would expect that responses from individuals with several previous visits would differ from individuals with only a few visits, and therefore including the number of prior visits as a covariate could control for these differences.

Model D is a standard mixed model that does not account for the visiting process, i.e. analogous to model B and C but assuming $\alpha_3 = 0$.

Models B, C, and D are fit using the `mixed` command in Stata; models A, B, C, D are fit assuming an independent structure for the variance-covariance matrix of the random effects.

Finally, model E is fitted following the two-stage procedure described in Section 7.3.1, and following Van Ness *et al.* [222]. The model used to estimate weights is an Andersen-Gill recurrent events model [94] for the observation process, assuming a gap-time scale:

$$r(\tilde{t}_{ij}) = r_0(\tilde{t}_{ij}) \exp(z_i \eta), \tag{7.17}$$

where $\tilde{t}$ are gap-times between consecutive observations, $r_i(\tilde{t})$ is the intensity of visit for individual i at gap-time $\tilde{t}$, $r_0(\tilde{t})$ is the unspecified baseline intensity at gap-time $\tilde{t}$, and z_i is a vector of coefficients that are assumed to accurately describe the observation process for individual i. η is a vector of regression coefficients that is estimated using the Cox partial likelihood method and a robust jack-knife estimator for the variance of the regression coefficients. The inverse intensity of visit weights are estimated by taking the inverse of the linear predictor $\exp(z_i \hat{\eta})$ at each time point, and further normalised by subtracting the mean inverse weight and adding the value 1 to each weight; the distribution of the weights is therefore centred on the value 1. Next, two further adjustments are needed. First, since the last data point for each individual represents the end of follow-up of the study, each weight is shifted by one time point. Second, given that each individual is observed at least once (i.e. at baseline), a weight of one is assigned to the first observation of each individual.

The marginal model for the longitudinal outcome is then fit using generalised estimating equations and including the normalised inverse intensity of visit weights as probability weights in the model. The model has the form:

$$E(y_{ij}) = \alpha_0 + Z_i \alpha_1 + t_{ij} \alpha_2, \tag{7.18}$$

and can be fit using readily available statistical software; for instance, I use the Stata command `glm`.

In conclusion: the simulation study is coded and run using Stata version 15, built-in functions (such as `mixed`, `glm`, `gsem`), and the user-written commands `survsim` and `merlin` [68, 217]. Results of the simulation study are summarised using R [86] and the R package `rsimsum` [223]. All the code required to simulate data, fit each model, and produce summaries is available online (`https://github.com/ellessenne/infobsmcsim`).

7.5.5 *Performance Measures*

I will assess average estimates and standard errors, empirical standard errors, bias, and coverage probability of $\hat{\alpha}_m$, with $m = \{0, 1, 2\}$. However, the main performance measures of interest are bias and coverage probability: the former quantifies whether an estimator targets the true value on average, while the latter represents the proportion of times that a confidence interval based on $\hat{\alpha}_{m,k}$ and $\hat{\text{SE}}(\hat{\alpha}_{m,k})$ contains the true value α_m, with k indexing each replication. Monte Carlo standard errors, useful to quantify the uncertainty in estimating bias and coverage, are estimated and reported as well [58].

7.5.6 *Number of Simulations*

The process of defining the number of simulations is analogous to that of Section 5.5.6. If assuming that $\text{Var}(\hat{\alpha}_m) \leq 0.1$ (or, equivalently, $\text{SE}(\hat{\alpha}_m) \leq 0.32$) and requiring a Monte Carlo standard error for bias of 0.01 or lower, given that $\text{MCSE(Bias)} = \sqrt{\text{Var}(\hat{\alpha}_m)/K}$, I would require a number of replications K = 1,000. The assumed standard error is larger than the standard errors reported by Liu *et al.* for a model similar to model A [215]. The expected Monte Carlo standard error for coverage, assuming a worst-case scenario of 50% coverage, would be 1.58% - which I deem acceptable once again. Therefore, I proceed by simulating 1,000 independent datasets for this simulation study.

Table 7.1: Number and percentage of models converging under each data-generating mechanism, Monte Carlo simulation study on modelling the observation process. Model A is the joint model, model B is the mixed model adjusting for the total number of measurements, model C is the mixed model adjusting for the cumulative number of measurements, model D is the mixed model with no further adjustment, and model E is the marginal model fitted using GEE and IIVW

Data-Generating Mechanism	Model A	Model B	Model C	Model D	Model E
JM ($\gamma = 0.00, \lambda = 0.10$)	100.00%	100.00%	100.00%	100.00%	100.00%
JM ($\gamma = 0.00, \lambda = 0.30$)	100.00%	100.00%	100.00%	100.00%	100.00%
JM ($\gamma = 0.00, \lambda = 1.00$)	100.00%	100.00%	100.00%	100.00%	100.00%
JM ($\gamma = 1.50, \lambda = 0.10$)	95.50%	100.00%	100.00%	100.00%	100.00%
JM ($\gamma = 1.50, \lambda = 0.30$)	100.00%	100.00%	100.00%	100.00%	100.00%
JM ($\gamma = 1.50, \lambda = 1.00$)	100.00%	100.00%	100.00%	100.00%	100.00%
Γ not depending on treatment	100.00%	100.00%	100.00%	100.00%	100.00%
Γ depending on treatment	100.00%	100.00%	100.00%	100.00%	100.00%
Γ depending on treatment and previous Y	100.00%	100.00%	100.00%	100.00%	100.00%
JM ($\gamma = 3.00, \lambda = 0.05$) with regular visits	99.90%	100.00%	100.00%	100.00%	100.00%

7.5.7 *Results*

Convergence Rates

Convergence rates for each model under each data-generating mechanism are presented in Table 7.1. Most models showed a perfect convergence rate of 100%, except for the joint model (model A). The joint model showed a lower convergence rate of 96% and 99% in two simulated scenarios, both with an informative observation process. However, the remaining scenarios showed perfect convergence rates for the joint model as well.

Results for Regression Coefficients

Bias of regression coefficients is presented in Figure 7.2, while coverage probability is presented in Figure 7.3; MSEs are included in Figure G.1. Bias, coverage probability, and MSEs for the variance components are included as Figures G.2, G.3, and G.4 in Appendix G.

When the observation process was not informative, all models estimated regression coefficients with null to negligible bias. Coverage probability of the regression coefficients was also optimal, with slight under coverage for the intercept term α_0 and the treatment effect α_1 for estimates originating from the marginal weighted model. Mean squared errors were similar across the range of scenarios with a non-informative observation process. Bias for the variance of the residual error term was null to negligible as well, with good coverage. Conversely, the variability of the random intercept v was estimated with slight negative bias from all models, with sub-par coverage (between 90% and 95 %); this is expected as all methods use maximum likelihood and not restricted maximum likelihood. Finally, the estimated variance of the random effect linking the two outcomes in the joint model was positively biased with coverage between 74% and 90%; the magnitude of bias decreased as the baseline intensity λ increased.

When the observation process was informative, the models included in this comparison performed quite differently. When generating data from a Γ distribution depending on treatment only, all models were able to estimate the regression coefficients with no bias, optimal coverage probability, and comparable mean squared errors. However, differences were marked in the remaining scenarios. In the scenario with observation times simulated from a Γ distribution depending on treatment and previous values of the longitudinal outcome, all models but model B (adjusting for the number of measurements) could estimate the treatment effect with null or minimal bias; model B overestimated the treatment effect. The same pattern was observed for coverage of the treatment effect, with model B under-covering (29%), and for the mean squared errors. The effect of time was estimated with small bias and good coverage from all models, with model E (IIVW model) performing slightly worse; mean squared errors were comparable. In scenarios simulated from a joint model, as expected, the joint model (model A) performed best overall, with minimal to no bias, optimal coverage, and the lowest mean squared errors. Model C (adjusting for the cumulative number of measurements) and model D (standard mixed model) overestimated the intercept term and underestimated the treatment effect while showing small bias when estimating the effect of time. Both models showed that the bias when estimating the effect of time decreased as the baseline intensity λ increased: as expected, the inclusion of more measurements allows to better estimate the effect of time. Model B performed worst when estimating the effect of treatment, with large negative bias. It also yielded biased intercept and effect of time, however, as with model C and D, bias for the estimate of time decreased as more measurements were available. Finally, model E slightly overestimated the effect of treatment. Model E showed increasing bias when estimating the intercept as the visiting process was denser, while (analogously as with model B, C, D) showed less biased estimates of the effect of time as the baseline intensity increased. All models with the largest biases showed also poor coverage and the largest standard errors. Overall, in settings simulated from a joint model, model B and model E performed worse and showed the largest biases. In the scenario simulated from a joint model with a sparse observation process and regular yearly visits, the joint model (model A) and the standard mixed model (model D) performed best, managing to recover the true values of all regression coefficients with no bias, and optimal coverage probabilities and mean squared errors. Model B managed to estimate the effect of time with small bias, but largely overestimated the intercept and underestimated the treatment effect. Model C managed to estimate the intercept and the treatment effect with small or no bias, but severely underestimated the effect of time. Coverage and mean squared errors followed the same pattern.

Results for the Association Parameter γ

Bias, coverage probability, and MSEs of the association parameter γ are presented in Figure 7.4.

The estimating procedure worked well when the two sub-models were not associated, with no bias, optimal coverage probabilities, and small mean squared errors - irrespectively of the baseline intensity of visit λ. Conversely, when the sub models were associated ($\gamma = 1.50$) the estimated association parameter was slightly negatively biased (-0.11 to -0.06), with sub-optimal coverage (75% to 83%). Mean squared error decreased when the baseline intensity of visit increased. Finally, the scenario simulated from a joint model with a strong association parameter $\gamma = 3.00$ and regular visits showed the worst performance, with large negative bias (-3.73), poor coverage (12%), and large mean squared error. Including regular visits caused γ to shrink towards the null, with a median estimate of -0.7289.

7.6 APPLICATION TO PSP-CKD DATA

The results of the Monte Carlo simulation study of Section 7.5 are illustrated in practice using data extracted from the PSP-CKD study [48]. In particular, I will be using the second dataset that was described in Section 3.2; it consists of 187,671 observations for 35,822 individuals, over approximately 3 years of follow-up since each practice was randomised to either regular or enhanced CKD care (the latter being the intervention studied with

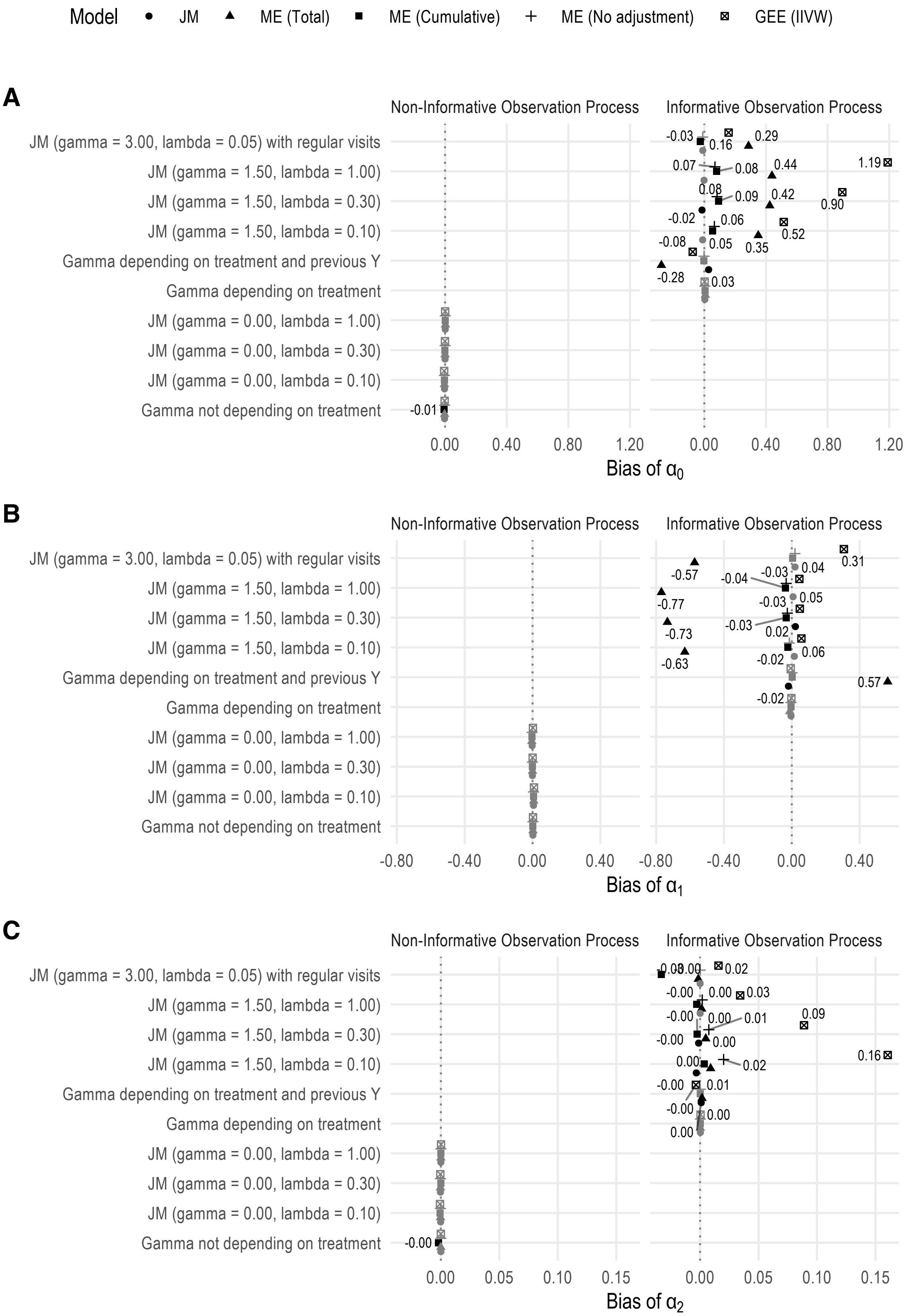


Figure 7.2: Bias of regression coefficients, Monte Carlo simulation study on modelling the observation process. Labelled values (with points in black) are statistically significant values, determined via Z tests based on Monte Carlo standard errors

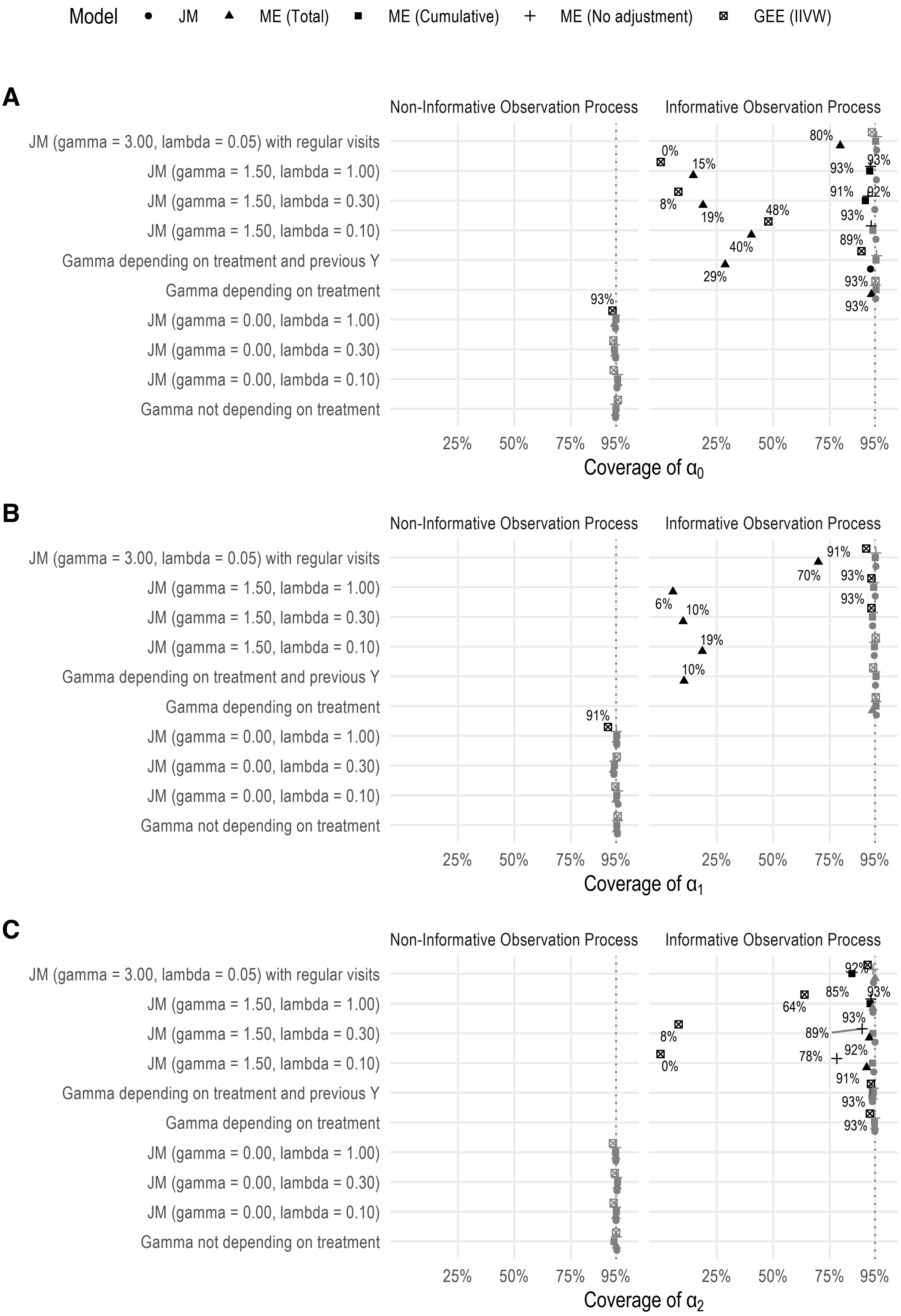


FIGURE 7.3: Coverage probability of regression coefficients, Monte Carlo simulation study on modelling the observation process. Labelled values (with points in black) are statistically significant values, determined via Z tests based on Monte Carlo standard errors

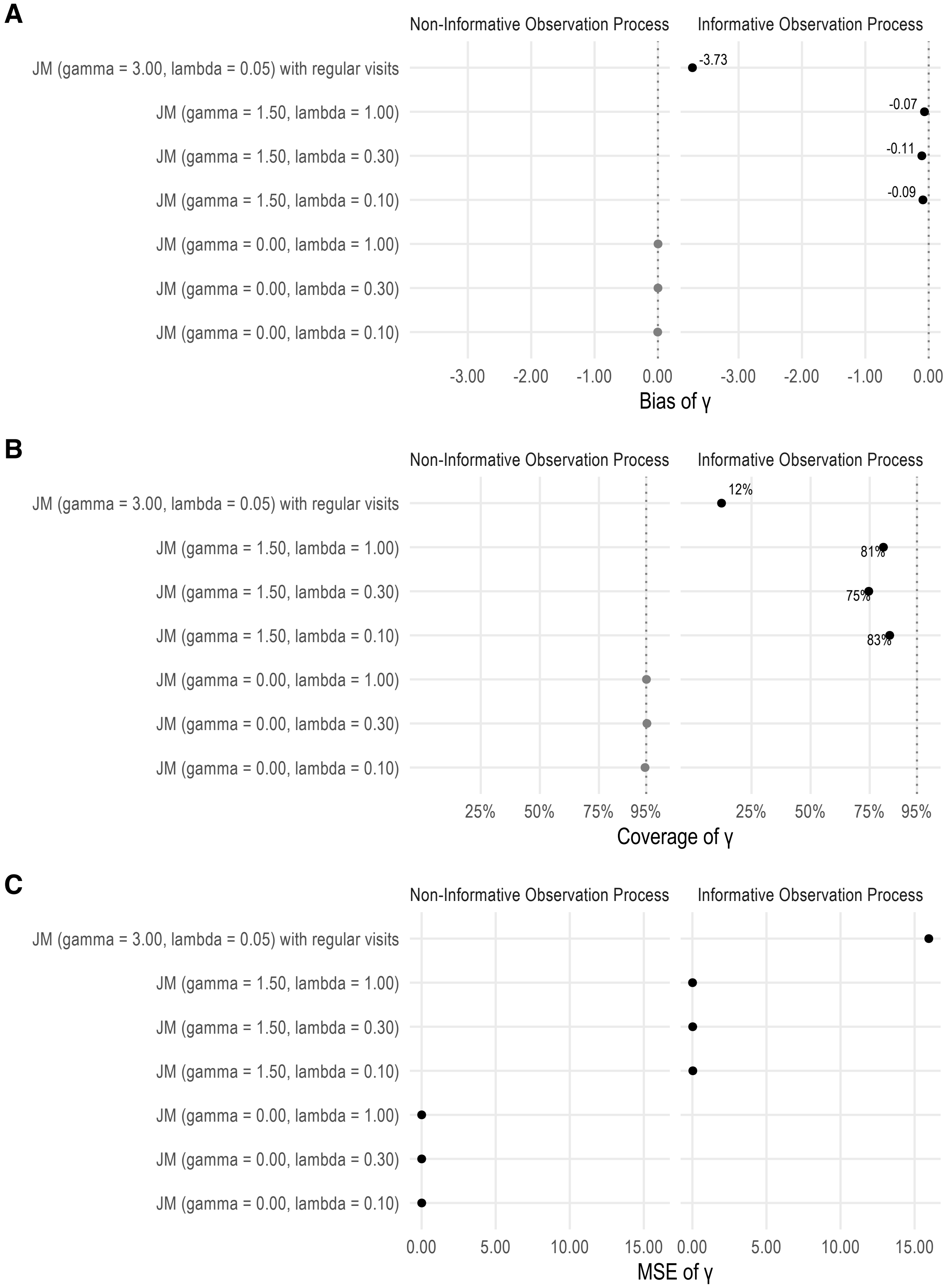


FIGURE 7.4: Bias, coverage probability, and MSEs of association parameter γ, Monte Carlo simulation study on modelling the observation process. Only results from the joint model are included, as it is the only model that estimates the association parameter γ. Labelled values (with points in black) are statistically significant values, determined via Z tests based on Monte Carlo standard errors

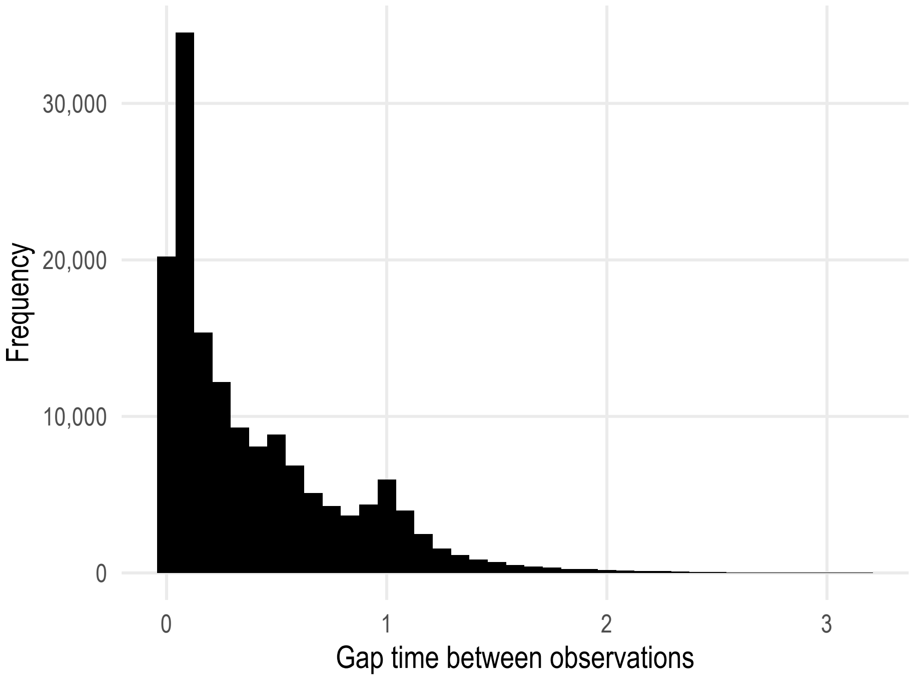


Figure 7.5: Distribution of gap times between longitudinal observations, application to PSP-CKD data.

PSP-CKD).

I start by evaluating whether the visiting process could be deemed informative. First, the median gap time between observations was 91 days (interquartile interval: 28 – 221 days); the distribution of gap times is depicted in Figure 7.5. Next, the Spearman's rank correlations between gap time and treatment ($\rho = -0.01$), gap time and age at baseline ($\rho = 0.04$), gap time and sex ($\rho = 0.01$) are statistically significant (all p-values ≤ 0.001), despite the small magnitude. Then, fitting a linear mixed model for gap time versus treatment, age at baseline and gender with a random intercept yielded significant associations. On average: females had 12.23-days longer gap times (95% C.I.: 9.44 – 15.02), treated individuals had 3.41-days shorter gap-times (95% C.I.: 0.68 – 6.14), and each 5-years difference in age at baseline was associated with 2.74-days shorter gap times (95% C.I.: 2.16 – 3.32). Finally, fitting the Andersen-Gill model for the observation process (as described above) with gender, age at baseline, and treatment as covariates included in the model yielded hazard ratios of 0.949 (95% C.I.: 0.926 – 0.973), 0.996 (95% C.I.: 0.995 – 0.997), and 1.058 (95% C.I.: 1.032 – 1.084), respectively. In conclusion, gap times seem to be associated with gender, age at baseline, and treatment modality; hence, the observation process is likely to be informative.

I fit the models included in the simulation study (and described in Section 7.5.4), including treatment as the exposure of interest and age at baseline and gender as covariates; I assume a linear effect of time for simplicity (and consistency with the simulation study). The joint model includes treatment, gender, and age at baseline as covariates included in the sub-model for the observation process, and so did the recurrent events model utilised to fit weights for the IIVW model. The estimated fixed effects for the longitudinal trajectory from each model are presented in Figure 7.6.

All five models estimated a similar, non-statistically significant effect of treatment at baseline, after adjusting for the remaining factors; the effect of time was negative, which is expected as kidney function generally declines over time. The magnitude of the interaction between time and treatment was negligible and non-statistically significant, showing that enhanced CKD care did not significantly alter the longitudinal eGFR trajectory compared to ordinary CKD care. In other words, the longitudinal decline of eGFR did not differ significantly between treatment arms.

The marginal model estimated an intercept, effect of gender, and effect of time noticeably different than the other four models; furthermore, the direction of the estimated coefficient for the interaction between time and treatment was reversed for the marginal model compared to the other models, although (as above-mentioned) the magnitude of the interaction remained negligible.

The estimated effect of time was similar between all models (approximately -1.10 per unit of time), except for the marginal model; predicted longitudinal trajectories based on the fixed effects (and for females with 75 years of age) are depicted in Figure 7.7. Overall all models estimated a similar longitudinal trajectory, with the exception of the IIVW model and the mixed-effects model adjusting for the total number of measurements; this result is consistent with the results of the simulations of Section 7.5: the mixed model adjusting for the total number of measurements performed worst overall, while the IIVW model yielded biased results for the exposure and the intercept of the longitudinal model under a variety of scenarios. This difference can be observed in these applied setting as well.

The difference between methods can be further appreciated when comparing the estimated difference between treatment arms, as depicted in Figure 7.8. The mixed model adjusting for the centred number of measurements over-estimates the difference

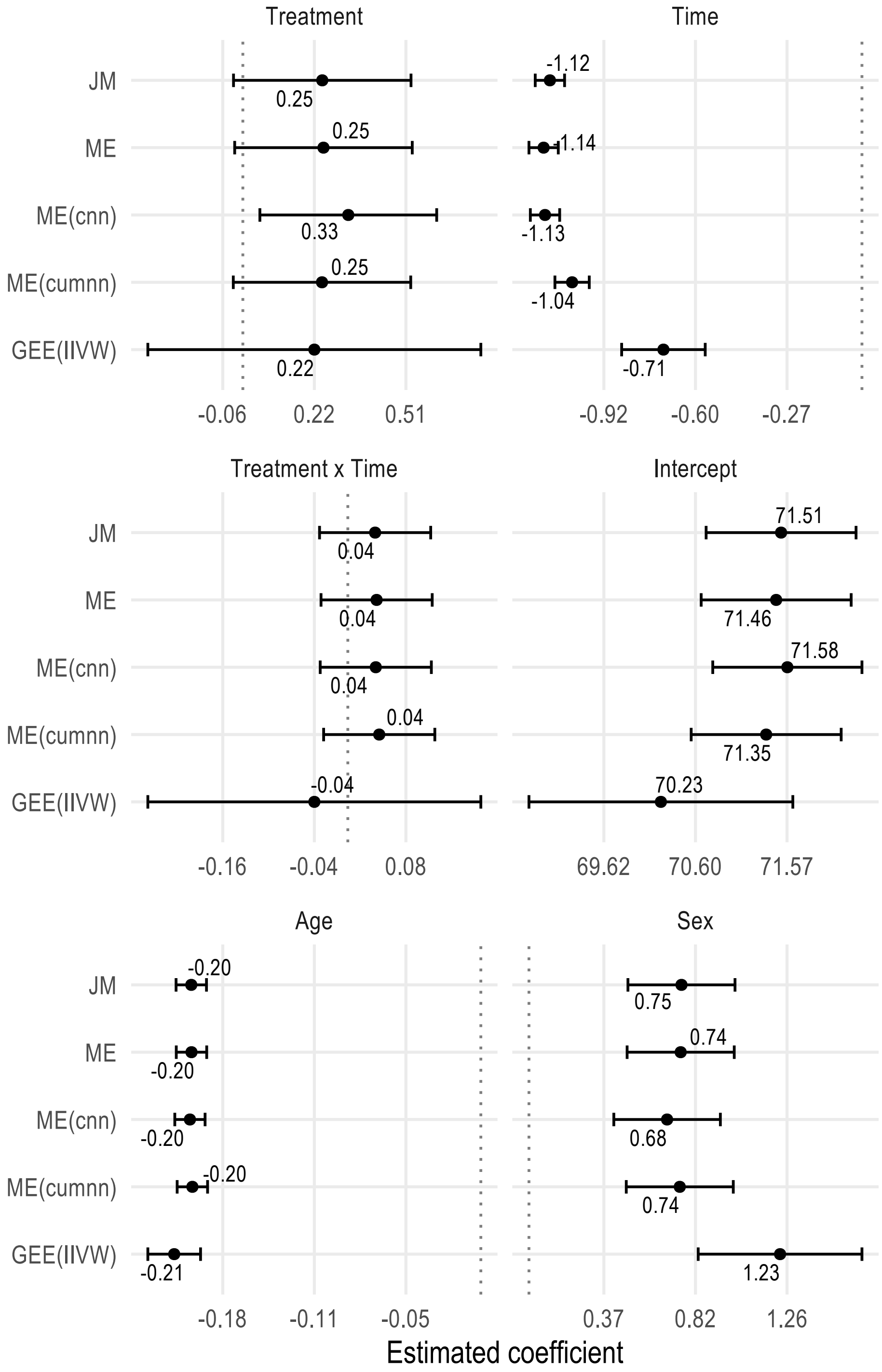


FIGURE 7.6: Estimated fixed effects for the longitudinal trajectory, application to PSP-CKD data; the vertical grey dotted line is placed at a value of zero

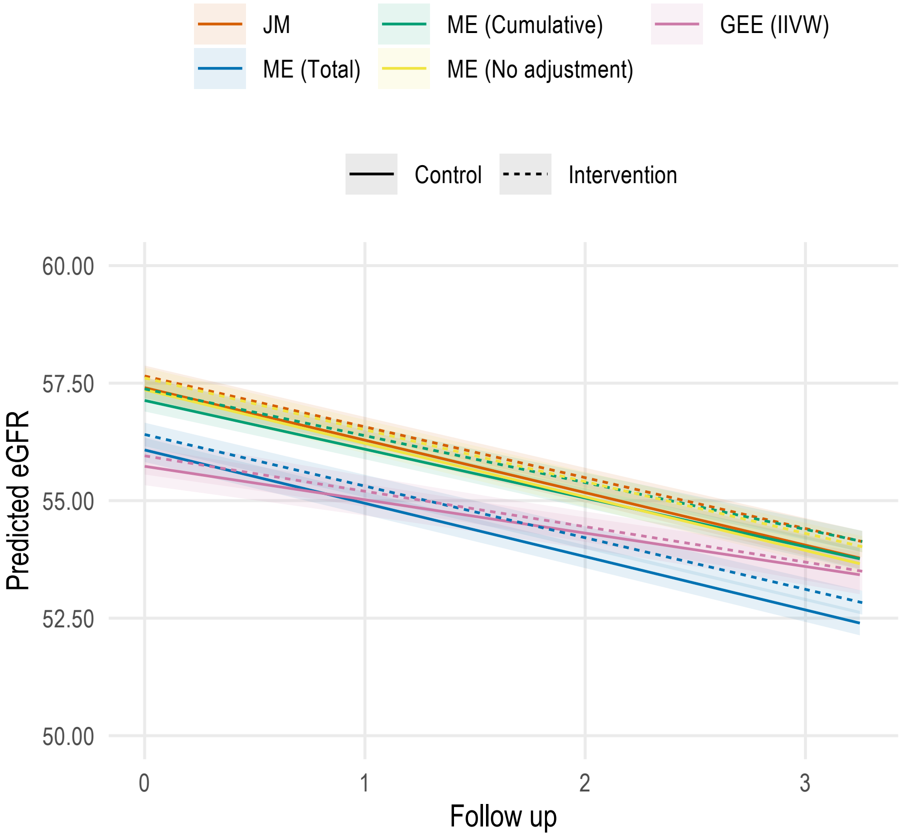


FIGURE 7.7: Longitudinal trajectories by treatment arm based on the fixed effects, application to PSP-CKD data. The predicted trajectories assume 75 years of age, female gender, and an average centred number of measurement or total number of measurements (when relevant)

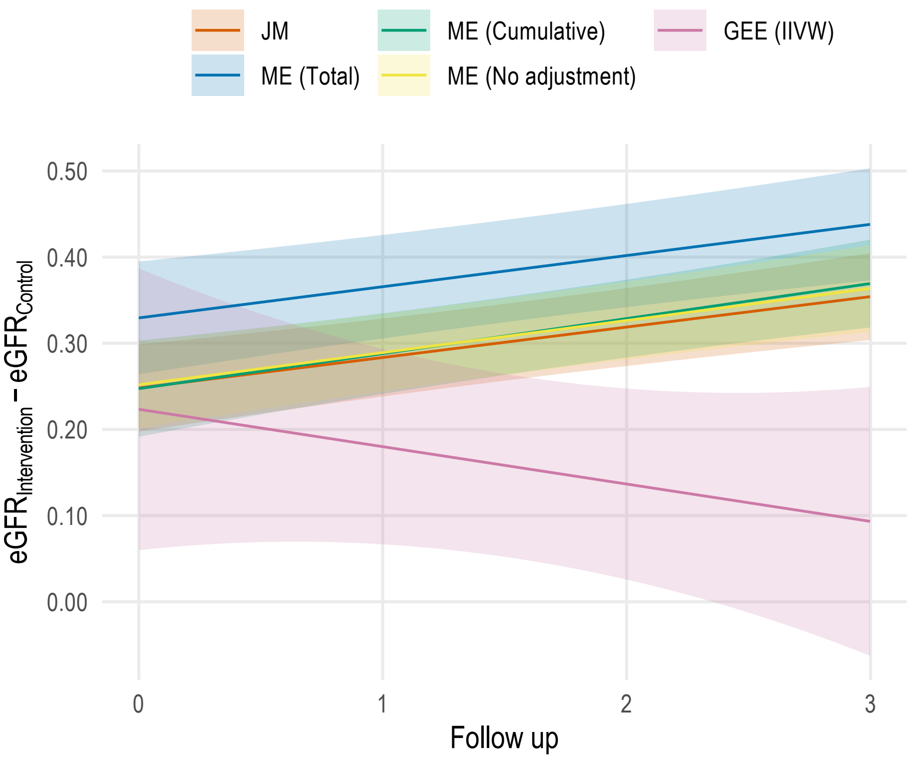


FIGURE 7.8: Difference between treatment arms over time, application to PSP-CKD data

between treatment arms compared to the joint model and the remaining mixed-effects models; despite that, the direction seems to be consistent, with an increasing difference between treatment arms as time goes by. Conversely, the difference between trajectories estimated by the marginal model has the opposite direction, decreasing rather than increasing over time.

The estimated coefficient for the observation process from the joint model shows a reduced risk of having a new observation for females compared to males (approximately 8%, hazard ratio of 0.920 with 95% C.I.: 0.902 − 0.938), an increased risk for treated individuals (7%, hazard ratio of 1.069 with 95% C.I.: 1.048 − 1.090), and a reduced risk for higher age at baseline (hazard ratio for a 5-years increase: 0.993, 95% C.I.: 0.989 − 0.997). This estimated hazard ratios together with the estimated value of the association parameter γ (-2.682, 95% CI: -2.900 to -2.468) seem to confirm that the observation process is non-random in these settings.

The joint model used so far can be further extended: in particular, I will describe how additional random effects and flexibility in modelling the longitudinal trajectory can be easily incorporated within the framework outlined in Section 7.4.

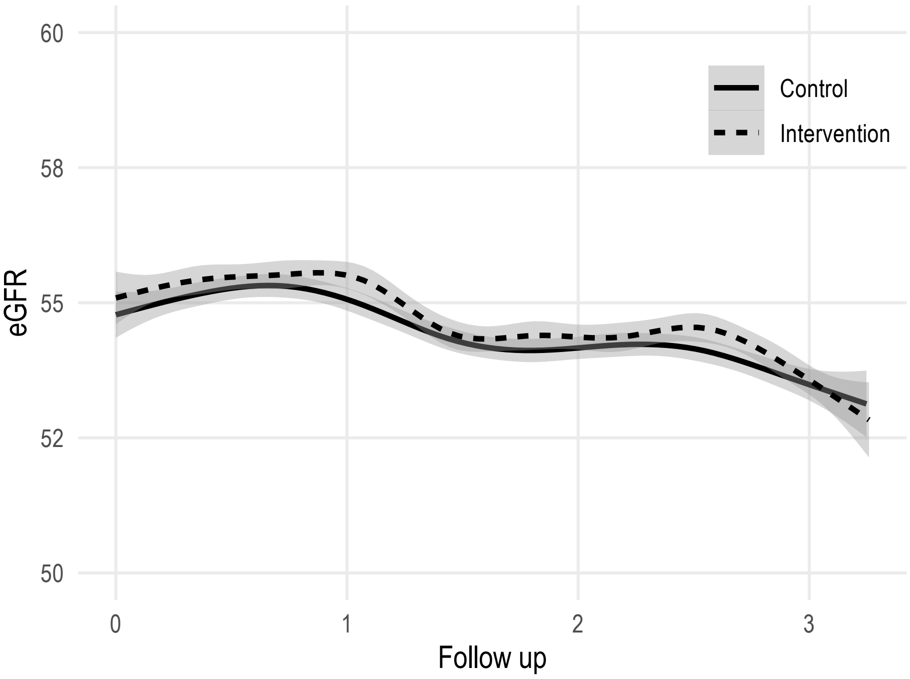


Figure 7.9: Observed eGFR trajectories, application to PSP-CKD data. Smoothed trajectories were obtained by fitting a generalised additive model with a penalised cubic spline smoother, as implemented in mgcv::gam with bs = "cs"

First, I include a random slope of time in the joint model, allowing for correlation between the random intercept and slope; despite that, both random effects are independent with the random effect linking the sub-model for the longitudinal outcome and the sub-model for the observation process. As described in previous Chapters of this Thesis, including a random slope as well as a random intercept allows modelling a second source of between-subjects heterogeneity.

Second, I relax the assumption of linearity for the fixed effect of time by using a restricted cubic spline with 4 degrees of freedom: in fact, the raw trajectories of eGFR by treatment arm, smoothed by fitting a generalised additive model with a penalised cubic spline (as implemented in `mgcv::gam` with `bs = "cs"`), show a certain degree of nonlinearity (Figure 7.9).

Finally, I fit a joint model that incorporates both a random slope of time and a spline function for the fixed effect of time.

The resulting predicted longitudinal trajectories for an individual with given characteristics (75 years old and female, analogously as before) from each of the three above-mentioned joint models, including the original joint model for comparison

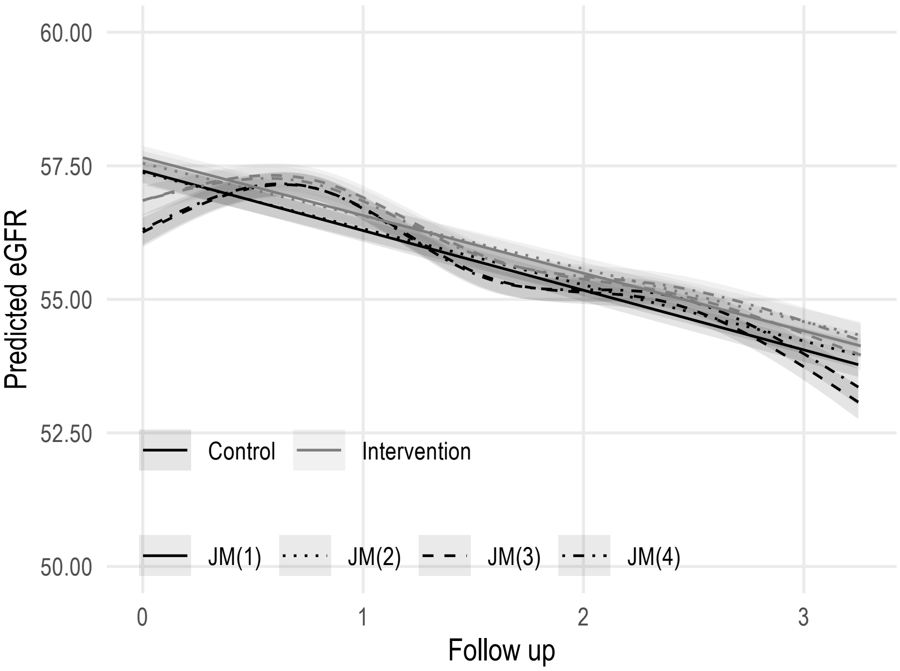


Figure 7.10: Longitudinal trajectories by treatment arm based on the fixed effects, application to PSP-CKD data and comparison of competing joint models. JM(1) is the original joint model, JM(2) is the original joint model plus a random slope of time, JM(3) is the original joint model with the fixed effect of time modelled via a restricted cubic spline with 4 degrees of freedom, and JM(4) is the original joint model plus the random slope of time and the splined fixed effect of time. The predicted trajectories assume 75 years of age, female gender, and an average centred number of measurement or total number of measurements (when relevant)

purposes, are plotted in Figure 7.10. The difference between treatment arms is depicted in Figure 7.11.

AIC and BIC of each of the four joint models have been calculated, and are included in Table 7.2: the best fitting model seems to be the most complex one, the joint model with a random effect of time and the fixed effect of time modelled using a restricted cubic spline with 4 degrees of freedom. In fact, the best fitting model yields predicted longitudinal trajectories that match more closely the observed trajectories (despite being constantly and slightly higher), as depicted in Figure 7.12. This result highlights once again the importance of appropriately modelling the longitudinal trajectory, a common issue in the analysis of longitudinal data irrespectively of the method being used.

7.7 Discussion

Throughout this Chapter, I discussed the issue of informative observation times and I introduced methods that have been proposed in the literature to account for it. Further

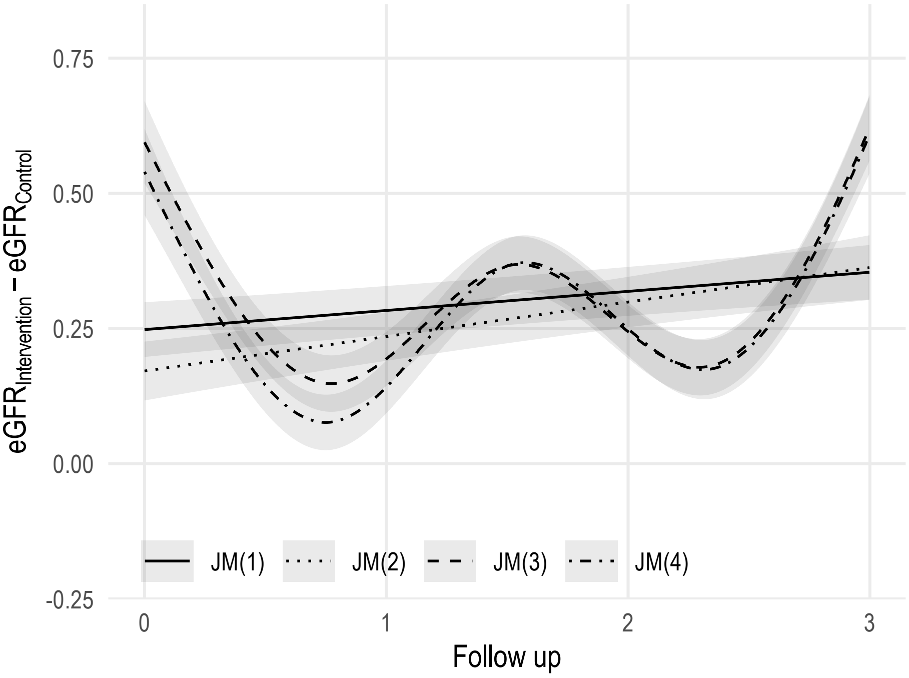


Figure 7.11: Difference between treatment arms over time, application to PSP-CKD data and comparison of competing joint models. JM(1) is the original joint model, JM(2) is the original joint model plus a random slope of time, JM(3) is the original joint model with the fixed effect of time modelled via a restricted cubic spline with 4 degrees of freedom, and JM(4) is the original joint model plus the random slope of time and the splined fixed effect of time

Table 7.2: Joint models comparison in terms of AIC and BIC, application to PSP-CKD data

Model	AIC	BIC
Original JM	1,416,709	1,416,862
Original JM + random slope	1,404,603	1,404,775
Original JM + splined effect of time	1,416,386	1,416,599
Original JM + random slope + splined effect of time	1,404,276	1,404,510

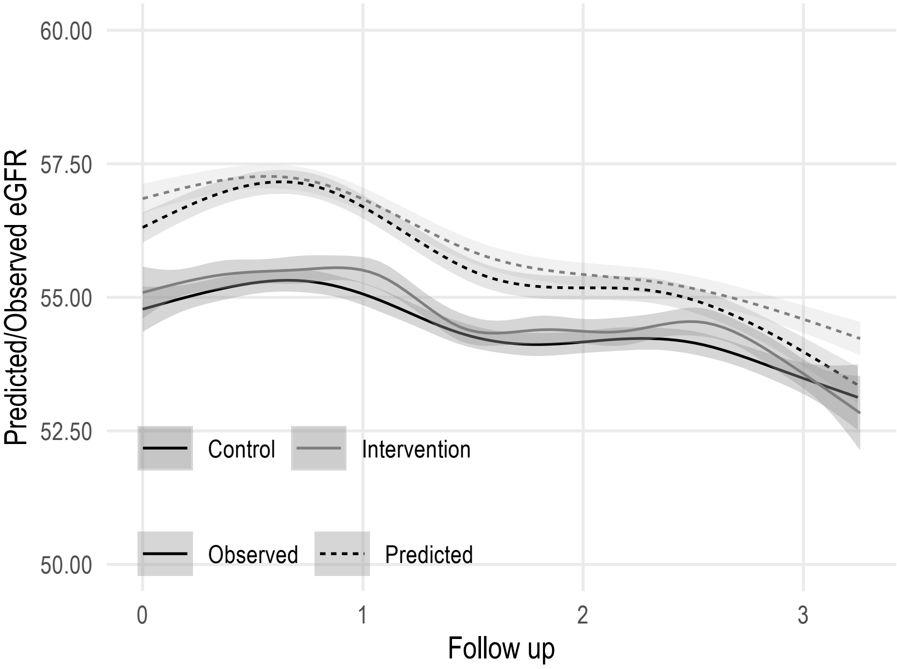


FIGURE 7.12: Observed eGFR trajectories by treatment arm versus predicted trajectories using the best fitting joint model, application to PSP-CKD data

to that, I introduced a joint modelling approach that can be formalised within a larger mixed-effects modelling framework [216]. Then, I compared some of the methods that have been previously suggested, the joint modelling approach, and the standard linear mixed-effects model (that completely disregards the observation process) via Monte Carlo simulation. I generated longitudinal data with an informative observation schedule by using three distinct approaches:

1. The observation and longitudinal processes were simulated jointly from a given joint model;
2. The observation process was simulated from a Gamma distribution with parameters depending on the characteristics of each study subject;
3. The observation and longitudinal processes were simulated jointly from a given joint model first, with pre-planned observations (once every year) added later on.

Then, I ran the simulation study by fitting the above-mentioned models to each of the simulated scenarios, with 1,000 replications per scenario.

The results of the Monte Carlo simulation study of Section 7.5 show that ignoring an informative visiting process leads to biased estimates of the regression coefficient of a longitudinal model; further to that, not all models included in this comparison performed equally.

To the best of my knowledge, the issue of informative observation times did not receive as much attention as the issue of informative drop-out. For instance, there is only another comparison currently published in the literature [219], albeit I include a different set of models in this simulation study and I simulate the observation process in continuous time, while Neuhaus *et al.* simulate an informative observation process by first generating a grid of potential observation times and then relating the probability of being observed to a given functional form of current (or lagged) covariates. Notably, the joint modelling approach that I described in Section 7.4 is not considered by Neuhaus *et al.* [219] and had therefore never been compared to other methods before.

As expected, the joint model that accounts for the informative observation process by modelling it via a recurrent events survival model performed best - especially in the scenarios simulated from a joint model. An interesting point is that the mixed-effects model that disregarded completely the observation process performed worse than the joint model but outperformed other methods; the inflation in the variance of the random intercept of the standard mixed model seemed to capture part (if not most) of the variability due to the observation process, although this result needs to be thoroughly tested in more complex scenarios (e.g. with random effects of time, etc.). The performance of the standard mixed model confirms the results of Neuhaus *et al.* [219].

The mixed models adjusting for the total number of measurements or the cumulative number of measurements (as a time-varying covariate) performed worst, and I would not recommend their usage in practice. This finding contrasts the findings of Goldstein *et al.*, although their settings are quite different than those of this simulation study [220]. Further to that, they acknowledged the potential for collider bias (due to conditioning on a collider, the number of measurements) when the phenotyping algorithm for determining the exposure has high sensitivity; indeed, in my simulations the sensitivity is perfect as there is no misspecification of the exposure. An additional possible explanation could be that in my settings the model adjusting for the total number of measurements is in fact conditioning on the future, as the total number of observations is not determined at the beginning of the study. This may be explaining the poor performance of this method in the simulations of Section 7.5.

The marginal model fitted using generalised estimating equations and inverse intensity of visit weights performed in-between the standard mixed model and the other mixed models; furthermore, its performance seemed to improve when the observation pattern became denser, except for the intercept term α_0. This pattern was generally observed throughout all scenarios and models, as the performance seemed to increase with more frequent observation patterns; this finding is consistent with other results published in the literature [203] and with those of Neuhaus *et al.* [219]: the IIVW approach showed bias in all scenarios of their simulation where the observation process was informative, even when adding regular visits to the study. To compute the weights of the IIVW approach, applied researchers need to correctly specify the model for the visit process, a challenging task - especially when not all the information required to fit the correctly specified model is observed (or known).

Most importantly, a key finding of this simulation study is that under the null all the approaches compared in this study produce unbiased estimates of the regression coefficients, the implication being that over modelling the observation process does not seem to introduce bias in the analysis. In settings where it is not clear whether the observation process is informative or not, fitting the joint model of Section 7.4 would provide applied researchers with a method for estimating (and testing) the association between the two outcomes. This could be especially useful e.g. as a sensitivity analysis of standard mixed-effects models.

The results of the simulation study can be appreciated in practice with the application of 7.6. Interestingly, in the settings of the PSP-CKD study, all models performed somewhat similarly except for the IIVW approach, which yielded completely different fitted trajectories. This highlights once again the importance of (1) choosing an appropriate analysis method (2) sensitivity analyses to evaluate whether the method of choice affects the results of a study. The results of the application are consistent with the results of the PSP-CKD investigators [48] and with the results of the application of Chapter 5: enhanced CKD care did not significantly alter the longitudinal loss of renal function.

The joint model for the observation process and a longitudinal outcome described in Section 7.4 can be further extended, as previously described. For instance, the results of the simulation study of Chapter 5 highlighted the importance of modelling appropriately the baseline hazard: the best fitting joint model from the application could be extended by using flexible parametric models (on the log- or log-cumulative scale) for the observation process to better capture the baseline intensity of the observation process. The baseline hazard for the observation process (estimated using the same smoothing approach described in Chapter 5) shows that a more flexible baseline hazard formulation that could accommodate turning points as well could be more appropriate in the settings of the application (Figure 7.13). Additional random effects could be introduced in the model to account for, say, heterogeneity in the trajectory of the longitudinal outcome over time; in the application, I included a random effect of time and that improved model fit. The functional form of the effect of time (both fixed and random) could also be further generalised by using fractional polynomials or splines; the longitudinal trajectories need to be modelled appropriately and best fit could be assessed via information criteria such as the AIC/BIC or their decomposition into additive components that allow assessing the goodness of fit for each component [197]. In the applied example of Section 7.6, modelling the fixed effect of time using splines vastly improved model fit. Time-varying treatments could also be included in both the observation process and longitudinal outcome sub-models, although the performance of the joint model would need to be assessed in these settings. Finally, the joint model could also be extended by modelling the drop-out process as well, as described in Chapter 6. Most of these extensions (and several others) are discussed in more detail in Chapter 8.

This simulation study has also some limitations. First, I assumed the treatment to be constant over time for simplicity; in real-life settings, however, individuals are likely to start and stop treatment when deemed necessary by their physician. I assumed the baseline hazard of the recurrent events model for the observation process to follow a Weibull distribution: this assumption could be further relaxed, and one could assume any parametric function, or even use flexible, spline-based formulations. Additionally, for diseases with a high mortality rate, a terminal event that truncates observation of the longitudinal process is likely to be informative in the sense that it likely correlates with disease severity (as described in Section 6.6). That is, drop-out is likely to be informative as the tendency to drop out after the occurrence of a terminal event is related to the current level of the longitudinally recorded biomarker. The proposed model could be easily extended to include a third equation with a second survival sub-model for the

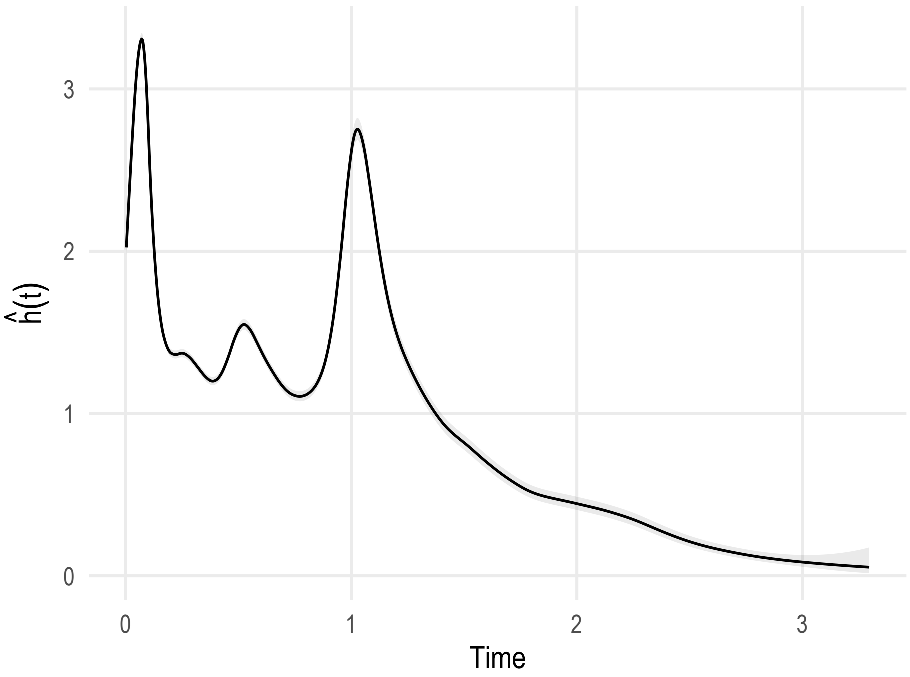


Figure 7.13: Smoothed non-parametric baseline hazard estimate, application to PSP-CKD data

drop-out process, as in the trivariate model of Liu *et al.* [215]. All of these extensions (and more) can be fit within the general framework of `merlin`, as described in Chapter 8 [216, 217]. One could also explore and extend the association structure between the two sub-models. For instance, the association structure could be reversed to include γ in the observation submodel: in that setting, assuming a positive association, higher values of the longitudinal process would lead to a more frequent visiting process (and vice-versa in the setting of negative association). The observation process could also depend on lagged values of the longitudinal outcome or of the exposure; this would relax the semi-Markov assumption of some of the data-generating mechanisms of this simulation study. More biologically (and clinically) plausible association structures (such as those described in Section 6.4) could also be investigated. Finally, taking the pattern of informative observation processes (where healthy individuals are under-represented) to the extreme, healthy individuals may not appear in health records at all, leading to cohort selection bias.

In conclusion, it is important to account for the visiting process when analysing health care utilisation data and I showed that ignoring it leads to biased estimates. Given the wide range of applied settings in which this could be relevant, the review of Farzanfar *et al.* [200] points towards a lack of awareness of the problem and the lack of readily available, user-friendly software to fit more complex joint models. With the extended joint modelling framework described in Section 7.4 I outlined a framework in which `merlin` could be easily used to fit complex joint models to help to reduce this translational gap.

A manuscript based on the content of this Chapter has been published in Statistica Neerlandica, and is included in Appendix F [201].

8 Discussion

8.1 OUTLINE

This Chapter concludes the Thesis by summarising the developments outlined in previous Chapters; in particular, I will describe the main results of each Chapter in Section 8.2. I will also consider the limitations of the work included in this Thesis in Section 8.3, and potential extensions and future developments in Section 8.4. Finally, I will provide some closing remarks in Section 8.5.

8.2 SUMMARY OF THE THESIS

In this Thesis, I investigated statistical methods that can account for the multilevel structure commonly encountered in electronic health records (EHRs); the methods have been studied through simulation and applications to real-world data in the settings of chronic kidney disease and intensive care medicine.

EHRs, their characteristics, and their use for research purposes have been described in Chapter 1. Potentials and pitfalls of using EHRs in research have been described as well: in particular, I focussed on the potential to answer innovative and more detailed clinical questions, and on the opportunity to study interventions (such as novel medications or medical devices) in real-life settings. Among the pitfalls, notably, I mentioned the requirement of ad-hoc methodologies that allow accommodating the characteristics of EHRs and avoid biases that would otherwise arise. Traditional statistical methods need to be thoroughly studied and tested in the settings of EHRs to assess whether underlying assumptions are met and whether their use is possible with EHR data.

Chapter 2 follows, where foundations of standard survival analysis and longitudinal data analysis are introduced. Fundamental relationships and terminology in survival analysis are described, with additional details on non-parametric methods and regression models (parametric, semi-parametric, and flexible parametric). Analogously, the notation for longitudinal data is introduced alongside the most common regression-based methods for the analysis of longitudinal data: marginal models estimated using generalised estimating equations (GEE), and mixed-effects models.

In Chapter 3 I introduced two motivating examples that I used throughout the Thesis: the PSP-CKD study and the VASST trial [48, 55]. The PSP-CKD study is a cluster randomised controlled pragmatic trial comparing enhanced CKD care against routine care in Northamptonshire, UK. PSP-CKD investigators concluded that the intervention (i.e. enhanced CKD care) did not affect the rate of renal function decline, but it did lead to significant improvements in processes and quality of care [48]. Data extracted from the PSP-CKD study has been used in the applied examples of Chapters 5 and 7, with results closely matching those of the PSP-CKD investigators. Conversely, VASST is a randomised controlled trial that compared vasopressin versus norepinephrine in patients with septic shock; the results of the trial were negative, showing that administering vasopressin did not reduce mortality rates [55]. One of the issues when analysing longitudinal data from trials in the settings of intensive care medicine is drop-out that truncates the longitudinal trajectories. Therefore, data extracted from VASST has been used in the applied example of Chapter 6 to illustrate the use of joint modelling to account for non-random drop-out in the analysis of longitudinal outcomes.

This Thesis broadly relies on Monte Carlo simulation methods. In Chapter 4, I introduced and described newly developed open-source software in R to aid the analysis of Monte Carlo simulation studies, the `rsimsum` package and the INTEREST Shiny app. A manuscript on `rsimsum` has been published in the Journal of Open Source Software [223], while a manuscript on INTEREST is currently under review in the Journal of Data Science, Statistics, and Visualisation. `rsimsum` supports simulation studies with a single or multiple estimands, with several methods being compared, and with any number of data-generating mechanisms. All performance measures described in Chapter 4 are supported, and their Monte Carlo standard errors are computed and reported by default. `rsimsum` also provides support for a variety of automated and opinionated data-visualisation methods that enable quick explorations of results and that can be

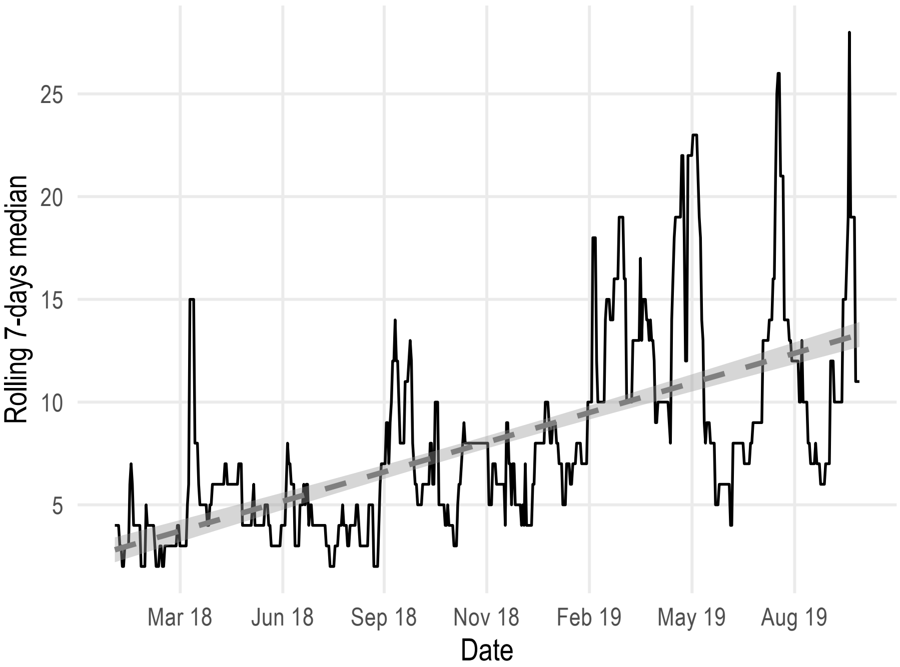


FIGURE 8.1: Rolling 7-days median of `rsimsum` downloads since its publication on CRAN, with super-imposed linear trend

further customised by the user. Both software packages are freely available, and have been presented at national and international conferences; `rsimsum` is also published on the Comprehensive R Archive Network (CRAN), and it has been downloaded more than 5,000 times since its availability with an increasing trend (Figures 8.1 and 8.2). Furthermore, `rsimsum` and INTEREST fit well within the ADEMP structured approach introduced by Morris, White, and Crowther [58] and also described in Chapter 4: ADEMP argues that by carefully designing, describing, and reporting Monte Carlo simulation studies clarity and reproducibility could be greatly improved. `rsimsum` assists with the analysis of simulation studies, relaxing the requirement of ad-hoc (and error-prone) code to compute all performance measures (and Monte Carlo errors) of interest by automating the whole procedure. INTEREST supplements the reporting and reproducibility of simulation studies by adding a layer of interactivity; this feature is extremely important, especially in simulation studies with several methods and simulated scenarios where dissemination of results is challenging at best. Chapter 4 also described methodologies to simulate biologically plausible, complex survival data; such methods are used extensively in Chapters 5 and 7.

In Chapter 5, I introduced multilevel survival data and described analysis methods that can take into account the hierarchical structure. In particular, I described in Section 5.3

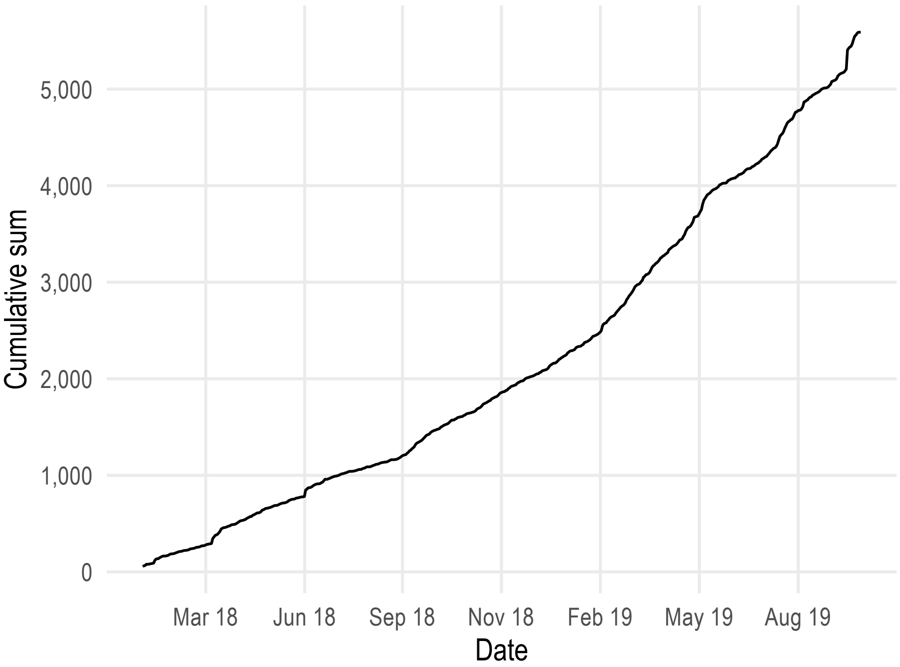


Figure 8.2: Cumulative sum of `rsimsum` downloads since its publication on CRAN

methods that are commonly used to analyse recurrent events data (such as infections, admissions to the hospital, etc.). Next, I introduced the frailty approach; I focussed on this approach as it ties naturally with the joint modelling framework discussed in later Chapters, and because it can easily accommodate different clustering structures. I also described how the frailty approach could be extended to accommodate multiple levels of the hierarchical structure, such as patients nested within hospitals nested within regions. One of the main challenges in regression modelling consists of choosing the correct model specification; otherwise, biases and/or inefficiencies may arise. The topic of model misspecification in shared frailty survival models has been tackled by mostly focussing on the distribution of the frailty: the literature on misspecification of both the baseline hazard and the distribution of the frailty is somewhat scarce. In Chapter 5, I designed and conducted the most extensive simulation study (to the best of my knowledge) on the topic; this study was published in Statistics in Medicine [93]. I studied misspecification of the baseline hazard, misspecification of the frailty distribution, or both. 90 distinct simulation scenarios were included, and several model formulations were explored: parametric, semi-parametric, or flexible parametric baseline hazard functions (including penalised approaches), and either Gamma or log-Normal frailty distributions. The results of these simulations highlight the importance of properly modelling both the baseline hazard and the frailty distribution, depending on the aims of the analysis. For instance, when the interest is on measures of relative risk, it is most important to properly model the baseline hazard; regression coefficients were confirmed to be robust to misspecification of the frailty distribution, with a very marginal impact on relative risk estimates. Conversely, measures of absolute risk (e.g. the loss in life expectancy) and measures of heterogeneity (e.g. the estimated variance of the frailty) were affected by misspecification of both the baseline hazard and the frailty distribution. The advantages of using flexible parametric models with a shared frailty term (as implemented in the R package `rstpm2`, for instance [132]) resulted clear, given their overall robustness (e.g. to the number of degrees of freedom used to model the baseline hazard) and the ease of extending standard model formulations (e.g. with time-dependent effects). Semi-parametric models retain some of the advantages of flexible parametric models when estimating measures of relative risk; however, obtaining absolute risk estimates is noticeably easier with flexible parametric models, aiding with the translation of relative risk measures to an absolute scale.

Chapter 6 introduces the topic of joint modelling of longitudinal and survival data. This class of models is particularly interesting in the settings of EHRs, given that repeated longitudinal measurements and time to event outcomes are often recorded in EHRs. I introduced the standard joint model formulation and described the estimation procedure (including computational challenges - some of which are shared with frailty models as described in Section 5.4.2) and several possible association structures linking the longitudinal and the survival components of the joint model. Joint longitudinal-survival models have traditionally focussed on the survival sub-model; however, the main focus of this Thesis was on the longitudinal component. When analysing longitudinal data, one of the underlying assumptions is that the longitudinal outcome and drop-out from the study (e.g. truncation of the longitudinal trajectory) are independent. This may not always be the case with EHRs data: individuals with abnormal values of the longitudinal outcome may be at higher risk of dropping out of the study, e.g. because of death. I described this issue in more detail in Section 6.6. Joint models for longitudinal and survival data can accommodate this scenario by jointly modelling the longitudinal outcome and the drop-out process: the two sub-models account for each other in the analysis and yield unbiased estimated coefficients for the longitudinal trajectory. As an illustration, the analysis of longitudinal organ failure (SOFA) scores from VASST - while accounting for the fact that individuals with higher SOFA scores are at much higher risk of death -

highlights how joint modelling could be a useful methodology to apply when drop-out is believed to be non-random. The results of this application have been submitted for publication in Critical Care Medicine, including a tutorial on the use of joint modelling in the settings of intensive care medicine; the manuscript is currently under review.

Finally, in Chapter 7 I investigate the violation of another common assumption when modelling longitudinal data, the assumption of independence between values of the longitudinal outcome and the timing between measurements. Compared to the problem of informative drop-out, the problem of informative observation times has received considerably less attention in the current literature. Nevertheless, in the settings of EHRs, this assumption is often violated: for instance, individuals with worse disease status are likely to visit their doctor more often compared to healthy individuals. I discuss the problem of informative observation in more detail in Section 7.2, describing characteristics of the observation process and formalising the observation process using counting process notation. Several methods have been proposed in the literature to account for informative observation times, methods that can be broadly categorised in methods based on inverse probability weighting and methods based on joint modelling. However, their comparative performance is unclear: at the time of writing, I could only find one simulation study in the literature, with results showing that completely ignoring the observation process yielded the smallest bias [219]. In Chapter 7, I first formalised a joint modelling approach that was previously proposed by Liu *et al.* [215] within a multivariate mixed-effects modelling framework that easily allows incorporating several extensions [216, 217]. The joint model assumes a mixed-effects model for the longitudinal trajectory, and a frailty model for the observation process; the two components are then linked via a shared random effect that captures the association between the two processes. Notably, this joint model can be fit using readily available statistical software. Then, I designed and conducted a Monte Carlo simulation study that compared the aforementioned joint modelling approach, the approach based on inverse probability weighting, the approach that completely ignores the observation process, and two pragmatic approaches that had been proposed and based on adjusting the standard mixed model for either the total or the cumulative number of observations per individual. The results of my simulations show how the joint modelling approach performed best, managing to estimate the coefficients of the longitudinal model with the smallest bias while at the same time allowing to do inference on the observation process. In fact, another advantage of the joint modelling approach is that it returns an estimated value of the association between the observation process and the longitudinal outcome. One of the most interesting results of my simulations in Chapter 7 is that the joint model does not induce a spurious association when the two processes are independent: thus, the joint modelling approach could be useful whenever it is not clear whether the observation process is informative or not, e.g. as a sensitivity analysis. A manuscript with the results of the work described in Chapter 7 has been published in Statistica Neerlandica [201].

8.3 LIMITATIONS

The methodological developments introduced throughout this Thesis and summarised in the previous Section present some limitations.

First, the development of the R package `rsimsum` and the INTEREST Shiny app (described in Chapter 4) is not concluded. New functionalities are to be added, and both packages could be improved in terms of computational speed, efficiency, and robustness. Further to that, it is not possible to exclude the presence of bugs that have not been detected yet.

The results of the Monte Carlo simulation study of Chapter 5 could be further generalised by adding more data-generating scenarios: for instance, one could include more sample size scenarios (in terms of individuals and clusters) and vary whether clusters are balanced (in terms of subjects per cluster) or not. One could also explore more complex model formulations (with multiple covariates), additional frailty distributions (such as the positive stable and inverse Gaussian distribution, although I included the two most commonly used frailty distributions in practice), and different censoring mechanisms or delayed entry. Nevertheless, the number of simulated scenarios (90 distinct scenarios) and the variety of methods being compared yields - to the best of my knowledge - the most extensive simulation study on the topic to date. Further to that, all methods use maximum likelihood which returns negatively biased estimates of the variance components with bias decreasing as the number of clusters increases. The restricted maximum likelihood method could be used with a small number of clusters to obtain unbiased estimates of the variance components [156]. The results are also highly dependent on the R implementation of each method: as such, they could vary if using a different statistical software package (e.g. Stata). Despite that, the packages that were used are well established and commonly used in practice (Figure 8.3), and I applied these methods as they are intended to be used i.e. without modifying convergence criteria and/or starting values of the estimation procedure; more accurate results could be obtained by requiring more stringent convergence criteria. Interestingly, though, some of the methods compared with this simulation study have been implemented only in R: the flexible parametric models with penalised splines (on either the log-hazard or log-cumulative hazard scale), the Cox model with a log-normal frailty, and the flexible parametric models with a Gamma frailty. This comparison would have not been possible if using a different statistical software package.

The applications of Chapters 5 and 6 relied on the traditional AIC and BIC information criteria to select the model that fitted the data best. Vaida and Blanchard [142] suggested the use of conditional AIC (cAIC) for model selection in linear mixed models, demonstrating that the traditional AIC and its small sample correction are inappropriate when the interest is on clusters; unfortunately, the cAIC is not routinely reported and thus it was not possible to use this criterion. Further to that, in the joint modelling settings, the AIC and BIC do not provide separate assessments of each component of the joint model. Zhang *et al.* [197] developed an additive decomposition of AIC and BIC that enables the assessment of the fit of each component of the joint model separately, and in the settings of Chapter 6, using this decomposition would allow assessing the fit of the longitudinal component (the outcome of interest) regardless of the survival sub-model. Unfortunately, the methodology of Zhang *et al.* is not implemented in most joint modelling software, with code developed using SAS software only and not publicly available.

The Monte Carlo simulation study of Chapter 7 comparing methods for accommodating the observation process in the analysis presents some limitations as well. For instance, the treatment was assumed to be constant over time for simplicity; in real-life settings, however, individuals are likely to be on and off treatment, as deemed necessary. The baseline hazard of the model for the observation process was assumed to follow a Weibull distribution: this assumption could be relaxed by assuming any other fully parametric or flexible formulation. The simulated scenarios of Chapter 7 also assumed that the drop-out

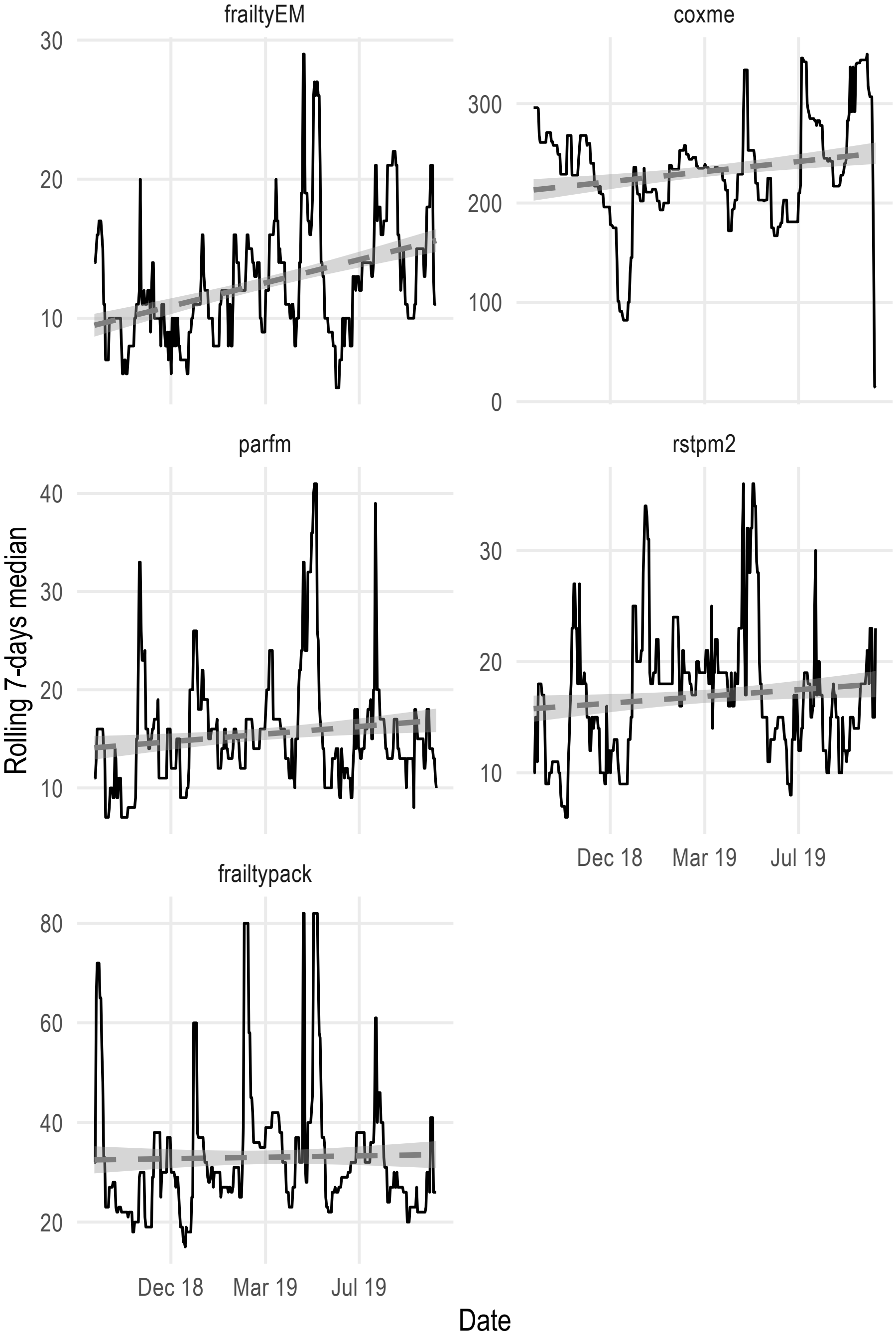


Figure 8.3: Rolling 7-days median of downloads from CRAN for the packages included in the simulation study on model misspecification of shared frailty models during the last 12 months, with super-imposed linear trend

process was not informative (as described in Chapter 6); this assumption is unlikely to hold in the settings of EHRs, especially for diseases with a high mortality rate where a terminal event that truncates observation of the longitudinal process is likely correlated with disease severity. I assumed all of the above for simplicity, and to focus on modelling the observation process; however, the joint modelling framework described in Chapter 7 could easily be extended to accommodate the drop-out process. More details on future work and extensions are included in Section 8.4.

Another assumption of the simulation study of Chapter 7 concerns the association structure of the joint model; specifically, the joint model being investigated assumed that the sub-models are linked through a shared random effect. Of course, several alternative association structures could be explored. For instance, the association structure could be reversed to include the association parameter γ in the observation sub-model, or it could depend on lagged values of the longitudinal outcome or of the exposure. More biologically (and clinically) plausible association structures - such as those described in Section 6.4 - could also be investigated.

All Monte Carlo simulation studies of this Thesis have been run following a fully factorial design. Fully factorial designs for simulation studies affect generalisability of results, as the number of simulated scenarios needs to be constrained to keep the computational cost acceptable. Despite having simulated a large number of scenarios (especially in the simulations of Chapter 5), incomplete designs and meta-modelling could be implemented to further increase the external validity and the ability to generalise the results. These approaches are also described in more detail when discussing future work and extensions of this Thesis in Section 8.4.

Finally, the main limitation of the methods studied throughout this Thesis consists of the computational complexity required to fit them. Shared frailty models (as discussed in Chapter 5) require numerical integration when assuming a numerically intractable frailty distribution, while joint longitudinal-survival models require different layers of numerical integration (for instance, when integrating over the random effects). In particular, the complexity of the numerical integration required to fit joint longitudinal-survival models grows exponentially as the number of random effects and/or outcomes included in the model increases, as discussed in Section 6.5. This issue is further aggravated by the sample size: the thousands (if not millions) of measurements commonly available in the settings of EHRs cause the computational burden to rapidly grow, even when applying traditional statistical methods. Methodological and computational improvements need to be sought to spread the use of more advanced methods in practice. For instance, the use of efficient sampling strategies (e.g. by exploiting the case-cohort study design, as in Baart *et al.* [224]) could be instrumental in reducing the amount of data required for a given analysis, and the development of more efficient algorithms and estimation procedures could yield additional computational benefits. Ultimately, these improvements would greatly contribute to spreading the adoption and use of advanced statistical methods that can accommodate the complexities of EHRs.

8.4 Future Work

The research presented throughout this Thesis focusses mainly on the assessment and application of multilevel modelling techniques in the settings of EHRs. However, there is still much scope for further work as described in the next few Sections.

8.4.1 Improvements of `rsimsum` *and INTEREST*

Although `rsimsum` and INTEREST are fully functional in their current state, several future developments are being planned. First and foremost, both packages could be made more robust and computationally efficient. `rsimsum` requires a few seconds at most to analyse Monte Carlo simulation studies with hundreds of scenarios and thousands of replications per scenario, but every improvement in terms of computational speed and memory usage is always welcome. Then, I aim to include support for multiple estimands at once in INTEREST as currently supported by `rsimsum` via the `multisimsum` function. I also aim to improve the flexibility of `rsimsum` and INTEREST in terms of customisation (of tables and plots), e.g. by displaying the raw R code used to generate the plots behind the scenes; that code could be then used by the user to fully re-build each plot and further modify it if needed. More flexible methods to produce LaTeX tables could also be implemented, allowing users to define the structure of a table and customise it as needed. Finally, additional interactive features could be added to the app via HTML widgets, D^3 [84], or other approaches; several R packages allow incorporating interactive graphs into Shiny apps (e.g. `htmlwidgets`, `plotly`, `r2d3` [225–227]).

8.4.2 Meta-modelling Results of Monte Carlo Simulation Studies

Much of the work of this Thesis has been conducted by utilising Monte Carlo simulation methods. Given the current computational possibilities (e.g. the increasing availability of cluster computing servers) it is tempting to simulate an ever-growing number of scenarios by varying several data-generating mechanisms (DGMs) at once. For instance, in the simulation study of Chapter 5 90 distinct DGMs were simulated; consequently, 22 models were fitted to each DGM. This yielded 1,980 summary data points that need to be purposefully presented to the reader to communicate the full results of the simulation study.

Rücker and Schwarzer [73] introduced the nested loop plot as a data visualisation that could be used to present several data-generating scenarios at once (e.g. 768 in their motivating example) by ordering all scenarios and then arranging them sequentially on the horizontal axis of a plot, with the performance measure (or estimate) of interest on the vertical axis. However, despite the nested loop plot being a promising alternative, several shortcomings still persist: most importantly, it is cumbersome to accommodate the performance measures of several methods in a single nested loop plot, and there is no obvious way to include uncertainty (in the form of Monte Carlo standard errors) within the data visualisation.

Skrondal [228] proposed to *attack the conventional wisdom* by specifying a meta-model when analysing the output of a Monte Carlo simulation study, by exploiting incomplete factorial designs, and by using variance reduction techniques. These techniques jointly combined could allow, according to Skrondal, investigating an increased number of experimental scenarios and improving external validity at the cost of the conventionally excessive precision. However, despite being introduced almost 20 years ago, this approach had not been widely adopted (if at all) in practice; interestingly, a similar approach has been used in the settings of health technology assessment (HTA) e.g. using Gaussian process regression and generalised additive models [229, 230]. The trade-off between precision and external validity remains a fundamental topic in Monte Carlo simulation studies: increased computational capabilities allow investigating an ever-growing number of scenarios (and with a sufficient number of replications to successfully control the Monte Carlo error). The downside of this *brute force* approach consists of having to deal with results that are cumbersome to summarise, describe, and disseminate (as in the example that motivated the nested loop plot of Rücker and Schwarzer). Given the increasing popularity of simulation studies (as outlined in Chapter 4 and Figure 4.1 specifically), the meta-modelling approach of Skrondal could, therefore, be revised and made more accessible to researchers making use of Monte Carlo simulation studies. Practically speaking, this would require describing incomplete fractional design with applied examples, designing an example simulation study using both approaches, and comparing the conclusions that one would draw from such experiments. Further to that and most importantly, the benefits and limitations of each approach need to be thoroughly discussed and contextualised.

8.4.3 Extensions of the Joint Modelling Approach for Longitudinal Electronic Health Records

The joint modelling approach for analysing longitudinal data in the settings of EHRs, as described in Chapter 7, could be further extended. The general framework that I illustrate here enables the analysis of longitudinal EHRs, potentially multivariate and potentially of different types. Several layers of nesting are supported, e.g. patients nested within hospitals nested within regions, and follows from the extended mixed-effects modelling framework of Crowther [216].

The first, natural extension consists of modelling both the observation and the drop-out process. As discussed in Chapter 6, modelling the drop-out process allows accommodating informative drop-out processes in the analysis; this is particularly relevant in the settings of EHRs, as individuals with worse recorded values of the longitudinal biomarker are likely at higher risk of dropping out, e.g. because of death. The drop-out process could be modelled using a parametric or flexible parametric survival model, accommodating a wide variety of settings. The resulting joint model, describing the sub-models with a simplified notation for illustration purposes, follows in Equation (8.1):

$$\begin{cases} y & = X_Y \beta_y + Z_Y b_y + \epsilon_y \\ h(t) & = h_0(t) \exp(X_h \gamma + W_h(t|\beta, b)\eta_h) \\ r(t) & = r_0(t) \exp(X_r \alpha + W_r(t|\beta, b)\eta_r) \end{cases} \tag{8.1}$$

where y is the longitudinal outcome, $h(t)$ is the drop-out process, $r(t)$ is the observation process, $\{X_y, X_h, X_r\}$ and Z_y are (potentially overlapping) covariates - fixed and random effects, respectively. $W_h(\cdot)$, $W_r(\cdot)$ are any (possibly multivariate) function of the random effects b describing the association between the longitudinal outcome and the drop-out and observation processes.

The next extension of the joint model from Equation (8.1) consist of jointly modelling a second longitudinal outcome, as in Equation (8.2):

$$\begin{cases} y_1 & = X_{y_1} \beta_{y_1} + Z_{y_1} b_{y_1} + \epsilon_{y_1} \\ y_2 & = X_{y_2} \beta_{y_2} + Z_{y_2} b_{y_2} + \epsilon_{y_2} \\ h(t) & = h_0(t) \exp(X_h \gamma + W_h(t|\beta, b)\eta_h) \\ r(t) & = r_0(t) \exp(X_r \alpha + W_r(t|\beta, b)\eta_r) \end{cases} \tag{8.2}$$

where the longitudinal outcomes y_1 and y_2 can be thought to originate from a multivariate distribution (e.g. bivariate normal), and the residual errors ϵ_{y_i} are allowed to be correlated. The joint model of Equation (8.2) is also assuming that the two longitudinal outcomes share the same drop-out process $h(t)$ and observation process $r(t)$. This assumption could be further relaxed, e.g. by allowing distinct observation processes (Equation (8.3)):

$$\begin{cases} y_1 & = X_{y_1} \beta_{y_1} + Z_{y_1} b_{y_1} + \epsilon_{y_1} \\ y_2 & = X_{y_2} \beta_{y_2} + Z_{y_2} b_{y_2} + \epsilon_{y_2} \\ h(t) & = h_0(t) \exp(X_h \gamma + W_h(t|\beta, b)\eta_h) \\ r_1(t) & = r_{0,1}(t) \exp(X_{r_1} \alpha + W_{r_1}(t|\beta_{y_1}, b_{y_1})\eta_{r_1}) \\ r_2(t) & = r_{0,2}(t) \exp(X_{r_2} \alpha + W_{r_2}(t|\beta_{y_2}, b_{y_2})\eta_{r_2}) \end{cases} \tag{8.3}$$

In this settings, there are two distinct $r_i(t)$ sub-models for each i^{th} observation process relative to each longitudinal outcome y_i. The model structure for each observation process can therefore vary, e.g. different subsets of covariates could be included in each sub-model. Analogously, the joint model of Equation (8.3) can be modified to allow outcome-specific drop-out processes, e.g. by including two sub-models $h_1(t)$ and $h_2(t)$ instead of $h(t)$. Finally, any number of longitudinal outcomes could be included in the joint model, with sub-models for the drop-out and observation process that are outcome-specific (or not); the correlation between the various sub-models is governed by the association structure W, with the possibility of selecting outcome-specific formulations (e.g. any combination of the association structures previously described in Section 6.4).

Extensions of the joint modelling approach of Chapter 7 that I introduced and described so far focussed on extending the number of outcomes that could be included in the joint model, allowing (potentially) outcome-specific sub-models for the drop-out and observation processes. The next obvious extension consists of allowing more general formulations in the time to event sub-models: for instance, fully parametric and most importantly flexible parametric formulations could be selected to model the baseline hazard functions $h_0(t)$ and $r_0(t)$.

Another extension consists of modelling the within-subject variability of longitudinal biomarkers (e.g. blood pressure) within the joint model [231, 232]: the approaches currently published in the literature disregard the observation and the drop-out processes, which could lead to biased results (as discussed in Chapters 6 and 7). By using the joint modelling framework described in Chapter 7 all of the above could be accommodated within the analysis.

The joint modelling approach could also be used to plan additional longitudinal measurements using personalised screening intervals [233, 234].

Finally, missing data in the joint modelling framework is an issue closely related to that of modelling the drop-out process; in fact, as described in Chapter 6, drop-out corresponds to a specific missingness pattern. There is considerable research interest on the topic (e.g. Moreno-Betancur *et al.* [235]), especially in the settings of EHRs, and further investigations are warranted.

A fundamental step of this future work requires the careful definition of the joint modelling framework within which all the above-mentioned extension are formulated [216]. Next, the more complex joint models need to be fully and thoroughly compared with more simple approaches (such as mixed-effects models that completely ignore the drop-out and the observation process), e.g. via Monte Carlo simulation; models that should be included in this comparison are also the joint models that disregard either the drop-out or the observation process. This comparison is of primary interest, as it is important to understand in practice the consequences of failing to account for informative drop-out, informative observation times, or both in the analysis; it is also important to understand the robustness of this joint modelling framework to model misspecification and modelling assumptions.

Further to that, the joint modelling approach needs to be compared with established approaches such as the methods based on inverse probability weighting described by Hernán *et al.* [203]. In brief, the inverse probability weighting approach consists of estimating weights for the informative drop-out process and weights for the observation process, obtaining overall weights by multiplying them together and using the overall weights in the analysis of the longitudinal outcome of interest. This approach did not perform well in the settings described in Chapter 7; hence, it would be interesting to see if its performance improves in more complex settings.

Nevertheless, to the best of my knowledge, there is no other comprehensive framework that would allow incorporating the drop-out process and the observation process directly within the analysis of (potentially multivariate) longitudinal data, allowing to flexibly define the association structure between the outcomes and the formulation of each sub-model. Most importantly, the joint models described in this Section can be fitted using readily available statistical software [217, 218], despite the heavy computational requirements (especially when the number of random effects increases). Alternative approaches to Gaussian quadrature (e.g. Monte Carlo integration) need to be evaluated in these settings to assess potential computational benefits.

Finally, applications using real-world data are necessary to illustrate the joint modelling approach in practice and disseminate its use.

8.4.4 *Simulation of Realistic Longitudinal Electronic Health Records Data*

The future developments highlighted as necessary in the previous Section relies on being able to simulate longitudinal outcomes that realistically mimic the settings of EHRs. This is a non-trivial task, as there is no established approach currently published in the literature - especially when it comes to simulating clinically plausible observation processes. Interestingly, the drop-out process could be simulated using the approaches described by Crowther and Lambert and based on an underlying joint longitudinal-survival model [68].

Neuhaus *et al.* [219] highlighted features of the observation process that clinicians-researchers identified as relevant when distinguishing between informative and non-informative observation processes in the settings of EHRs:

1. Many visits are unplanned, and correlated with the health status of a patient (e.g. feeling ill), with this information typically not available to data analysts;
2. Visits are missed, commonly for reasons related to their condition. This information is not available to data analysts either;
3. The timing between visits is highly irregular, and seemingly not following any stochastic process easily identifiable;
4. Visiting patterns are often clustered, e.g. individuals with a given disease and treatment pattern seem to follow a more homogeneous visiting schedule (between individuals).

Consequently, they introduced an approach for simulating longitudinal observations with an informative (or not) observation process that can accommodate all the characteristics outlined above. Their approach begins by generating a grid of possible observations for a given number of study subjects, e.g. weekly measurements for five years of follow-up; the true longitudinal trajectories are then simulated using e.g. a mixed-effects model. Then, they define a model for the probability of each observation being observed:

$$\operatorname{logit}[P(R_{ij} = 1|\cdot)] = f(\cdot), \tag{8.4}$$

where R_{ij} is an indicator variable that takes the value one when the i^{th} observation for the j^{th} individual from the grid of possible observations is observed, zero otherwise. The probability of being observed then depends on any characteristic $(\cdot)$ of the true underlying longitudinal process: as an example, it could include current values of the longitudinal trajectory, lagged values, or a function of them (e.g. a rolling mean or the rate of change). The parameters of $f(\cdot)$ govern the strength and the direction of the association between characteristics of the longitudinal outcome and the observation process.

Despite allowing a large amount of flexibility when simulating longitudinal data with an informative observation process, the approach of Neuhaus *et al.* requires pre-specifying a grid of possible observation times (and values) and therefore yields an observation process that is defined (and simulated) in discrete time. Conversely, in the simulation study of Chapter 7, I simulated the observation process in continuous time via a recurrent events survival model that modelled the within-individual correlation via a frailty term. Compared to the approach of Neuhaus *et al.*, the approach of the simulations of Chapter 7 did not require the pre-specification of a grid of possible observation times and still retained flexibility in defining the association structure. Most interestingly, the approach used in Chapter 7 could be traced to a well-defined joint model (i.e. the true data-generating model), yielding the additional benefit of being able to fit the true data-generating model to the simulated data.

To allow the methods comparison of Section 8.4.3, a flexible framework that builds on the approaches outlined so far needs to be developed. This framework needs to allow simulating (potentially multivariate) longitudinal outcomes with a flexible specification of drop-out and observation processes and association structures. Ideally, the drop-out and observation processes could depend on the characteristic of the longitudinal outcomes (and/or a function of them), as in the approach of Neuhaus *et al.*; the implications of simulating the observation process in discrete time versus continuous time should be assessed as well (if any). Finally - and most importantly -, the simulation framework should be implemented in readily available software packages to enable other researchers to use the methodology with ease.

8.5 FINAL CONCLUSIONS

Electronic health records (EHRs) are being used increasingly often for clinical and epidemiological research, providing the opportunity to answer innovative and more detailed clinical questions. Despite that, new challenges arise - especially when applying traditional statistical methods that may see their assumptions violated in the settings of EHRs. In this Thesis, I explored methods that can be used to accommodate the intrinsic hierarchical structure of EHRs; in particular, I investigated multilevel survival models and joint models for survival and longitudinal data (where the clustering unit is the individual). The former methods can be used to model the within-cluster correlation, while the latter can be used to account for a variety of violations in the assumptions of traditional mixed-effects models for longitudinal data, or jointly accommodate longitudinal and time to event outcomes. In particular, the joint modelling approach can be used to jointly model the longitudinal outcome of interest and the drop-out process, the observation process, or both. The use of these methods could help overcome the limitations of traditional methods - especially in the settings of EHRs - and avoid biases that would otherwise arise.

Acknowledging the need for advanced statistical methods to tackle more complex applied settings and illustrating their use with practical examples are fundamental steps, empowering epidemiologists, statisticians, and applied researchers alike to utilise EHRs to their full potential and answer the above-mentioned innovative clinical questions. Ultimately, this would lead to an increased understanding of disease conditions and (new or established) treatments for the highest benefit of patients and individuals within a given health-care system.

A *Research Paper:* `rsimsum`

The manuscript on the R package `rsimsum`, further described in Chapter 4, has been published in the Journal of Open Source Software and is available at the following DOI: 10.21105/joss.00739. The manuscript is also included in this Appendix.

AG developed the software package and wrote the manuscript. Dr. LeBeau and Dr. Leeper contributed to improving the manuscript and the software through the peer review process for the Journal of Open Source Software, which is openly available online (`https://github.com/openjournals/joss-reviews/issues/739`).

The Journal of Open Source Software

rsimsum: Summarise results from Monte Carlo simulation studies

Alessandro Gasparini[1]

1 Biostatistics Research Group, Department of Health Sciences, University of Leicester

DOI: 10.21105/joss.00739

Software

- Review
- Repository
- Archive

Submitted: 11 May 2018
Published: 20 June 2018

Summary

Monte Carlo simulation studies are numerical methods for conducting computer experiments based on generating pseudo-random observations from a known truth. Monte Carlo simulation studies - referred from now on as *simulation studies* for conciseness - represent a powerful tool and have several practical applications in statistical and biostatistical research: among others, evaluating new or existing statistical methods, comparing them, assessing the impact of modelling assumption violations, and helping with the understanding of statistical concepts. Establishing properties of current methods is necessary to allow using them with confidence; however, sometimes properties are very hard (if not impossible) to derive analytically: large sample approximation is possible, but evaluating the goodness of the approximation to finite samples is required. Approximations often require assumptions as well: what are the consequences of violating such assumptions? Simulation studies can help answer these questions. They can also help answer additional questions such as: is an estimator biased in a finite sample? Do confidence intervals for a given parameter achieve the desired nominal level of coverage? How does a newly developed method compare to an established one? What is the power to detect a desired effect size under complex experimental settings and analysis methods?

The increased availability of powerful computational tools (both personal and high-performance cluster computers), the perceived efficacy, and the emergence of specialist courses and tutorial papers on simulation studies (Morris, White, and Crowther 2017) contributed to the rise of simulation studies in the current literature. Despite that, simulation studies are often poorly designed, analysed, and reported (Morris, White, and Crowther 2017): information on data-generating mechanisms (DGMs), number of repetitions, software, estimands are often lacking or poorly reported, making critical appraisal and replication of published studies a difficult task. Another aspect of simulation studies that is often poorly reported or not reported at all is the Monte Carlo error of summary statistics, defined as the standard deviation of the estimated quantity over repeated simulation studies. Monte Carlo errors play an important role in understanding the role of chance in results of simulation studies and have been showed to be severely underreported (Koehler, Brown, and Haneuse 2009).

`rsimsum` is an R package that can compute summary statistics from simulation studies. `rsimsum` is modelled upon a similar package available in Stata, the user-written command `simsum` (White 2010), but - to the best of our knowledge - there is no similar package in R. The aim of `rsimsum` is to help to report simulation studies, including understanding the role of chance in results of simulation studies: Monte Carlo standard errors and confidence intervals based on them are computed and presented to the user by default. `rsimsum` can compute a wide variety of summary statistics: bias, empirical and model-based standard errors, relative precision, relative error in model standard error, mean squared error, coverage, bias. Further details on each summary statistic are presented elsewhere (White 2010; Morris, White, and Crowther 2017).

The main function of `rsimsum` is called `simsum` and can handle simulation studies with a single estimand of interest at a time. Missing values are excluded by default, and it is possible to define boundary values to drop estimated values or standard errors exceeding such limits (e.g. standardised values larger than 10). It is possible to define a variable representing methods compared with the simulation study, and it is possible to define factors that vary between the different simulated scenarios (data-generating mechanisms, DGMs). However, methods and DGMs are not strictly required: in that case, a simulation study with a single scenario and a single method is assumed. Finally, `rsimsum` provides a function named `multisimsum` that allows summarising simulation studies with multiple estimands as well.

An important step of reporting a simulation study consists in visualising the results; therefore, `rsimsum` exploits the R package `ggplot2` (Wickham 2009) to produce a portfolio of opinionated data visualisations for quick exploration of results, inferring colours and facetting by data-generating mechanisms. `rsimsum` includes methods to produce (1) plots of summary statistics with confidence intervals based on Monte Carlo standard errors (forest plots, bar plots, and lolly plots), (2) zip plots to graphically visualise coverage by directly plotting confidence intervals (Morris, White, and Crowther 2017), and (3) heat plots. The latter is a visualisation type that has not been traditionally used to present results of simulation studies, and consists in a mosaic plot where the factor on the x-axis is the methods compared with the current simulation study and the factor on the y-axis is one of the data-generating factors, as selected by the user: see for instance Figure 1, which can be obtained via a single function call with `rsimsum`. Each tile of the mosaic plot is coloured according to the value of the summary statistic of interest, with a red colour representing values above the target value and a blue colour representing values below the target.

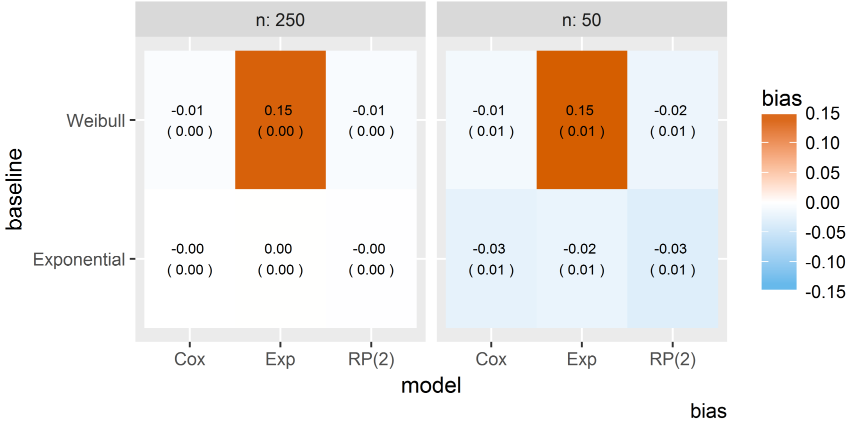


Figure 1: example of heat plot that can be obtained with `rsimsum` via a single function call. The example data comes from a simulation study on model misspecification in survival models, and it is bundled with `rsimsum` (see `help("relhaz", package = "rsimsum")`).

References

Koehler, Elizabeth, Elizabeth Brown, and Sebastien JPA Haneuse. 2009. "On the Assessment of Monte Carlo Error in Simulation-Based Statistical Analyses." *The American Statistician* 63 (2):155–62. https://doi.org/10.1198/tast.2009.0030.

Morris, Tim P, Ian R White, and Michael J Crowther. 2017. “Using Simulation Studies to Evaluate Statistical Methods.” *arXiv Preprint arXiv:1712.03198*.

White, Ian R. 2010. “Simsum: Analyses of Simulation Studies Including Monte Carlo Error.” *The Stata Journal* 10 (3):369–85.

Wickham, Hadley. 2009. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. http://ggplot2.org.

B Research Paper: INTEREST: INteractive Tool for Exploring REsults from Simulation sTudies

The manuscript on the interactive web app INTEREST, further described in Chapter 4, has been submitted for publication in the Journal of Data Science, Statistics, and Visualisation and it is currently under review. A pre-print is also available from arXiv (`https://arxiv.org/abs/1909.03813`), and is included in this Appendix.

All authors discussed the idea leading to the inception of INTEREST. AG led the development of INTEREST, with constructive feedback from all co-authors. AG drafted the manuscript, with input and feedback from all co-authors. TP and MJC wrote the Stata code for the example simulation study used as a case study. All authors read and approved the final manuscript.

JDSSV

Journal of Data Science, Statistics, and Visualisation

MMMMMM YYYY, Volume VV, Issue II. *doi: XX.XXXXX/jdssv.v000.i00*

INTEREST: INteractive Tool for Exploring REsults from Simulation sTudies

Alessandro Gasparini
University of Leicester

Tim P. Morris
MRC Clinical Trials Unit at UCL

Michael J. Crowther
University of Leicester

Abstract

Simulation studies allow us to explore the properties of statistical methods. They provide a powerful tool with a multiplicity of aims; among others: evaluating and comparing new or existing statistical methods, assessing violations of modelling assumptions, helping with the understanding of statistical concepts, and supporting the design of clinical trials. The increased availability of powerful computational tools and usable software has contributed to the rise of simulation studies in the current literature. However, simulation studies involve increasingly complex designs, making it difficult to provide all relevant results clearly. Dissemination of results plays a focal role in simulation studies: it can drive applied analysts to use methods that have been shown to perform well in their settings, guide researchers to develop new methods in a promising direction, and provide insights into less established methods. It is crucial that we can digest relevant results of simulation studies. Therefore, we developed **INTEREST**: an *INteractive Tool for Exploring REsults from Simulation sTudies*. The tool has been developed using the **Shiny** framework in R and is available as a web app or as a standalone package. It requires uploading a tidy format dataset with the results of a simulation study in R, Stata, SAS, SPSS, or comma-separated format. A variety of performance measures are estimated automatically along with Monte Carlo standard errors; results and performance summaries are displayed both in tabular and graphical fashion, with a wide variety of available plots. Consequently, the reader can focus on simulation parameters and estimands of most interest. In conclusion, **INTEREST** can facilitate the investigation of results from simulation studies and supplement the reporting of results, allowing researchers to share detailed results from their simulations and readers to explore them freely.

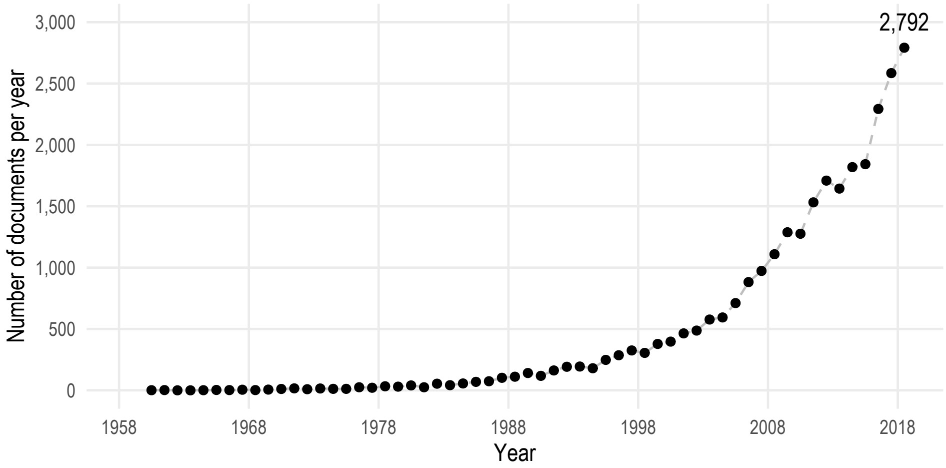
Keywords: Simulation study, Monte Carlo, Visualisation, Reporting, R, **Shiny**, Replicability.

1. Background

Monte Carlo simulation studies are computer experiments based on generating pseudo-random observations from a known truth. Statisticians usually mean *Monte Carlo simulation study* when they say *Simulation study*; throughout this article, we will just use *simulation study* but this encapsulates Monte Carlo simulation studies. Simulation studies have several applications and represent an invaluable tool for statistical research nowadays: in statistics, establishing properties of current methods is key to allow them to be used – or not – with confidence. Sometimes it is not possible to derive exact analytical properties; for example, a large sample approximation may be possible, but evaluating the approximation in finite samples is required. Approximations often require assumptions as well: what are the consequences of violating such assumptions? Monte Carlo simulation studies come to the rescue and can help to answer these questions. They also can help answer questions such as: is an estimator biased in a finite sample? What are the consequences of model misspecification? Do confidence intervals for a given parameter achieve the advertised/nominal level of coverage? How does a newly developed method compare to an established one? What is the power to detect a desired effect size under complex experimental settings and analysis methods?

Simulation studies are being used increasingly in a wide variety of settings. For instance, searching on the database of peer-reviewed research literature Scopus (`https://www.scopus.com`) with the query string `TITLE-ABS-KEY ("simulation study") AND SUBJAREA (math)` yields more than 25000 results with a 25-fold increase during the last 30 years, from 111 documents in 1988 to 2792 in 2018 (Figure 1). The increased availability of powerful computational tools and ready-to-use software to researchers surely contributed to the rise of simulation studies in the current literature.

Figure 1: Trend in published documents on simulation studies from 1960 onwards. The number of documents was identified on Scopus via the search key `TITLE-ABS-KEY ("simulation study") AND SUBJAREA (math)`, and the number of documents identified in 2018 is labelled on the plot.



Despite the popularity of simulation studies, they are often poorly designed, analysed, and reported. Morris *et al.* reviewed 100 research articles published in Volume 34 of *Statistics in Medicine* (2015) with at least one simulation study and found that information on data-generating mechanisms (DGMs), number of repetitions, software, and estimands were often lacking or poorly reported, making critical appraise and replication of published studies a difficult task (Morris et al. 2019) . Another aspect of simulation studies that is often poorly reported or not reported at all is the Monte Carlo error of estimated performance measures, defined as the standard error of estimated performance, owing to the fact that a finite number of repetitions are used and so performance is estimated with uncertainty. Monte Carlo errors play an important role in understanding the role of chance in the results of simulation studies and have been showed to be severely underreported (Koehler et al. 2009).

The possibility of independently verifying results from scientific studies is a fundamental aspect of science (Laine et al. 2007); as a consequence, several reporting guidelines have emerged under the banner of the EQUATOR Network (`http://www.equator-network.org`) (Schulz et al. 2010; von Elm et al. 2007). Despite similar calls for harmonised reporting to allow for greater reproducibility in the area of computation science (e.g. Peng (Peng 2011)) and several articles advocating for more rigour in specific aspects of simulation studies (Hoaglin and Andrews 1975; Hauck and Anderson 1984; Díaz-Emparanza 2002; Burton et al. 2006; White 2010; Smith and Marshall 2011), design and reporting guidelines for simulation studies are lacking; Morris *et al.* introduced the ADEMP framework (Aims, Data-generating mechanisms, Estimands, Methods, Performance measures) aiming to fill precisely that gap. In the *Reporting* section they compared the several ways of reporting results that they observed in their reviews, including results in text for small simulation studies, tabulating and plotting results, and even the nested-loop plot proposed by Rücker and Schwarzer for fully-factorial simulation studies with many data-generating mechanisms (Rücker and Schwarzer 2014). They concluded by arguing that *there is no correct way to present results, but we encourage careful thought to facilitate readability, considering the comparisons that need to be made.*

As outlined in Spiegelhalter *et al.*, there is little experimental evidence on how different types of visualisations are perceived (Spiegelhalter et al. 2011); despite that, they highlight the ease of improving understanding via interactive visualisations that can be adjusted by the user to best fit specific requirements. The recent advent of tools such as Data-Driven Documents (**D³**, or **D3.js**) (Bostock et al. 2011) and **Shiny** (Chang et al. 2019) has further facilitated the development of interactive visualisations.

The increased availability of powerful computational tools has not only contributed to a rise in the popularity of simulation studies, it has also allowed researchers to simulate an ever-growing number of data-generating mechanisms and include several estimands and methods to compare: up to 4.2×10^{10}, 32, and 33, respectively, in the aforementioned review (Morris et al. 2019). With a large number of data-generating mechanisms, estimands, or methods, analysing and reporting the results of a simulation study becomes cumbersome: what results shall we focus on so as not to bewilder readers? Which estimands and methods should we include in our tables and plots? How should we plot or tabulate several data-generating mechanisms at once?

In an attempt to address these questions, we developed **INTEREST**, an *INteractive Tool for Exploring REsults from Simulation sTudies*. **INTEREST** is a browser-based interactive tool, and it requires first uploading a dataset with results from a simulation study; then, it estimates performance measures and it displays a variety of tables and plots automatically. The user can focus on specific data-generating mechanisms, estimands, and methods: tables and plots are updated automatically. This article will introduce the implementation details of **INTEREST** in the *Implementation* section and the main features in the *Results and discussion* section, where we will further discuss its relevance. We also present a case study to motivate the use of INTEREST and illustrate its use in practice. Finally, we conclude the manuscript with some ending remarks in the *Conclusions* section.

2. Implementation

INTEREST was developed using the free statistical software R (R Core Team 2019) and the R package **Shiny** (Chang et al. 2019). **Shiny** is an R package (and framework) that allows building interactive web apps straight from within R: the resulting applications can be hosted online, embedded in reports and dashboards, or just run as standalone apps.

The front-end of **INTEREST** has been built using the **shinydashboard** package (Chang and Borges Ribeiro 2018); **shinydashboard** is based upon **AdminLTE** (`https://adminlte.io/`), an open-source admin control panel built on top of the Bootstrap framework (Version 3.x) and released under the MIT license.

The back-end functionality of **INTEREST** is published as a standalone R package named **rsimsum** for easier long-term maintainability (Gasparini 2018); **rsimsum** is freely available on the Comprehensive R Archive Network (CRAN) under the GNU General Public License Version 3 (`https://www.gnu.org/licenses/gpl-3.0`).

INTEREST is available as an online application and as a standalone version for offline use. The online version is hosted at `https://interest.shinyapps.io/interest/`, and can be accessed via any web browser on any device (desktop computers, laptops, tablets, smartphones, etc.). The standalone offline version can be obtained from GitHub (`https://github.com/ellessenne/interest`) and can be run on any desktop computer and laptop with a local instance of R; if required, R can be downloaded for free from the website of the R project (R Core Team 2019). INTEREST (as **rsimsum**) is published under the GNU General Public License Version 3.

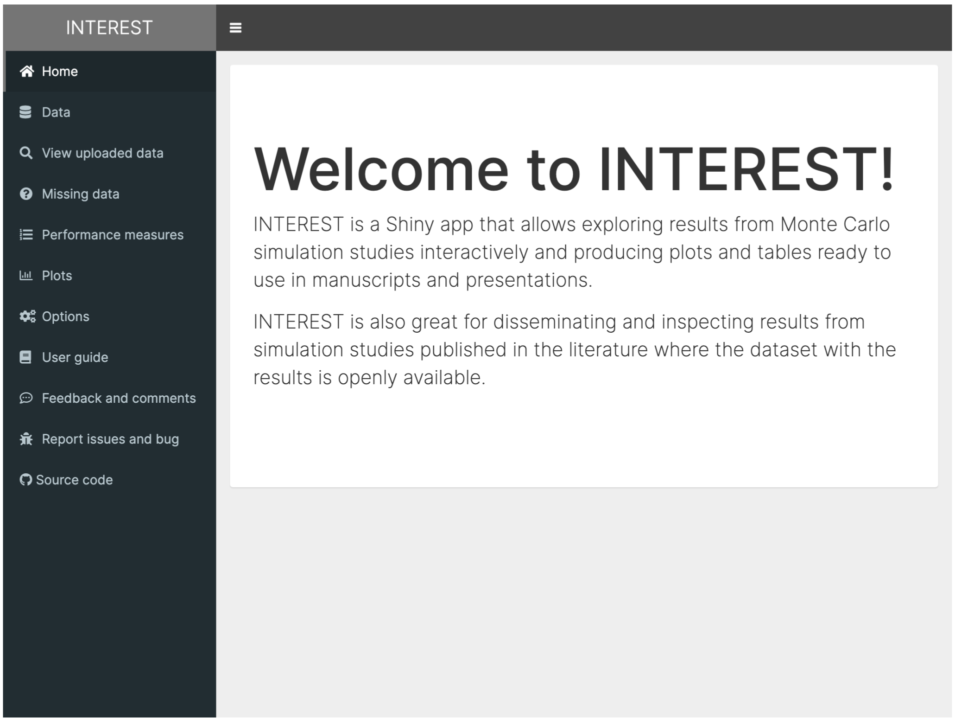
3. Results and discussion

The main interface of **INTEREST** is presented in Figure 2. The interface is composed of a main area on the right and a navigation bar on the left; the navigation bar includes sub-menus for customising plots or modifying the default behaviour of **INTEREST**. We now introduce and describe the functionality of the application.

3.1. Data

The use of **INTEREST** starts by providing a tidy dataset (also known as long format, with variables in columns and observations in rows (Wickham 2014); an example of

Figure 2: Homepage of **INTEREST**. On the left, there is a navigation bar with sub-menus useful to tune the default behaviour of the app. On the right, the main window of **INTEREST**.



Welcome to INTEREST!

INTEREST is a Shiny app that allows exploring results from Monte Carlo simulation studies interactively and producing plots and tables ready to use in manuscripts and presentations.

INTEREST is also great for disseminating and inspecting results from simulation studies published in the literature where the dataset with the results is openly available.

Table 1: Example of dataset in tidy format, with each row identifying a replication for each combination of data-generating me table directly exported from **INTEREST**, case study DGM 2: true Weibull baseline hazard function.

Replication	DGM	Method	Estimate
1	1	1	$\hat{\theta}_{1,1,1}$
2	1	1	$\hat{\theta}_{2,1,1}$
3	1	1	$\hat{\theta}_{3,1,1}$
1	2	1	$\hat{\theta}_{1,2,1}$
2	2	1	$\hat{\theta}_{2,2,1}$
3	2	1	$\hat{\theta}_{3,2,1}$
1	1	2	$\hat{\theta}_{1,1,2}$
2	1	2	$\hat{\theta}_{2,1,2}$
3	1	2	$\hat{\theta}_{3,1,2}$
1	2	2	$\hat{\theta}_{1,2,2}$
2	2	2	$\hat{\theta}_{2,2,2}$
3	2	2	$\hat{\theta}_{3,2,2}$
⋮	⋮	⋮	⋮

tidy data is included in Table 1) with results from a simulation study via the *Data* tab from the side menu. A dataset can be provided to **INTEREST** in three different ways:

1. The user can upload a dataset. The uploaded file can be a comma-separated file (`.csv`), a Stata dataset (version 8-15, `.dta`), an SPSS dataset (`.sav`), a SAS dataset (`.sas7bdat`), or an R serialised object (`.rds`); the format will be inferred automatically from the extension of the uploaded file, and the auto-detection is case-insensitive. It is also possible to upload compressed files (ending in `.gz`, `.bz2`, `.xz`, or `.zip`) that are automatically decompressed;

2. The user can provide a URL link to a dataset hosted elsewhere. All considerations relative to the file format from point (1) are also valid here;

3. Finally, the user can paste a dataset (e.g. from Microsoft Excel) in a text box. The pasted data is assumed to be tab-separated.

Once a dataset has been uploaded via one of the three methods outlined, the user will have to define the variables required by **INTEREST** and some optional variables, depending on the structure of the input dataset. The names of each column (i.e. variable) from the uploaded dataset automatically populate a set of select-list inputs to assist the user. A variable defining a point estimate from the simulation study and a variable representing the standard error of such estimates are required, and the user has to define the true value of the estimand of interest as well. Additionally, a user can define a variable representing methods being compared with the current simulation study (and choose the comparator), and one or more variables defining data-generating mechanisms (DGMs, e.g. sample size, true correlation, true baseline hazard function for survival models, etc.).

The *View uploaded data* side tab in **INTEREST** displays the dataset uploaded by the user using the R package **DT**, an R interface to the **DataTables** plug-in for jQuery (Xie et al. 2019). The resulting table is interactive and can be sorted and filtered by the user. It is good practice to verify that the uploaded dataset is as expected before continuing with the analysis and any visual exploration.

3.2. Missing data

INTEREST includes a section for exploring missingness of estimates and/or standard errors from each repetition of a simulation study, which may occur, for example, due to non-convergence of some repetitions. Missing values need to be carefully explored and handled at the initial stage of any analysis. Missingness may originate as a consequence of software failures: if so, the code could (or should) be made more robust to ensure fewer or no failures. Conversely, missing data may arise as a consequence of characteristics of the simulated data, yielding to non-convergence of the estimation procedures. In other words, missing values may not be missing completely at random. A discussion on the interpretation of missing values can be found elsewhere (White et al. 2011; Morris et al. 2019).

The missing data functionality is based on the R package **naniar** (Tierney et al. 2019), and can be accessed via the *Missing data* tab. It comprises visual and tabular summaries; missing data visualisations available in **INTEREST** are the following:

- Bar plots of number (or proportion) of missing values by method and data-generating mechanism (if defined). Number and proportion of missing values are produced for each variable included in the data uploaded to **INTEREST**;
- A plot to visualise the amount of missing data in the whole dataset;
- A scatter plot with missing status depicted with different colours; to be able to plot missing values, they are replaced with values 10% lower than the minimum value in that variable. This plot allows identifying trends and patterns between variables in missing values (e.g. all estimates with a very large standard error have a missing point estimate);
- A heat plot with methods on the horizontal axis and the data-generating mechanisms on the vertical axis, with the colour fill representing the percentage of missingness in each tile.

Each plot can be further customised and exported (e.g. for use in slides and reports): more details in the *Plots* section below. Finally, **INTEREST** computes and outputs a table with the number, proportion, and the cumulative number of missing values per variable, stratifying by method and data-generating mechanisms; the table can be easily exported to $\LaTeX$ format for further use (via the R package **xtable** (Dahl et al. 2019)).

3.3. Performance measures

INTEREST estimates performance measures automatically as soon as the user defines the required variables via the *Data* tab. Supported performance measures are presented

Table 2: Overview of performance measures estimated by **INTEREST**.

Performance measure	Description
Bias	Deviation between estimate and the true value
Empirical standard error	Log-run standard deviation of the estimator
Relative precision against a reference	Precision of a method B compared to a reference method A
Mean squared error	The sum of squared bias and variance of the estimator
Model standard error	Average estimated standard error
Coverage	Probability that a confidence interval contains the true value
Bias-eliminated coverage	Coverage after removing bias, i.e. with confidence intervals centered on the estimated value rather than the true value of the estimand
Power	Power of a significance test

in Table 2, and discussed in more detail elsewhere (Burton et al. 2006; White 2010; Morris et al. 2019). In addition to that, **INTEREST** returns mean and median estimate, and mean and median squared error of the estimate. Finally, **INTEREST** computes and returns Monte Carlo standard errors by default. The list of performance measures estimated by **INTEREST** can be customised via the *Options* tab: by default, all are included.

3.4. Tables

Estimated performance measures are presented in tabular form in the *Performance measures* side tab, once again using the R package **DT**. The table of estimated performance measures is relative to a given data-generating mechanism, which can be modified using a select list input on the side. It is also possible to customise the number of significant digits and to select whether Monte Carlo standard errors should be excluded in each table or not via the *Options* tab.

Finally, it is possible to export the tables in two ways:

1. Export the table in LaTeX format, e.g. for use in reports, articles, or presentations, via the *Export table* tab and the R package **xtable** (Dahl et al. 2019). The caption of the table can be directly customised;

2. Export estimated performance measures as a dataset, e.g. to be used with a different software package of choice. The table of estimated performance measures can be exported as displayed by **INTEREST** or in tidy format, and in a variety of formats: comma-separated (`.csv`), tab-separated (`.tsv`), R (`.rds`), Stata (version 8-15, `.dta`), SPSS (`.sav`), and SAS (`.sas7bdat`).

3.5. Plots

INTEREST can produce a variety of plots to automatically visualise results from simulation studies. Plots produced by **INTEREST** can be categorised into two broad groups: plots of estimates (and their estimated standard errors) and plots of performance, following analysis. Plots for method-wise comparisons of estimated values and standard errors are:

- Scatter plots;
- Bland-Altman plots (Altman and Bland 1983; Bland and Altman 1999);
- Ridgeline plots (Wilke 2018).

Each plot will include all data-generating mechanisms by default and allows comparing serial trends and the relative performance of methods included in the simulation study.

Conversely, the following plots are supported for estimated performance:

- Plots of performance measures with confidence intervals based on Monte Carlo standard errors. There are two variations of this plot: forest plots, and lolly plots. Both methods display the estimated performance measure alongside confidence intervals based on Monte Carlo standard errors; different methods are arranged side by side, either on the horizontal or on the vertical axis;
- Heat plots of performance measures: these plots are mosaic plots where the several methods being compared (if defined) are on the horizontal axis and the data-generating mechanisms are on the vertical axis. Then, each tile of the mosaic plot is coloured according to the value of a given performance measure. To the best of our knowledge, this is a novel way of visualising results from simulation studies, with an application in practice that can be found elsewhere (Gasparini et al. 2019);
- Zip plots to visually explain coverage probabilities by plotting the confidence intervals directly. More information on zip plots is presented elsewhere (Morris et al. 2019);
- Nested loop plots, useful to compare performance measures from studies with several DGMs at once. This visualisation is described in more detail elsewhere (Rücker and Schwarzer 2014).

Finally, all plots can be exported for use in manuscript, reports, or presentations by simply clicking the *Save plot* button underneath a plot; all plots are exported by default in `.png` format, but other options are available via the *Options* tab. For instance, to suit a wide variety of possible use cases, **INTEREST** supports several alternative image formats such as `pdf`, `svg`, and `eps`. Through the *Options* tab it is also possible to customise the resolution of the plot for non-vectorial format (in dots per inch, `dpi`) and the physical size (height and width) of the plots to be exported. The *Options* tab allows further customisations: for instance, it is possible to (1) define a custom label for the x-axis and the y-axis and (2) change the overall appearance of the plot by applying one of the predefined themes (which are described in more detail in the *User guide* tab).

3.6. INTEREST for exploring results

INTEREST allows researchers to upload a dataset with the results of their Monte Carlo simulation study obtaining estimates of performance in a quick and straightforward way. This is very appealing, especially with simulation studies with several data-generating mechanisms where it could be confusing to investigate all scenarios at once. Using the app it is possible to vary data-generating mechanisms and obtain updated tables and plots in real-time, therefore allowing to quickly iterate and take into consideration all possible scenarios.

3.7. INTEREST for disseminating results

One of the intended usage scenarios for **INTEREST** consists of supplementing reporting of simulation studies. This is especially useful with large simulation studies, where it is most cumbersome to summarise all results in a manuscript: it is common to include in the main manuscript only a subset of results for conciseness. The remaining results are then relegated to supplementary material, web appendices, or not published at all - undermining dissemination and replicability of a study.

Furthermore, given that it is becoming increasingly common to publish the code of simulation study, one could publish the dataset with the results alongside the code used to obtain it. That dataset could then be uploaded to **INTEREST** by readers, who could then explore the full results of the study as they wish. Given the ubiquity of web services like GitHub (`https://github.com`) and data-sharing repositories such as Zenodo (`https://zenodo.org/`), we encourage **INTEREST** users to publish online the full results of their simulation studies for other users to download and experiment with.

4. Future developments

Although **INTEREST** is fully functional in its current state, several future developments are being planned. For instance, we aim to include support for multiple estimands at once as currently supported by **rsimsum** via the `multisimsum` function. We also aim to improve the flexibility of **INTEREST** in terms of customisation (of tables and plots), e.g. by displaying the raw R code used to generate the plots behind the scenes. Finally, we are considering adding additional interactive features to the app via HTML widgets, **D³**, or other approaches; there are several R packages that allow incorporating interactive graphs into **Shiny** apps such as **htmlwidgets** (Vaidyanathan et al. 2018), **plotly** (Sievert 2018), and **r2d3** (Luraschi and Allaire 2018).

5. Case study

The case study included in this Section illustrates the use of **INTEREST** to analyse publicly available results of a simulation study. In particular, we will be using the results from the worked illustrative example included in Morris *et al.* (Morris et al. 2019).

The study dataset contains the results of a simulation study comparing three different methods for estimating the hazard ratio in a randomised trial with a time to event outcome. In particular, the methods being compared are proportional hazards survival models of the kind:

$$h_i(t) = h_0(t)\exp(X_i\theta),$$

where θ is the log hazard ratio for the effect of a binary exposure (e.g. treatment). This class of models requires an assumption regarding the shape of the baseline hazard function $h_0(t)$: it can be assumed to follow a given parametric distribution, or it can be left unspecified (yielding therefore a Cox model). The *aim* of this simulation study consists of assessing the impact of such an assumption on the estimation of the log hazard ratio.

Morris *et al.* consider two distinct *data-generating mechanisms*, varying the baseline hazard function:

1. An exponential baseline hazard with $\lambda = 0.1$ (DGM = 1);
2. A Weibull baseline hazard with $\lambda = 0.1, \gamma = 1.5$ (DGM = 2).

In both settings, data are simulated on 300 patients with a binary covariate (e.g. treatment) simulated using $X_i \sim \text{Bern}(0.5)$ - simple randomisation with an equal allocation ratio. The log hazard ratio is set to be $\theta = -0.50$; this is the true value of the *estimand* of interest.

Three distinct *methods* are fit to each simulated scenario: a parametric survival model that assumes an exponential baseline hazard, a parametric survival model that assumes a Weibull baseline hazard, and a Cox semi-parametric survival model.

Finally, the *performance measures* of interest are bias, coverage, empirical and model-based standard errors. Assuming that $\text{Var}(\hat{\theta}) \leq 0.04$, 1600 repetitions are run to ensure that the Monte Carlo standard error of bias (the key performance measure of interest) is lower than 0.005.

The dataset with the results of this simulation study is publicly available, and can be downloaded from GitHub: `https://github.com/tpmorris/simtutorial/raw/master/Stata/estimates.dta`. Within the dataset published on GitHub, the exponential, Weibull, and Cox models are coded as model 1, 2, and 3, respectively. The above-mentioned dataset is in Stata format; an R version is available as well (`https://github.com/tpmorris/simtutorial/raw/master/R/estimates.rds`), and **INTEREST** supports both.

The workflow of **INTEREST** starts by providing the dataset with the results of the simulation study. Given that the dataset is already available online, we can directly pass the URL above to **INTEREST** and then define the required variables (as illustrated in Figure 3); the uploaded dataset can then be verified via the *View uploaded data* tab (Figure 4).

We can also customise the performance measures reported by **INTEREST** via the *Options* tab (Figure 5), e.g. focussing on those outlined above as key performance

Table 3: Example of LaTeX table directly exported from **INTEREST**, case study DGM 2: true Weibull baseline hazard function.

Performance Measure	1	2	3
Bias in point estimate	0.0494 (0.0035)	0.0048 (0.0038)	0.0062 (0.0038)
Empirical standard error	0.1381 (0.0024)	0.1516 (0.0027)	0.1511 (0.0027)
Model-based standard error	0.1539 (0.0001)	0.1541 (0.0001)	0.1542 (0.0001)
Coverage of nominal 95% confidence interval	0.9600 (0.0049)	0.9556 (0.0051)	0.9575 (0.0050)

measures (bias, coverage probability, empirical standard errors, model-based standard errors).

The next step of the workflow consists of investigating missing values: this can be achieved via the *Missing data* tab. In particular, there is no missing data in the study dataset (Figure 6). We can, therefore, continue the analysis knowing that there is no pattern of serial missingness or non-convergence issues in our data.

The performance measures of interest are tabulated in the *Performance measures* tab, e.g. for DGM = 2 (Figure 7). We can see that bias for the exponential model is much larger than the Weibull and Cox models: approximately 10% of the true value (in absolute terms) compared to less than 1%. Empirical and model-based standard errors are quite similar for the Weibull and Cox models; conversely, the exponential model seemed to overestimate the model-based standard error. Coverage was as advertised for all methods, at approximately 95%. By comparison, all models performed equally in the other scenario (DGM = 1); these results are omitted from the manuscript for brevity, but we encourage readers to replicate this analysis and verify our statement.

The *Performance measures* tab provides a LaTeX table ready to be pasted e.g. in a manuscript: the resulting table is included as Table 3. A dataset with all the estimated performance measures here tabulated can also be exported to be used elsewhere (Figure 8).

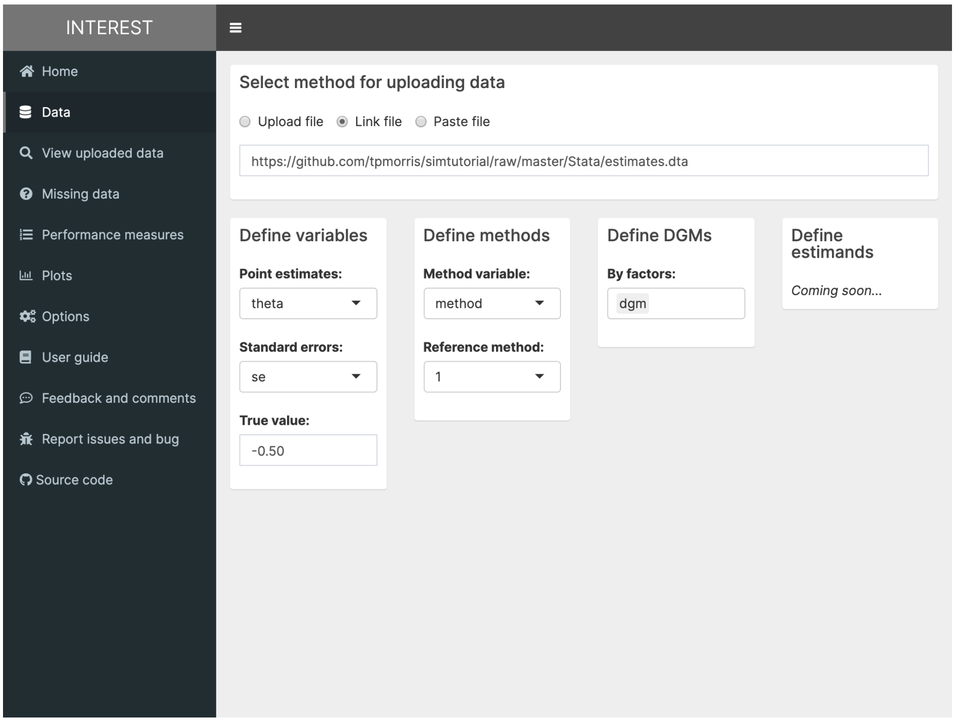
We can also visualise the results of this simulation study. First, we can produce a method-wise comparison of point estimates from each method using e.g. scatter plots (Figure 9) or Bland-Altman plots (Figure 10). With both plots, it is possible to appreciate that for the DGM with $\gamma = 1.5$ the exponential model yields point estimates that are quite different compared to the Weibull and Cox models. Analogous plots can be obtained for estimated standard errors.

The performance measures tabulated in the *Performance measures* tab can also be plotted via the *Plots* tab. For instance, it is straightforward to obtain a forest plot for bias (as illustrated in Figure 11) which can be exported by clicking the *Save plot* button. The plots' appearance can also be customised via the *Options* tab, e.g. by modifying the axes' labels and the overall theme of the plot (Figure 12); the resulting forest plot, exported in `.pdf` format, is included as Figure 13. Several other data visualisations are supported by **INTEREST**, as described in the previous Sections: lolly plots, zip plots, and so on.

6. Conclusions

As outlined in the introduction, Monte Carlo simulation studies are too often poorly

Figure 3: App interface to load the dataset for the case study. **INTEREST** can import datasets that are available online by simply pasting a link to it; then, the required variables can be defined via a list of pre-populated select inputs.



Select method for uploading data

○ Upload file ● Link file ○ Paste file

https://github.com/tpmorris/simtutorial/raw/master/Stata/estimates.dta

Define variables

Point estimates: theta

Standard errors: se

True value: -0.50

Define methods

Method variable: method

Reference method: 1

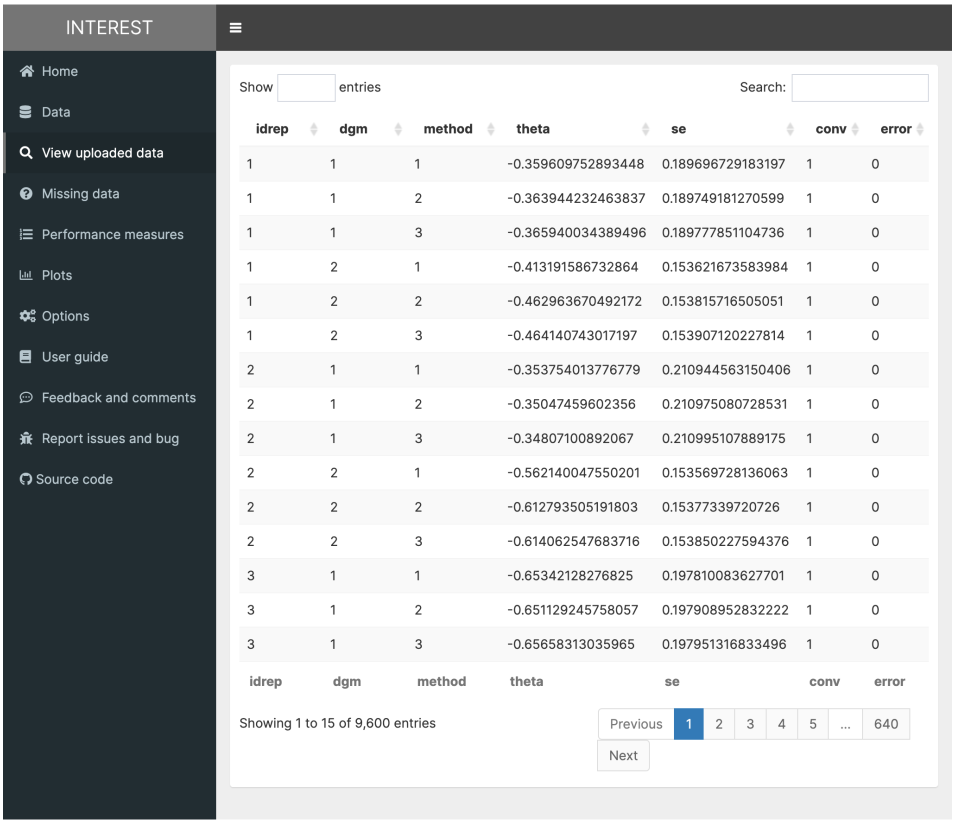
Define DGMs

By factors: dgm

Define estimands

Coming soon...

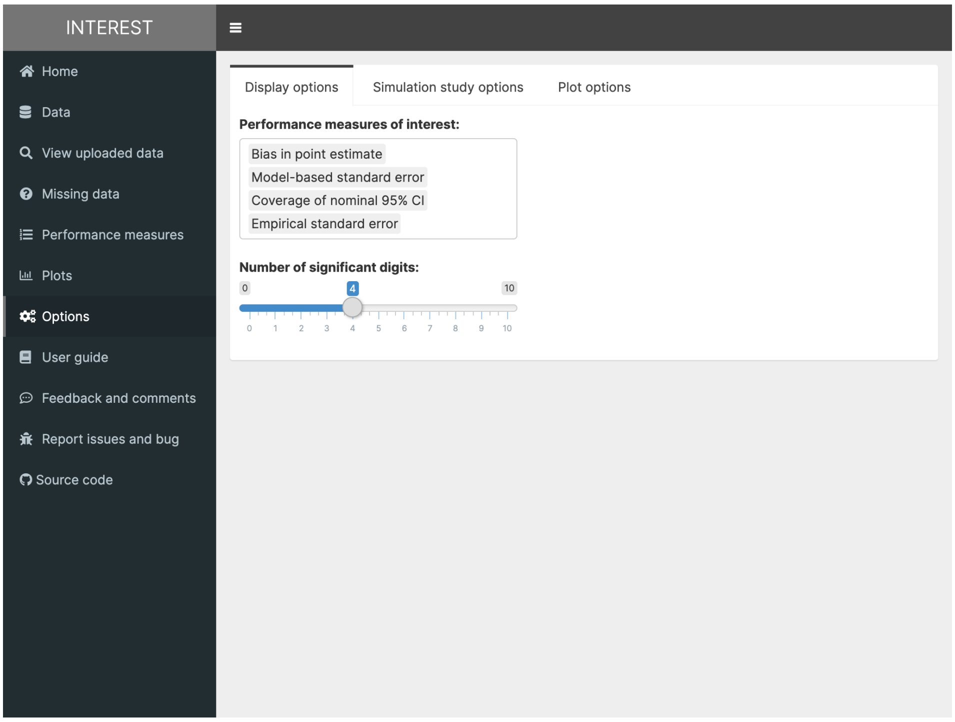
Figure 4: Verifying the dataset for the case study. After importing the study dataset, it is recommended to verify that the uploaded data is correct.



idrep	dgm	method	theta	se	conv	error
1	1	1	-0.359609752893448	0.189696729183197	1	0
1	1	2	-0.363944232463837	0.189749181270599	1	0
1	1	3	-0.365940034389496	0.189777851104736	1	0
1	2	1	-0.413191586732864	0.153621673583984	1	0
1	2	2	-0.462963670492172	0.153815716505051	1	0
1	2	3	-0.464140743017197	0.153907120227814	1	0
2	1	1	-0.353754013776779	0.210944563150406	1	0
2	1	2	-0.35047459602356	0.210975080728531	1	0
2	1	3	-0.34807100892067	0.210995107889175	1	0
2	2	1	-0.562140047550201	0.153569728136063	1	0
2	2	2	-0.612793505191803	0.15377339720726	1	0
2	2	3	-0.614062547683716	0.153850227594376	1	0
3	1	1	-0.65342128276825	0.197810083627701	1	0
3	1	2	-0.651129245758057	0.197908952832222	1	0
3	1	3	-0.65658313035965	0.197951316833496	1	0
idrep	dgm	method	theta	se	conv	error

Showing 1 to 15 of 9,600 entries

Figure 5: Customising the performance measures reported by **INTEREST**. It is possible to focus on a subset of key performance measures by selecting them via the *Options* tab.

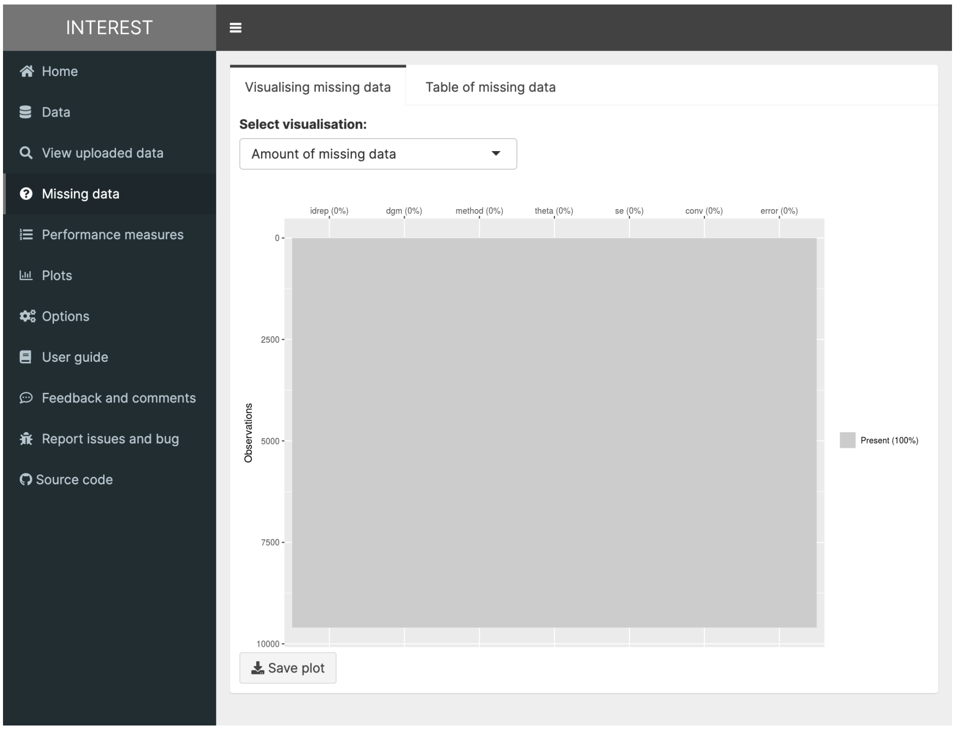


Performance measures of interest:

Bias in point estimate
Model-based standard error
Coverage of nominal 95% CI
Empirical standard error

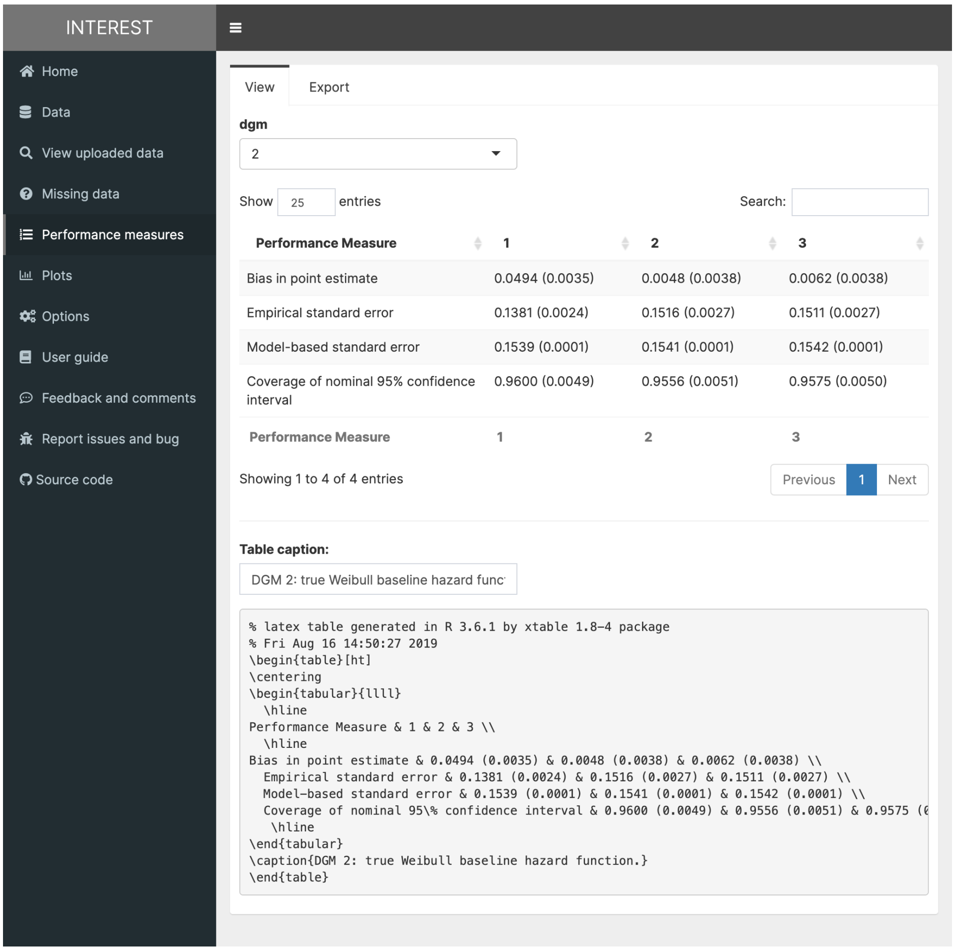
Number of significant digits: 4

Figure 6: Investigating missing data. Missingness patterns in the study dataset need to be assessed before continuing with the analysis. Several visualisations and tabular displays are available from the *Missing data* tab.



Select visualisation: Amount of missing data

Figure 7: Table of performance measures for a given DGM. Performance measures of interest are tabulated in the *Performance measures* tab, e.g. for the 2$^{\text{nd}}$ DGM (with a Weibull baseline hazard function).



dgm: 2

Show 25 entries

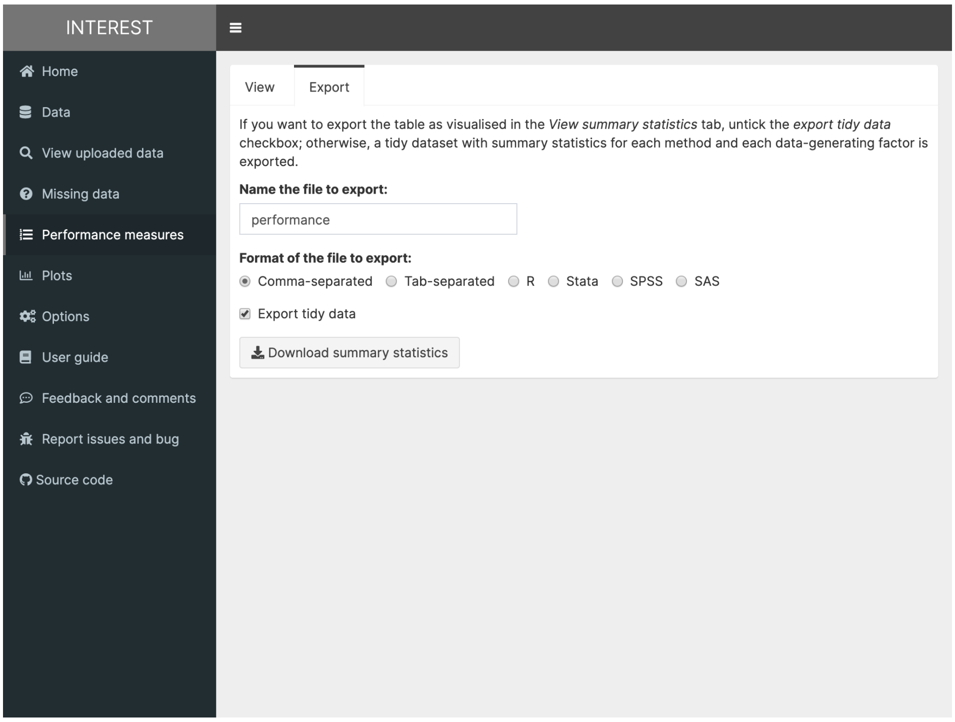
Performance Measure	1	2	3
Bias in point estimate	0.0494 (0.0035)	0.0048 (0.0038)	0.0062 (0.0038)
Empirical standard error	0.1381 (0.0024)	0.1516 (0.0027)	0.1511 (0.0027)
Model-based standard error	0.1539 (0.0001)	0.1541 (0.0001)	0.1542 (0.0001)
Coverage of nominal 95% confidence interval	0.9600 (0.0049)	0.9556 (0.0051)	0.9575 (0.0050)
Performance Measure	1	2	3

Showing 1 to 4 of 4 entries

Table caption: DGM 2: true Weibull baseline hazard func

```
% latex table generated in R 3.6.1 by xtable 1.8-4 package
% Fri Aug 16 14:50:27 2019
\begin{table}[ht]
\centering
\begin{tabular}{llll}
  \hline
Performance Measure & 1 & 2 & 3 \\
  \hline
Bias in point estimate & 0.0494 (0.0035) & 0.0048 (0.0038) & 0.0062 (0.0038) \\
  Empirical standard error & 0.1381 (0.0024) & 0.1516 (0.0027) & 0.1511 (0.0027) \\
  Model-based standard error & 0.1539 (0.0001) & 0.1541 (0.0001) & 0.1542 (0.0001) \\
  Coverage of nominal 95\% confidence interval & 0.9600 (0.0049) & 0.9556 (0.0051) & 0.9575 (
   \hline
\end{tabular}
\caption{DGM 2: true Weibull baseline hazard function.}
\end{table}
```

Figure 8: Exporting options for estimated performance measures. Performance measures of interest can be exported in a variety of formats ready to be used elsewhere (e.g. for dissemination purposes or to develop ad-hoc visualisations).



If you want to export the table as visualised in the *View summary statistics* tab, untick the *export tidy data* checkbox; otherwise, a tidy dataset with summary statistics for each method and each data-generating factor is exported.

Name the file to export:

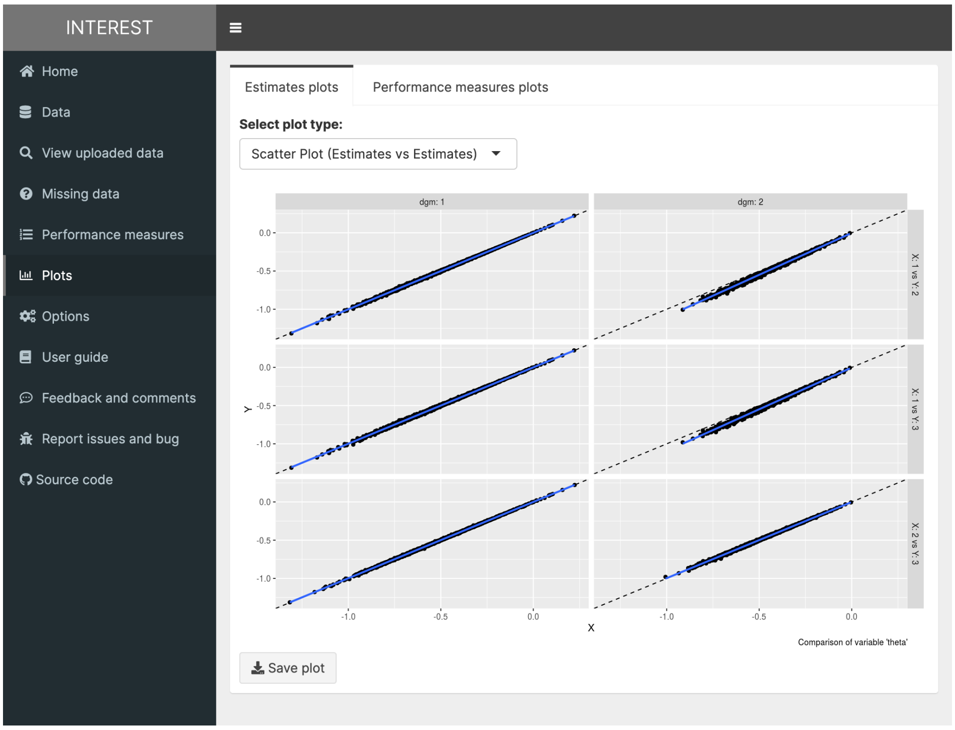
performance

Format of the file to export:

● Comma-separated ○ Tab-separated ○ R ○ Stata ○ SPSS ○ SAS

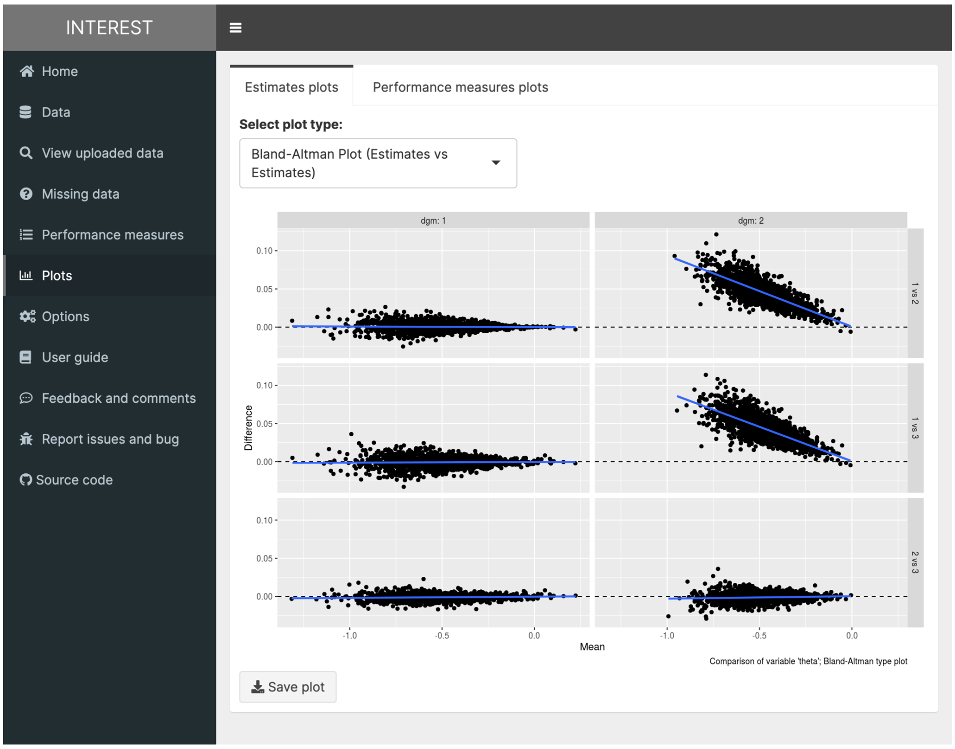
☑ Export tidy data

Figure 9: Visual comparison of point estimates via scatter plots. Points estimates for each method-DGM combination can be produced automatically using **INTEREST**.



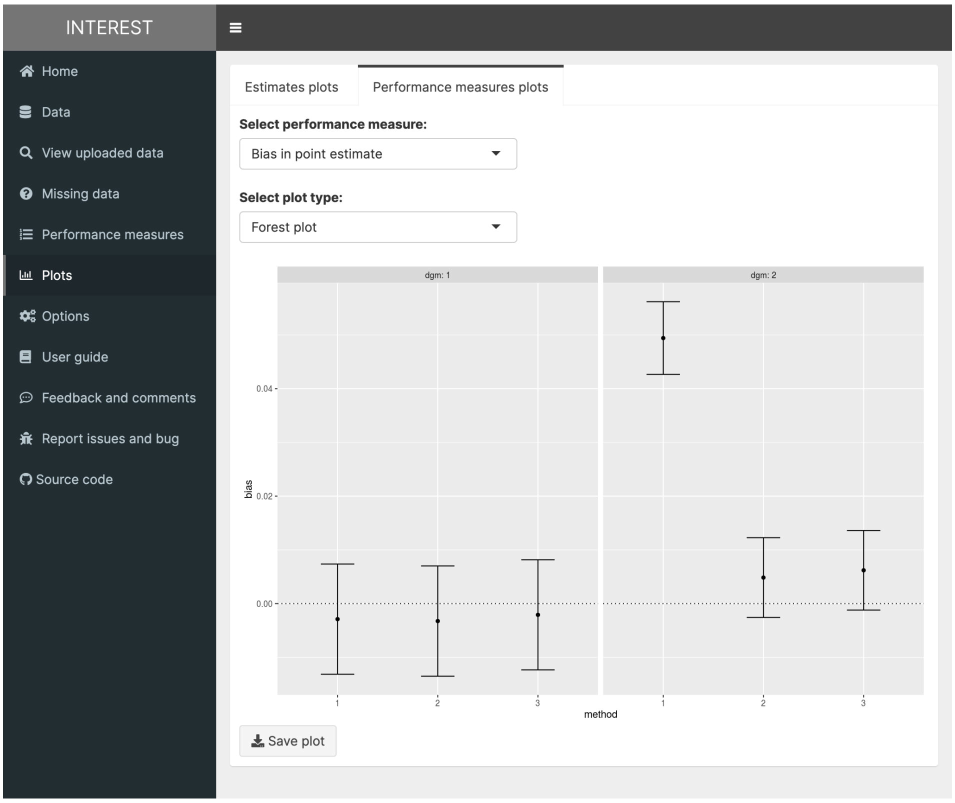
Select plot type: Scatter Plot (Estimates vs Estimates)

Figure 10: Visual comparison of point estimates via Bland-Altman plots. Points estimates for each method-DGM combination can be produced automatically using **INTEREST**.



Select plot type: Bland-Altman Plot (Estimates vs Estimates)

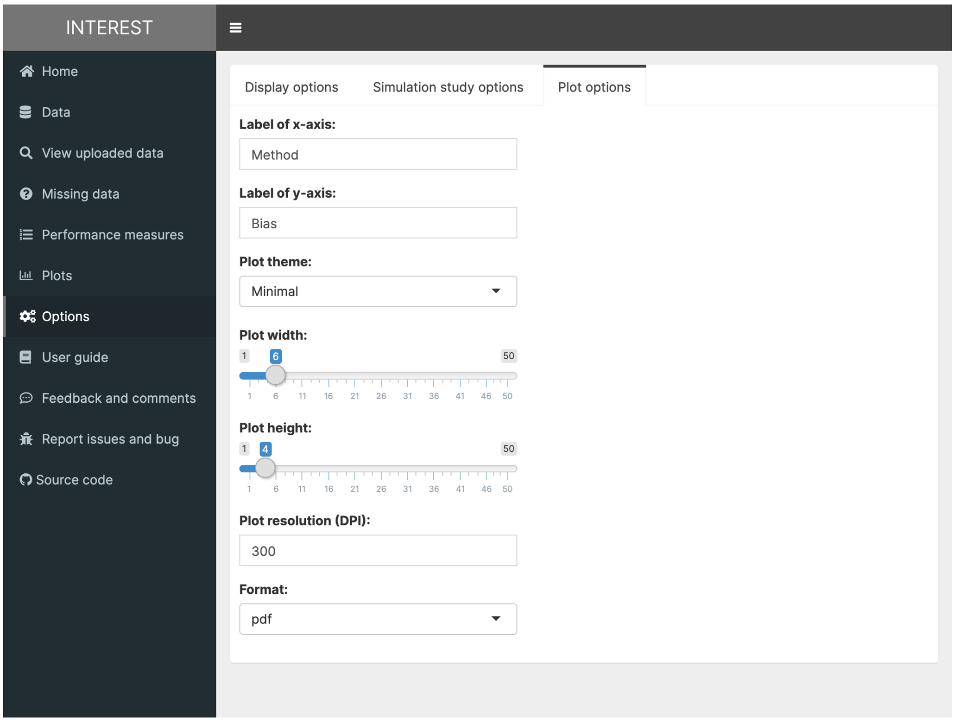
Figure 11: Visual comparison of performance measures via forest plots. Estimated performance measures such as bias can be easily plotted via the *Plots* tab.



Select performance measure: Bias in point estimate

Select plot type: Forest plot

Figure 12: Customising the visual appearance of plots. **INTEREST** allows customising the appearance of plots produced by the app via the *Options* tab, e.g. by modifying the axes' labels and/or the overall theme.



Label of x-axis: Method

Label of y-axis: Bias

Plot theme: Minimal

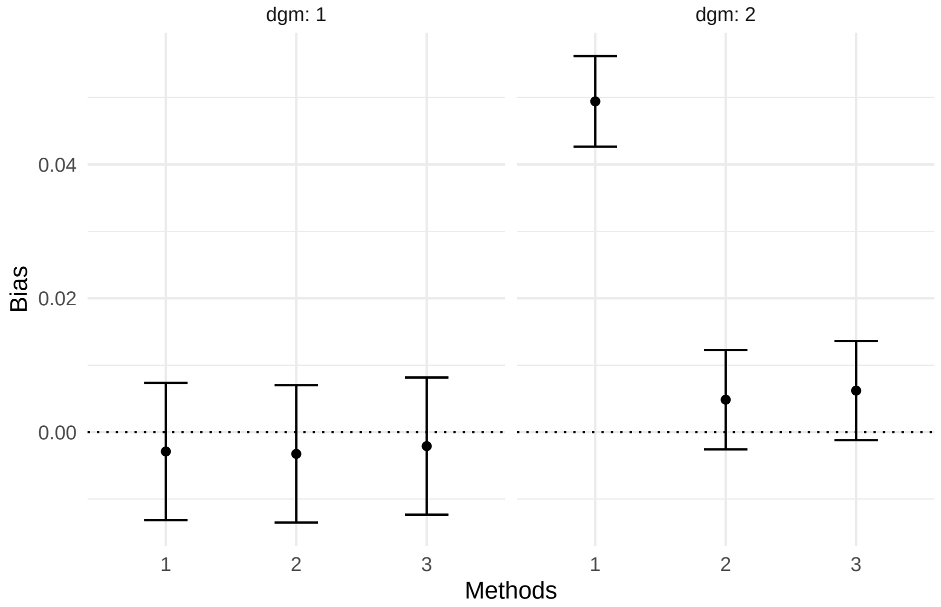
Plot width: 6

Plot height: 4

Plot resolution (DPI): 300

Format: pdf

Figure 13: Forest plot for bias, case study on survival regression modelling. This forest plot produced by **INTEREST** and further customised via the *Options* tab can be directly exported from the app.



analysed and reported (Morris et al. 2019). Given the increased use in methodological statistical research, we hope that **INTEREST** could improve reporting and disseminating results from simulation studies to a large extent. As illustrated in the case study, the exploration and analysis of the Monte Carlo simulation study of Morris *et al.* can be fully reproduced by using **INTEREST**. Estimated performance measures are tabulated automatically, and plots can be used to visualise the performance measures of interest. Moreover, the user is not constrained to a given set of plots and can fully explore the results with ease e.g. by varying DGMs to focus on or by choosing different data visualisations. Most interestingly, the only requirement to reproduce the simulation study described in the case study is a device with a web browser and connection to the Internet. To the best of our knowledge, there is no similar application readily available to be used by researchers and readers of published Monte Carlo simulation studies alike.

Acknowledgements

TPM is supported by the Medical Research Council (grant numbers MC_UU_12023/21 and MC_UU_12023/29). MJC is partially funded by the MRC-NIHR Methodology Research Panel (MR/P015433/1).

We thank Ian R. White for discussions that lead to the inception and development of **INTEREST**.

References

Altman, D. G. and Bland, J. M. (1983). Measurement in medicine: The analysis of method comparison studies. *The Statistician*, 32(3):307, DOI: 10.2307/2987937, https://doi.org/10.2307%2F2987937.

Bland, J. M. and Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, 8(2):135–160, DOI: 10.1177/096228029900800204, https://doi.org/10.1177%2F096228029900800204.

Bostock, M., Ogievetsky, V., and Heer, J. (2011). **D^3**: Data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, DOI: 10.1109/tvcg.2011.185.

Burton, A., Altman, D. G., Royston, P., and Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine*, 25(24):4279–4292, DOI: 10.1002/sim.2673.

Chang, W. and Borges Ribeiro, B. (2018). ***shinydashboard**: Create Dashboards with **shiny***, https://CRAN.R-project.org/package=shinydashboard. R package version 0.7.1.

Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2019). ***shiny**: Web Application Framework for R*, https://CRAN.R-project.org/package=shiny. R package version 1.3.2.

Dahl, D. B., Scott, D., Roosen, C., Magnusson, A., and Swinton, J. (2019). ***xtable**: Export Tables to LaTeX or HTML*, https://CRAN.R-project.org/package=xtable. R package version 1.8-4.

Díaz-Emparanza, I. (2002). Is a small Monte Carlo analysis a good analysis? *Statistical Papers*, 43(4):567–577.

Gasparini, A. (2018). **rsimsum**: Summarise results from Monte Carlo simulation studies. *Journal of Open Source Software*, 3(26):739, DOI: 10.21105/joss.00739, https://doi.org/10.21105/joss.00739.

Gasparini, A., Clements, M. S., Abrams, K. R., and Crowther, M. J. (2019). Impact of model misspecification in shared frailty survival models. *Statistics in Medicine*, DOI: 10.1002/sim.8309, https://doi.org/10.1002/sim.8309.

Hauck, W. W. and Anderson, S. (1984). A survey regarding the reporting of simulation studies. *The American Statistician*, 38(3):214–216.

Hoaglin, D. C. and Andrews, D. F. (1975). The reporting of computation-based results in statistics. *The American Statistician*, 29(3):122–126.

Koehler, E., Brown, E., and Haneuse, S. J. (2009). On the assessment of Monte Carlo error in simulation-based statistical analyses. *The American Statistician*, 63(2):155–162, DOI: 10.1198/tast.2009.0030.

Laine, C., Goodman, S. N., Griswold, M. E., and Sox, H. C. (2007). Reproducible research: Moving toward research the public can really trust. *Annals of Internal Medicine*, 146(6):450–453, DOI: `10.7326/0003-4819-146-6-200703200-00154`.

Luraschi, J. and Allaire, J. (2018). ***r2d3**: Interface to **D³** Visualizations*, `https://CRAN.R-project.org/package=r2d3`. R package version 0.2.3.

Morris, T. P., White, I., and Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, pages 1–29, DOI: `10.1002/sim.8086`.

Peng, R. D. (2011). Reproducible research in computational science. 334(6060):1226–1227, DOI: `10.1126/science.1213847`.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, `https://www.R-project.org/`.

Rücker, G. and Schwarzer, G. (2014). Presenting simulation results in a nested loop plot. *BMC Medical Research Methodology*, 14(1), DOI: `10.1186/1471-2288-14-129`.

Schulz, K. F., Altman, D. G., Moher, D., and for the CONSORT Group (2010). CONSORT 2010 Statement: Updated guidelines for reporting parallel group randomised trials. *PLOS Medicine*, 7(3):1–7, DOI: `10.1371/journal.pmed.1000251`.

Sievert, C. (2018). ***plotly** for R*, `https://plotly-r.com`.

Smith, M. K. and Marshall, A. (2011). Importance of protocols for simulation studies in clinical drug development. *Statistical Methods in Medical Research*, 20(6):613–622.

Spiegelhalter, D., Pearson, M., and Short, I. (2011). Visualizing uncertainty about the future. *Science*, 333(6048):1393–1400, DOI: `10.1126/science.1191181`.

Tierney, N., Cook, D., McBain, M., and Fay, C. (2019). ***naniar**: Data Structures, Summaries, and Visualisations for Missing Data*, `https://CRAN.R-project.org/package=naniar`. R package version 0.4.2.

Vaidyanathan, R., Xie, Y., Allaire, J., Cheng, J., and Russell, K. (2018). ***htmlwidgets**: HTML Widgets for R*, `https://CRAN.R-project.org/package=htmlwidgets`. R package version 1.3.

von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsch, P., Vandenbroucke, J. P., and for the STROBE Initiative (2007). The Strengthening the Reporting of Observational Studies in Epidemiology (Strobe) Statement: Guidelines for reporting observational studies. *PLOS Medicine*, 4(10):1–5, DOI: `10.1371/journal.pmed.0040296`.

White, I. R. (2010). **simsum**: Analyses of simulation studies including Monte Carlo error. *The Stata Journal*, 10(3):369–385.

White, I. R., Royston, P., and Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30(4):377–399, DOI: `10.1002/sim.4067`.

Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, 59(10), DOI: 10.18637/jss.v059.i10.

Wilke, C. O. (2018). ***ggridges**: Ridgeline Plots in **ggplot2***, https://CRAN.R-project.org/package=ggridges. R package version 0.5.1.

Xie, Y., Cheng, J., and Tan, X. (2019). ***DT**: A Wrapper of the JavaScript Library **DataTables***, https://CRAN.R-project.org/package=DT. R package version 0.8.

Affiliation:

Alessandro Gasparini
Biostatistics Research Group
Department of Health Sciences
University of Leicester
George Davies Centre
University Road
Leicester
LE1 7RH
United Kingdom
E-mail: ag475@leicester.ac.uk

Tim P. Morris
MRC Clinical Trials Unit at UCL
90 High Holborn
London
WC1V 6LJ
United Kingdom

Michael J. Crowther
Biostatistics Research Group
Department of Health Sciences
University of Leicester
George Davies Centre
University Road
Leicester
LE1 7RH
United Kingdom

Journal of Data Science, Statistics, and Visualisation https://jdssv.org/
published by the International Association for Statistical Computing
http://iasc-isi.org/
MMMMMM YYYY, Volume VV, Issue II *Submitted:* yyyy-mm-dd
doi:XX.XXXXX/jdssv.v000.i00 *Accepted:* yyyy-mm-dd

C R Code for the Case Study on Flexible Parametric Survival Models

The `R` code included in this Appendix is the code used to run the simulation study described in the case study of Section 4.6.

It requires loading the following packages:

```
library(tidyverse)
library(rstpm2)
library(simsurv)
library(foreach)
```

`rstpm2` [132] is used to fit flexible parametric models, and `simsurv` [69] is used to simulate complex survival data. `foreach` [236] provides the `foreach` looping construct, which is useful in the settings of simulation studies.

Then, I set the seed for reproducibility purposes:

```
set.seed(208779431)
```

I define a data frame with the DGMs, and the number of replications `B`:

```
dgms <- tibble::tibble(
  h = seq(2), # 1 = Weibull, 2 = Mixture Weibull
  dgm = seq(2) # Sequential number of DGMs: 1, 2
)
B <- 1000 # Number of replications
```

Then, I wrote the following code to iterate over the 2 DGMs (outer `foreach` loop) and then over B replications (inner `foreach` loop).

Within each replication:

1. A dataset is simulated;
2. The 3 models included in this comparison are fitted;
3. The estimated log treatment effect from each model is stored in tidy format.

Finally, results are assembled and stored within a data frame named `out`.

```
out <- foreach::foreach(
  dgm = dgms[["dgm"]],
  .combine = dplyr::bind_rows
) %do% {
  # Extract current DGM
  current <- dgms[dgms[["dgm"]] == dgm, ]

  # Run B replications for the current scenario
  out.in <- foreach::foreach(
    i = seq(B),
    .combine = dplyr::bind_rows
  ) %do% {
    # Simulate data
    covs <- tibble::tibble(
      id = seq(300),
      trt = stats::rbinom(n = 300, size = 1, prob = 0.5)
    )
    if (current[["h"]] == 1) {
      surv <- simsurv::simsurv(
        dist = "weibull",
        lambdas = 0.60,
        gammas = 0.80,
        betas = c(trt = -0.50),
        x = covs,
        maxt = 10,
        interval = c(0, 500)
      )
    } else {
      surv <- simsurv::simsurv(
        dist = "weibull",
        lambdas = c(1.00, 1.00),
        gammas = c(1.50, 0.50),
        betas = c(trt = -0.50),
        mixture = TRUE,
        pmix = 0.50,
        x = covs,
        maxt = 10,
        interval = c(0, 500)
      )
    }
    x <- dplyr::left_join(x = covs, y = surv, by = "id")

    # Transform simulated times < 1e-6 to avoid
    # problems with finite differences
    x[["eventtime"]][x[["eventtime"]] < 1e-6] <- 1e-6

    # Fit the  models included in this comparison
    m1 <- rstpm2::stpm2(
      formula = survival::Surv(eventtime, status) ~ trt,
      data = x,
      df = 2
    )
    m2 <- rstpm2::stpm2(
      formula = survival::Surv(eventtime, status) ~ trt,
      data = x,
      df = 5
    )
    m3 <- rstpm2::stpm2(
      formula = survival::Surv(eventtime, status) ~ trt,
      data = x,
      df = 10
    )

    # Export results
    res <- tibble::tibble(
      i = i,
      dgm = dgm,
      b = vapply(
        X = c(m1, m2, m3),
        FUN = function(x) summary(x)@coef[2, 1],
        FUN.VALUE = numeric(1)
      ),
      se = vapply(
        X = c(m1, m2, m3),
        FUN = function(x) summary(x)@coef[2, 2],
        FUN.VALUE = numeric(1)
      ),
      model = seq(3)
    )
    res
  }
  out.in
}
```

The data frame with all information on DGMs is then merged with the results data set:

```
out <- dplyr::left_join(x = out, y = dgms, by = "dgm") %>%
  dplyr::mutate(
    model = factor(model,
      levels = seq(3),
      labels = c("M1 (2df)", "M2 (5df)", "M3 (10df)")
    ),
    h = factor(h,
      levels = seq(2),
      labels = c("Weibull", "Mixture Weibull")
    )
  )
```

Finally, the structure of the dataset with the results of this simulation study is the following:

```
dplyr::glimpse(out)
```

```
# Observations: 6,000
# Variables: 6
# $ i     <int> 1, 1, 1, 2, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 5, 6, 6, 6, 7, 7, 7...
# $ dgm   <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
# $ b     <dbl> -0.6378815, -0.6422050, -0.6426656, -0.6571888, -0.6568290, -...
# $ se    <dbl> 0.1226920, 0.1233417, 0.1233309, 0.1231681, 0.1236395, 0.1236...
# $ model <fct> M1 (2df), M2 (5df), M3 (10df), M1 (2df), M2 (5df), M3 (10df),...
# $ h     <fct> Weibull, Weibull, Weibull, Weibull, Weibull, Weibull, Weibull...
```

This dataset is then exported, to be used in Chapter 4, Section 4.6.

```
saveRDS(object = out, file = "data/case-study.RDS")
```

D Research Paper: Impact of Model Misspecification in Shared Frailty Survival Models

The research manuscript investigating the impact of model misspecification in shared frailty survival models has been published in Statistics in Medicine; it is omitted from the e-Thesis due to copyright. The pre-prints history is available on arXiv (`https://arxiv.org/abs/1810.08140`).

AG and MJC conceived and planned the study, with input from MSC; notably, MSC suggested the interpretation of a mixture normal frailty as *hidden groups*. AG developed all the code required to simulate data, fit each models, and obtain the predictions of interest. AG analysed the results of the simulation study, and interpreted the results with critical input from all co-authors. AG wrote the manuscript, with input and feedback from all co-authors - especially during the revision process. Finally, the anonymous reviewers and the editorial team of Statistics in Medicine greatly improved the clarity and comprehensiveness of the manuscript during the peer review process.

E *Supplementary Results: Monte Carlo Simulation on Impact of Model Misspecification in Shared Frailty Survival Models*

This Appendix contains supplementary results from the Monte Carlo simulation on the impact of model misspecification in shared frailty survival models, introduced in Section 5.5.

Mean squared error for the regression coefficient is included in Figures E.1 and E.2, while mean squared error for the loss in life expectancy is included in Figures E.3 and E.4.

Finally, result for the estimated variance of the frailty (bias, coverage probability, and mean squared error) are included in Figures E.5 to E.10.

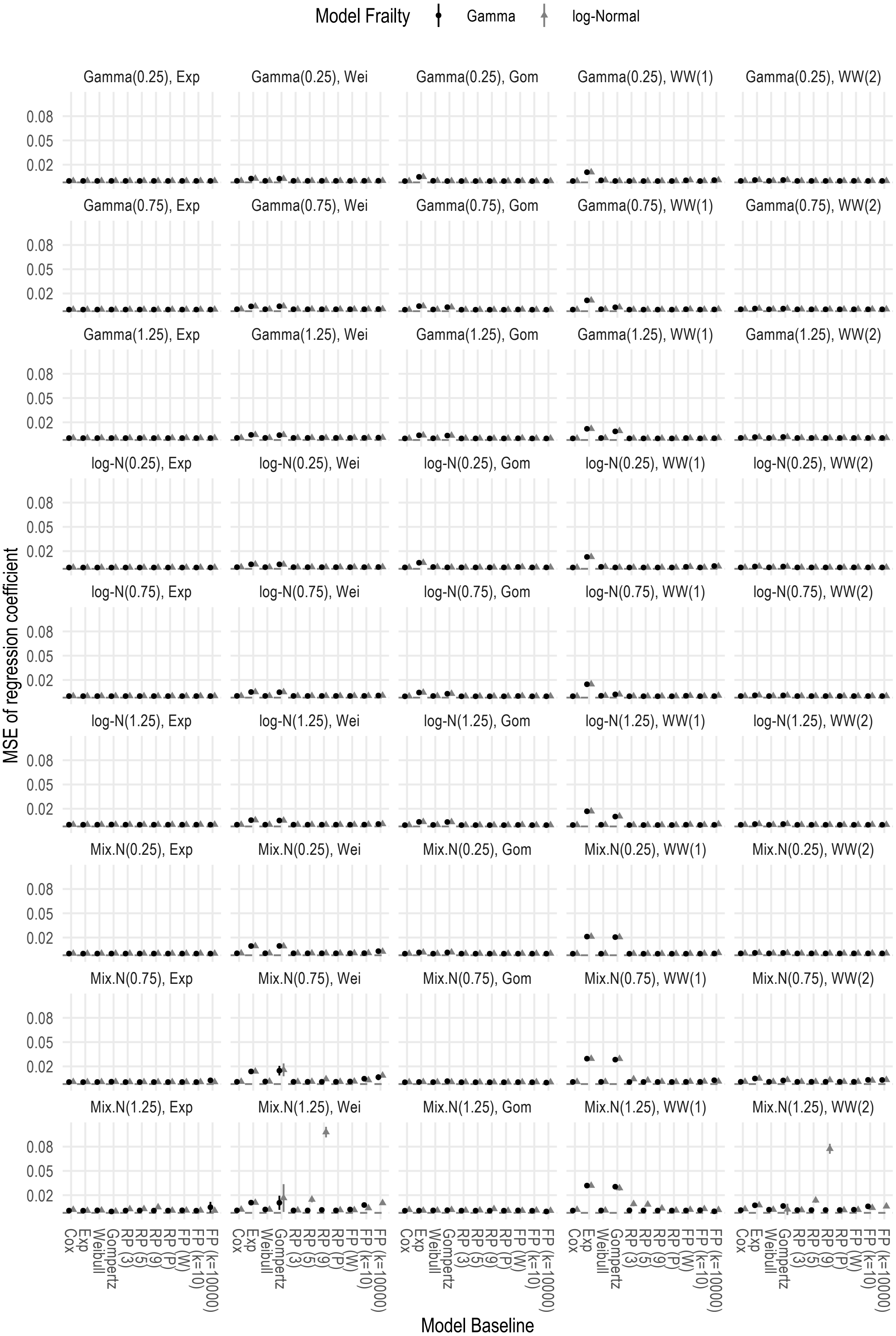


Figure E.1: MSE of regression coefficient, scenarios with 20 clusters of 150 individuals each. Colours represent the model frailty, and each subplot includes results for a given combination of data-generating baseline hazard and frailty

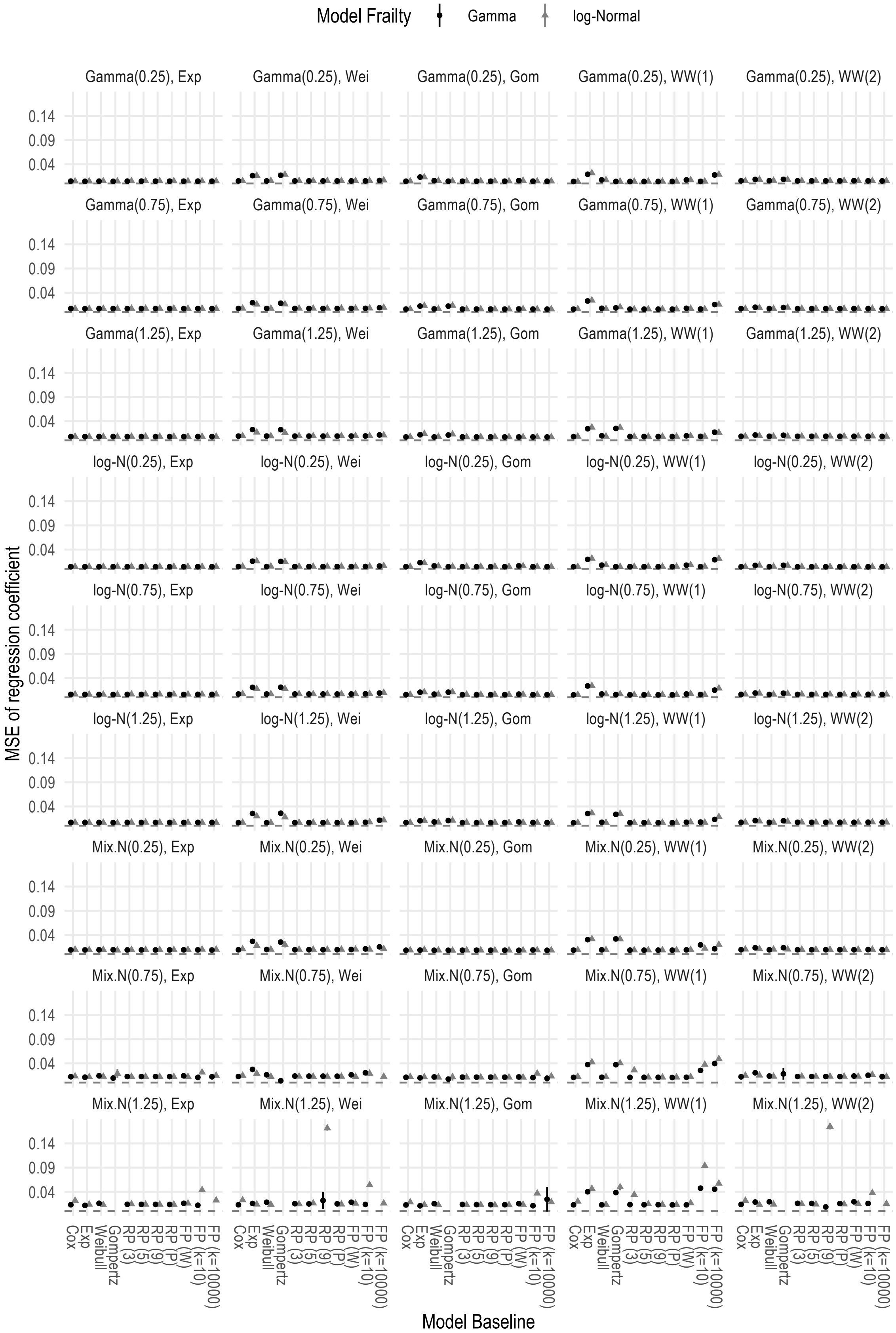


Figure E.2: MSE of regression coefficient, scenarios with 750 clusters of 2 individuals each. Colours represent the model frailty, and each subplot includes results for a given combination of data-generating baseline hazard and frailty

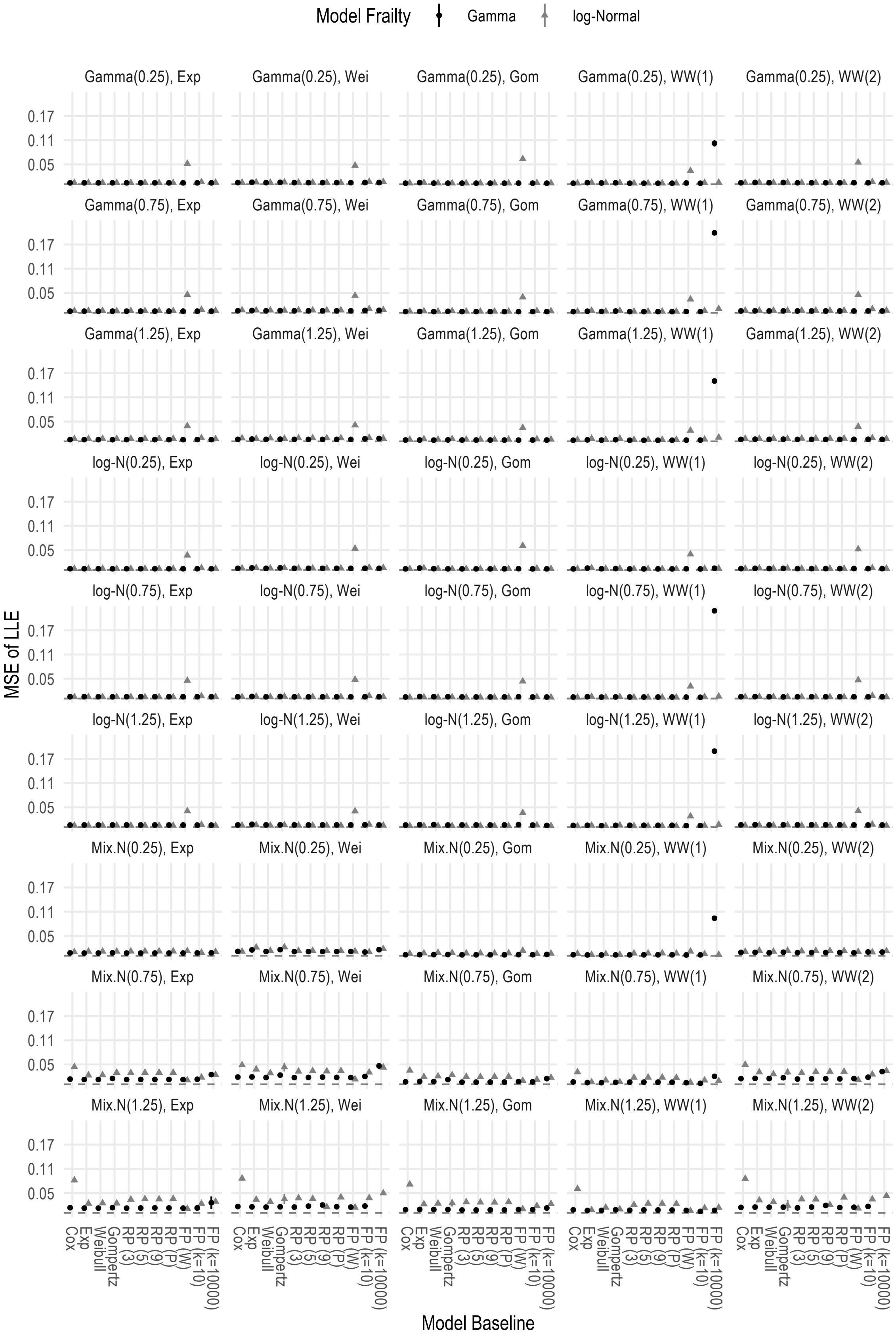


FIGURE E.3: MSE of LLE, scenarios with 20 clusters of 150 individuals each. Colours represent the model frailty, and each subplot includes results for a given combination of data-generating baseline hazard and frailty

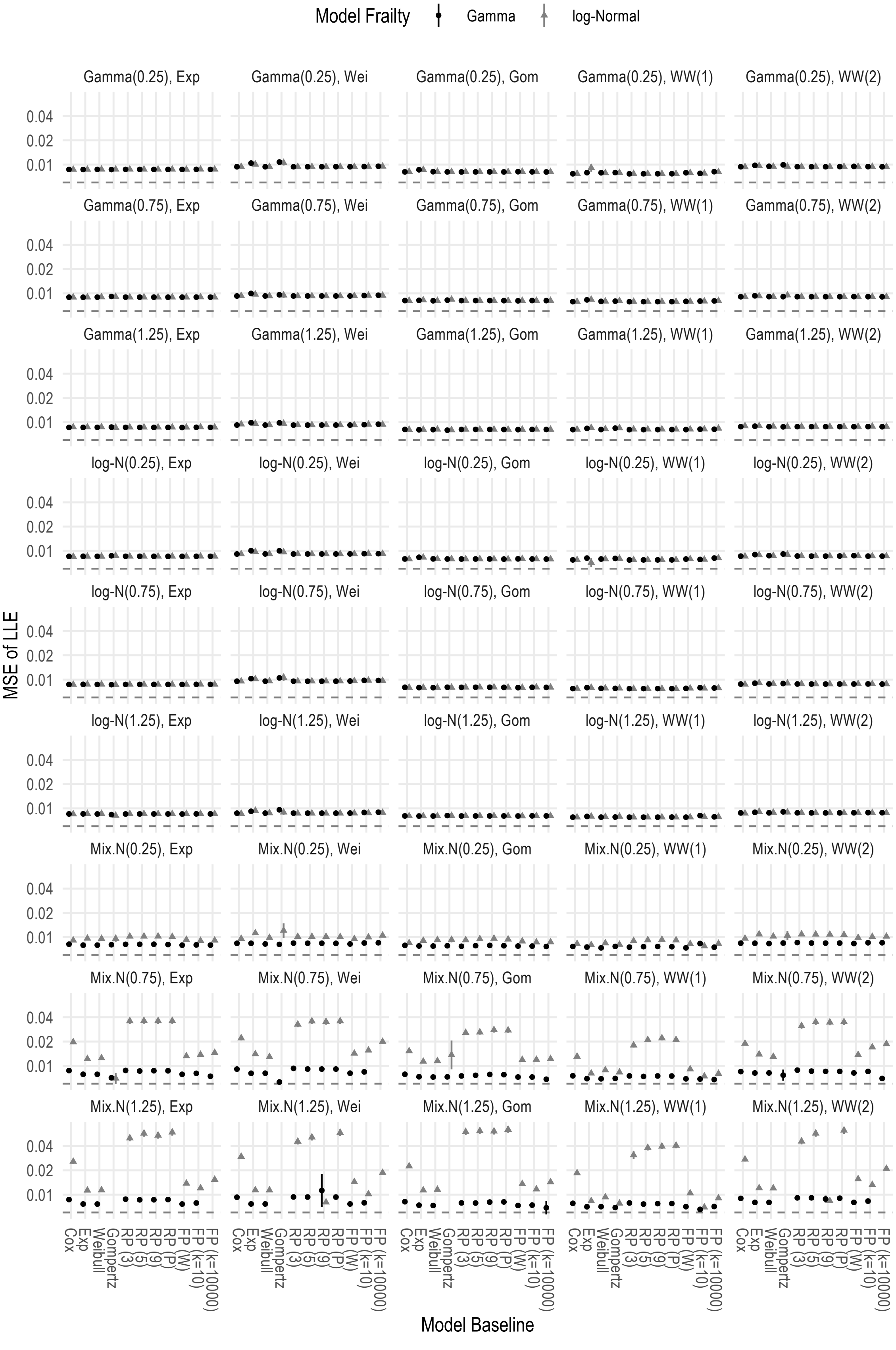


Figure E.4: MSE of LLE, scenarios with 750 clusters of 2 individuals each. Colours represent the model frailty, and each subplot includes results for a given combination of data-generating baseline hazard and frailty

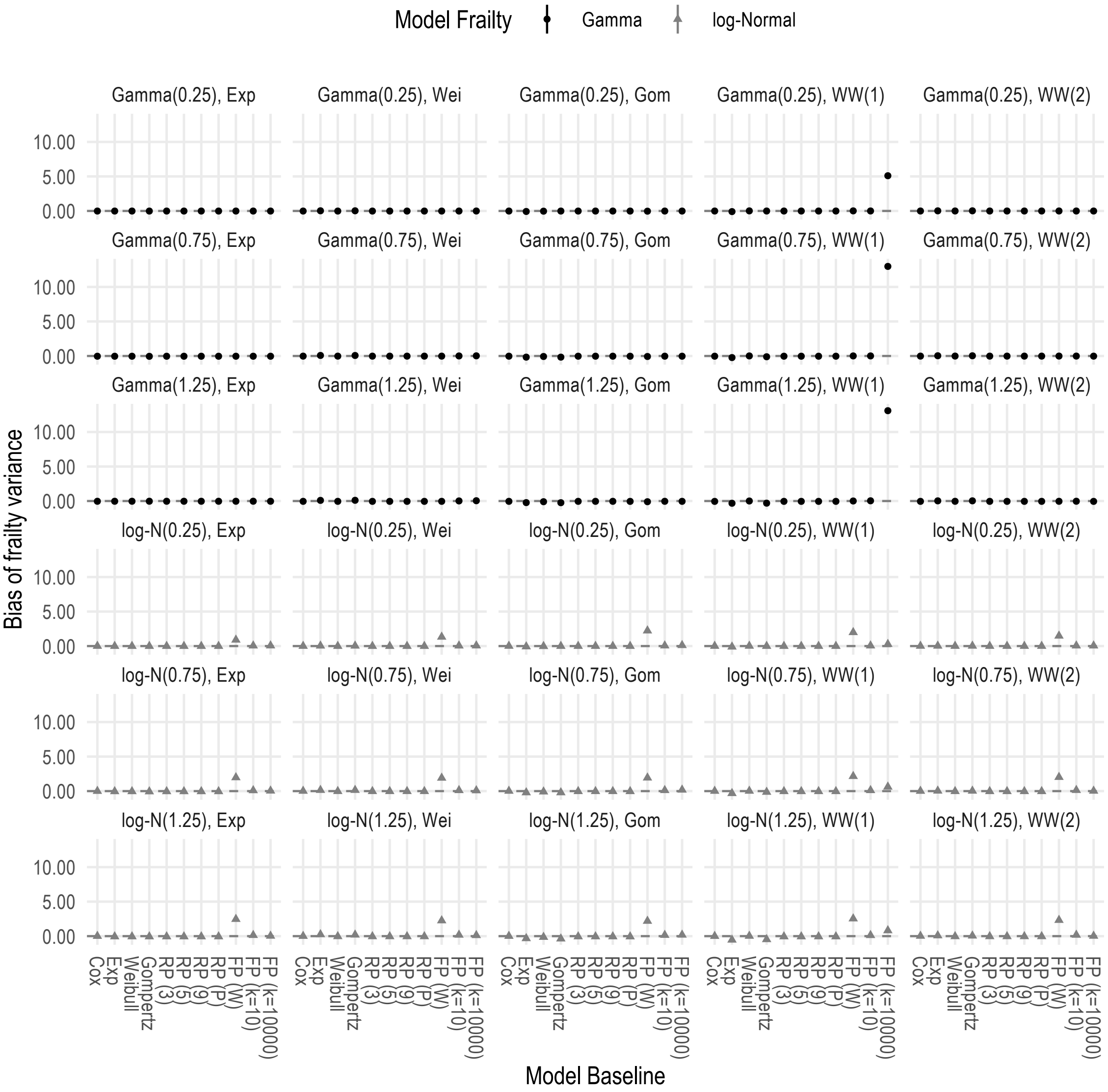


FIGURE E.5: Bias of frailty variance, scenarios with 20 clusters of 150 individuals each. Colours represent the model frailty, and each subplot includes results for a given combination of data-generating baseline hazard and frailty

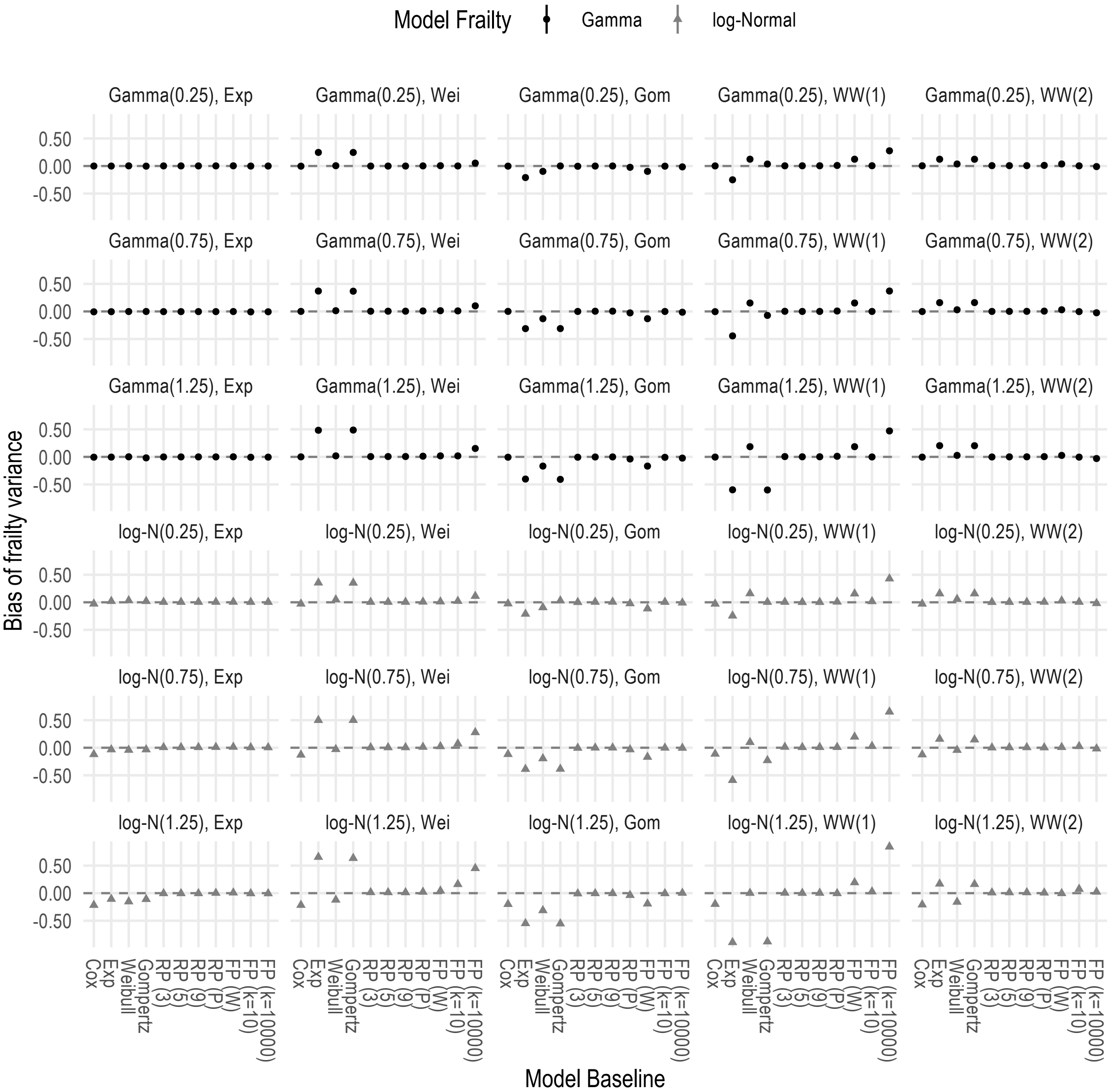


FIGURE E.6: Bias of frailty variance, scenarios with 750 clusters of 2 individuals each. Colours represent the model frailty, and each subplot includes results for a given combination of data-generating baseline hazard and frailty

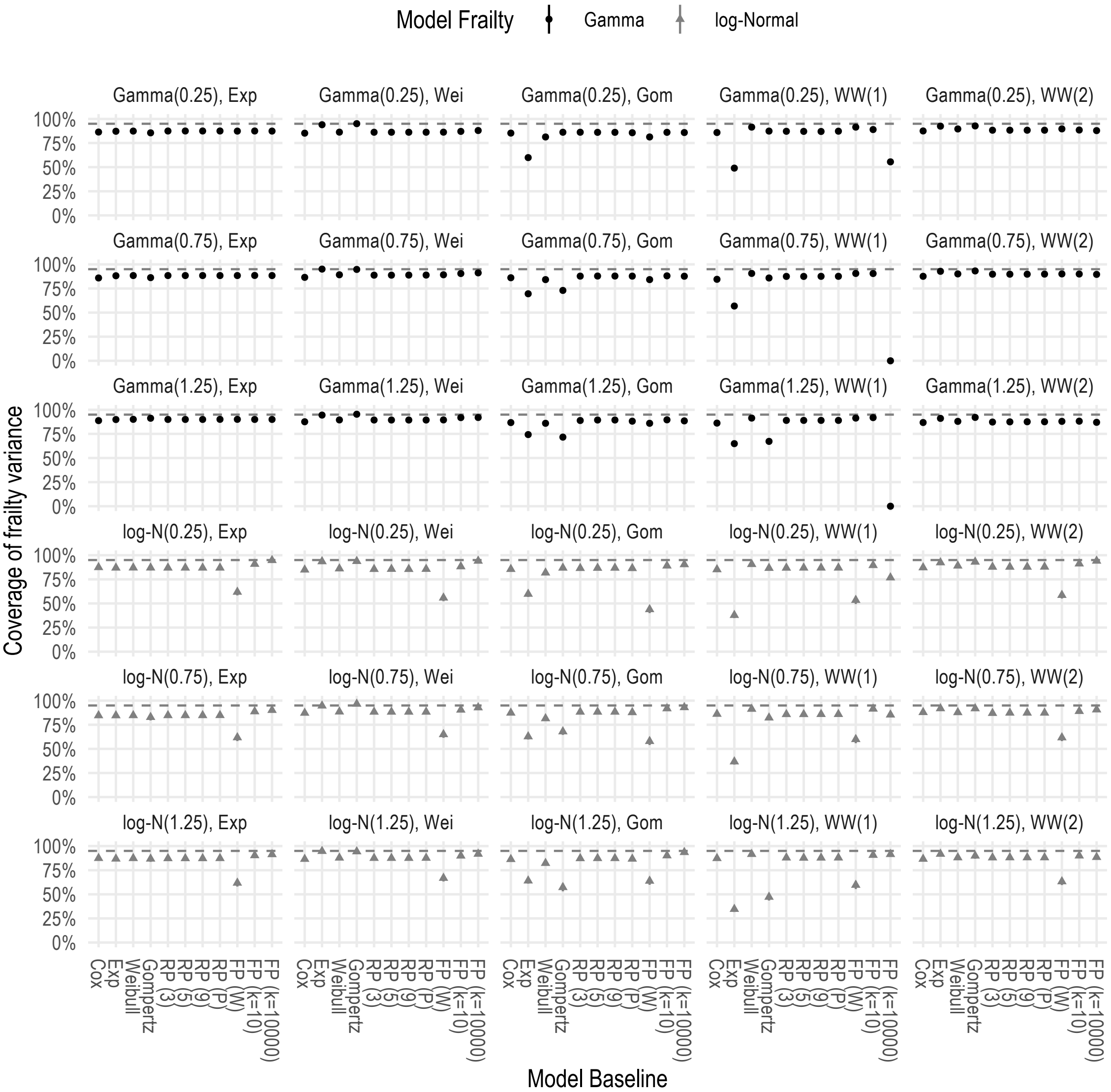


Figure E.7: Coverage of frailty variance, scenarios with 20 clusters of 150 individuals each. Colours represent the model frailty, and each subplot includes results for a given combination of data-generating baseline hazard and frailty

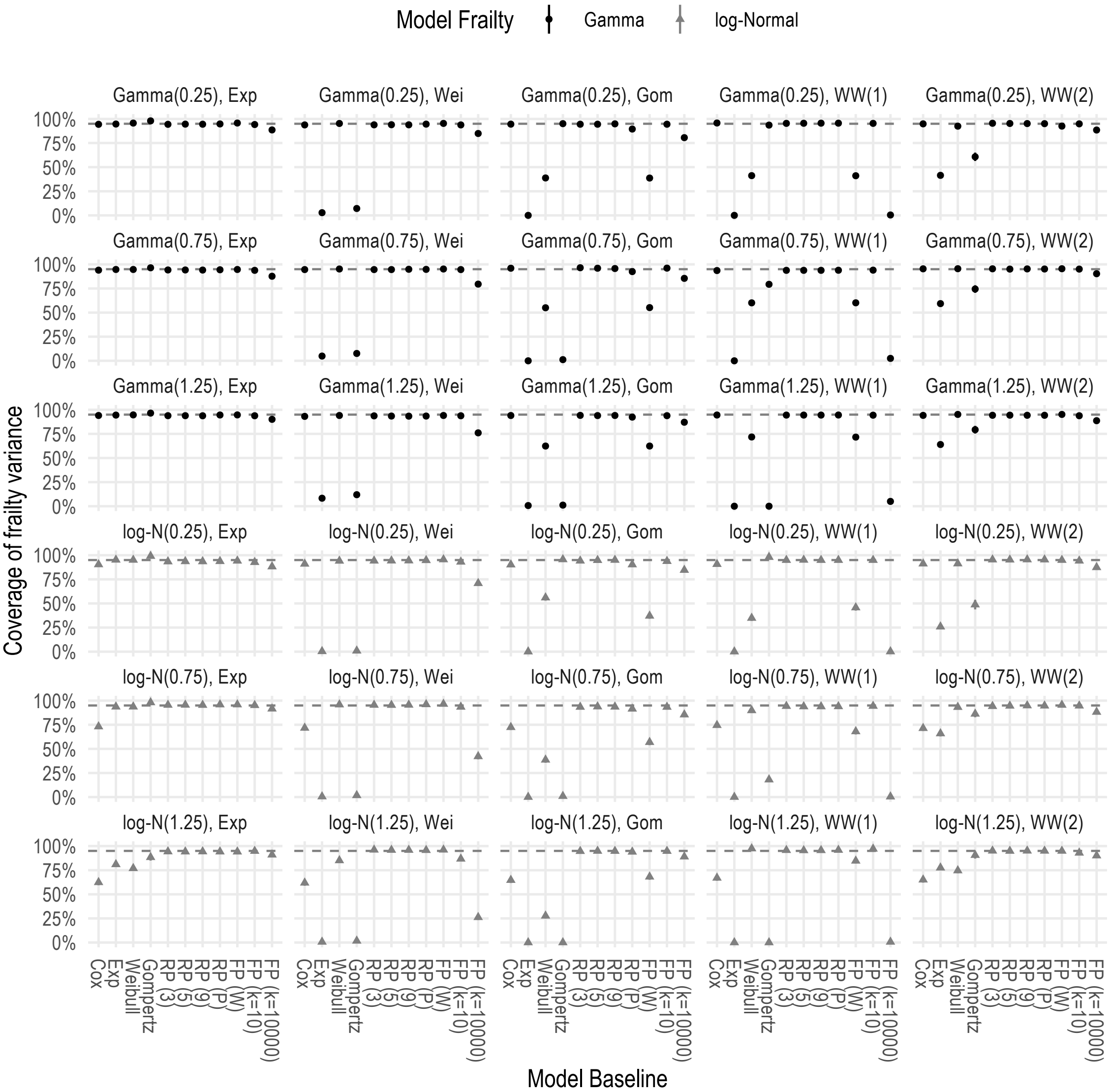


Figure E.8: Coverage of frailty variance, scenarios with 750 clusters of 2 individuals each. Colours represent the model frailty, and each subplot includes results for a given combination of data-generating baseline hazard and frailty

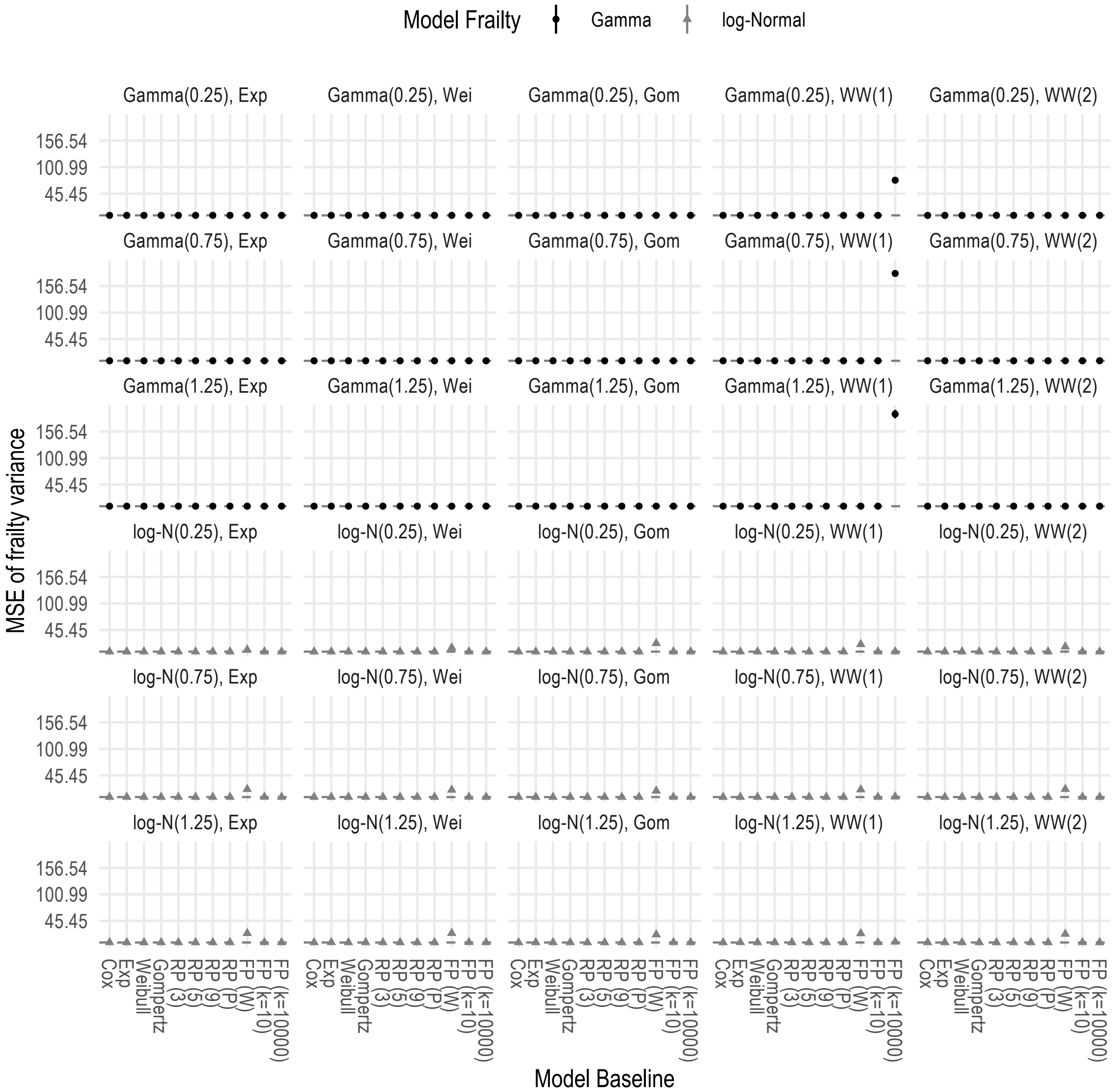


FIGURE E.9: MSE of frailty variance, scenarios with 20 clusters of 150 individuals each. Colours represent the model frailty, and each subplot includes results for a given combination of data-generating baseline hazard and frailty

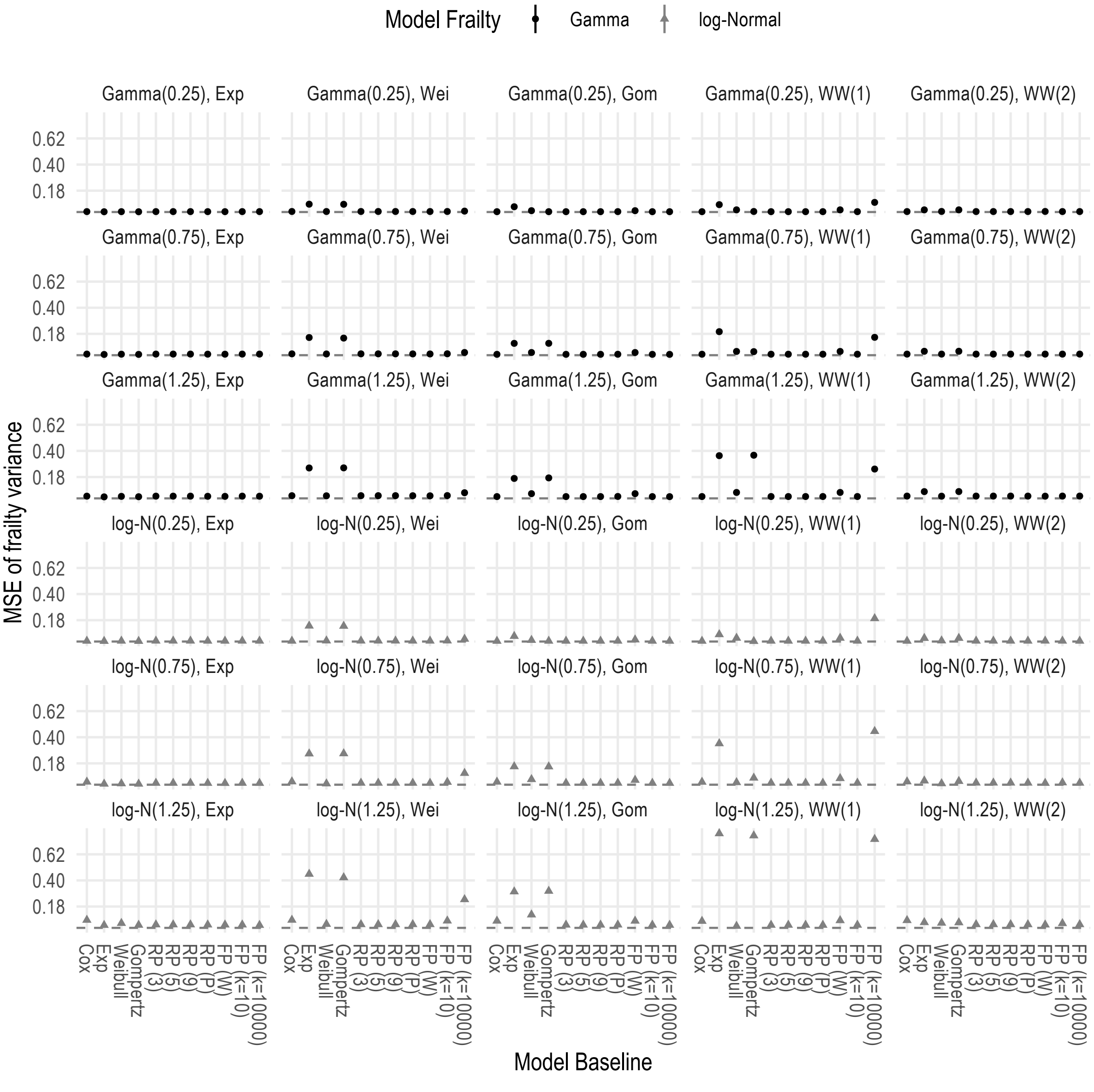


FIGURE E.10: MSE of frailty variance, scenarios with 750 clusters of 2 individuals each. Colours represent the model frailty, and each subplot includes results for a given combination of data-generating baseline hazard and frailty

F Research Paper: Mixed Effects Models for Healthcare Longitudinal Data with an Informative Visiting Process: A Monte Carlo Simulation Study

This Appendix contains the manuscript comparing methods to account for an informative observation process that have been proposed in the literature via Monte Carlo simulation. The manuscript has been published in Statistica Neerlandica featuring in a Special Issue on the 2018 Survival Analysis for Junior Researchers conference, and is included in this Appendix. The pre-print history for the manuscript is also available from arXiv (`https://arxiv.org/abs/1808.00419`).

AG, MJC, KRA, JKB, and MJS conceived and planned the study. AG developed all the code required for the simulation study, analysed its results, and interpreted them with critical contributions from all co-authors. AG wrote the manuscript, with input and feedback from all co-authors - especially during the revision process. RWM and NJB are the lead investigators of PSP-CKD, and provided access to data and constructive feedback on the applied example and interpretation of its results. Finally, the anonymous reviewers and the editorial team of Statistica Neerlandica vastly contributed to improving the manuscript throughout the peer review process.

Received: 1 August 2018 | Revised: 25 July 2019 | Accepted: 14 August 2019
DOI: 10.1111/stan.12188

SPECIAL ISSUE ARTICLE

WILEY

Mixed-effects models for health care longitudinal data with an informative visiting process: A Monte Carlo simulation study

Alessandro Gasparini[1] | **Keith R. Abrams**[1] | **Jessica K. Barrett**[2] | **Rupert W. Major**[1,3] | **Michael J. Sweeting**[1,4] | **Nigel J. Brunskill**[3,5] | **Michael J. Crowther**[1]

[1]Biostatistics Research Group, Department of Health Sciences, University of Leicester, Leicester, UK
[2]MRC Biostatistics Unit, University of Cambridge, Cambridge, UK
[3]Department of Nephrology, University Hospitals of Leicester NHS Trust, Leicester, UK
[4]Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK
[5]Department of Infection Immunity and Inflammation, University of Leicester, Leicester, UK

Correspondence
Alessandro Gasparini, Biostatistics Research Group, Department of Health Sciences, University of Leicester, Centre for Medicine, University Road, Leicester LE1 7RH, UK.
Email: ag475@leicester.ac.uk

Funding information
NIHR CLAHRC East Midlands and Kidney Research UK, Grant/Award Number: TF2/2015; MRC Unit

Electronic health records are being increasingly used in medical research to answer more relevant and detailed clinical questions; however, they pose new and significant methodological challenges. For instance, observation times are likely correlated with the underlying disease severity: Patients with worse conditions utilise health care more and may have worse biomarker values recorded. Traditional methods for analysing longitudinal data assume independence between observation times and disease severity; yet, with health care data, such assumptions unlikely hold. Through Monte Carlo simulation, we compare different analytical approaches proposed to account for an informative visiting process to assess whether they lead to unbiased results. Furthermore, we formalise a joint model for the observation process and the longitudinal outcome within an extended joint modelling framework. We illustrate our results using data from a pragmatic trial on enhanced care for individuals with chronic kidney disease, and we introduce user-friendly software that can be used

Programme, Grant/Award Number: MC_UU_00002/5; MRC New Investigator Research, Grant/Award Number: MR/P015433/1

to fit the joint model for the observation process and a longitudinal outcome.

KEYWORDS
electronic health records, informative visiting process, inverse intensity of visiting weighting, longitudinal data, mixed-effects models, Monte Carlo simulation, recurrent-events models, selection bias

1 | INTRODUCTION

The analysis of longitudinal data is essential to understand the evolution of disease and the effect of interventions over time. A source of longitudinally recorded data that is being used increasingly often in medical research is health care consumption data; that is, data sources that have been constructed by extracting and linking electronic health records from primary, specialist, and hospital care with other data sources such as nationwide registries for epidemiological surveillance. Several examples of cohorts constructed in such a way are emerging in a variety of medical fields: amongst others, kidney disease (Hemmelgarn et al., 2009; Runesson et al., 2016), cardiovascular disease (Denaxas et al., 2012), and end-of-life health care (Tanuseputro et al., 2015). Data cohorts constructed by extracting medical records have thousands—if not millions—of individuals with hundreds of measurements each: The availability to researchers of such vast amount of data allows answering more relevant and detailed clinical questions but poses new challenges. In terms of reporting, guidelines have emerged to improve discovery, transparency, and replicability of research finding utilising routinely collected data (Benchimol et al., 2015). In terms of methodological challenges, first and foremost, observation times are likely to be correlated with the underlying disease severity in health care consumption data sets. For instance, individuals tend to have irregular observation times as patients with more severe conditions (or showing early symptoms of a disease) tend to visit their doctor or go to the hospital more often than those with milder conditions (and no symptoms). Their worse disease status is also likely to be reflected in worse biomarkers being recorded at such visits, causing abnormal values of such biomarkers to be overrepresented and normal values to be underrepresented. Taking this pattern to the extreme, healthy individuals may not appear in health records at all, leading to cohort selection bias; this is a separate issue that is not dealt with in this manuscript.

Traditional methods used to analyse longitudinal data rely on the assumption that the underlying mechanism that controls the observation time is independent of disease severity; however, that is unlikely with health care consumption data. It can be shown that failing to account for informative dropout in a longitudinal study could yield biased estimates of the model parameters (McCulloch, Neuhaus, & Olin, 2016; Wu & Carroll, 1988), and so does näively applying traditional methods when the follow-up is irregular and related to the outcome (Pullenayegum & Lim, 2016). Despite the potential for bias, there is some evidence pointing toward a lack of awareness of the potential for bias in longitudinal studies with health care data irregularly collected over time: In a recent literature review on the topic, Farzanfar et al. (2017) showed that 86% of studies did not report enough information to evaluate whether the visiting process was informative or not, and only one study used a method capable of dealing with an informative observation process. This is concerning when the aim of a research project is aetiology.

Bias may arise when data on covariates and outcomes are collected at irregular, subject-specific intervals; in fact, when analysing data originating from electronic health records, data is collected only when study subjects consume health care (e.g., by visiting their doctor or going to the hospital). As a consequence, visit times are likely to be informative and to depend on the clinical history of an individual. The visiting process in this setting is therefore deemed to be informative (or dynamic, outcome dependent). The bias that one may encounter when the observation process is informative can be classified in two types: selection bias or confounding (Hernán, McAdams, McGrath, Lanoy, & Costagliola, 2009). Selection bias arises because of the selection of observed individuals only in the analysis. This bias is the same bias induced by informative censoring due to loss to follow-up (Hernán, Hernández-Díaz, & Robin, 2004): Censoring is the extreme case of an observation process where an individual is not observed ever again. Conversely, confounding arises when there are common causes of both the exposure and the outcome, for example, when the consequent visit times are decided by physicians or patients based on, for example, current health status, which itself is associated with the observed longitudinal outcome. Hernán et al. (2009) describe selection bias and confounding originating from dynamic observation processes more in detail, including directed acyclic graphs (DAGs) that illustrate the underlying causal mechanism.

In the past years, several methods have been developed to deal with longitudinal data terminated by informative dropout (Kurland, Johnson, Egleston, & Diehr, 2009); conversely, the problem of informative visit times has received considerably less attention. Despite that, a few methods emerged that can be broadly categorised in two families: methods based on inverse intensity of visit weighting (IIVW, an extension of inverse probability of treatment weighting [IPW]; Robins, Rotnitzky, & Zhao, 1995) and methods based on shared random effects (Liu, Huang, & O'Quigley, 2008). An introduction to the various methods is presented elsewhere (Pullenayegum & Lim, 2016). Nevertheless, to the best of our knowledge, there is only one comparison existing in the current literature that yielded negative results: Neuhaus et al. (2018) conclude that fitting ordinary linear mixed models disregarding the observation process yielded the smallest bias and showed that adding regular visits to the observation schedule (if possible) reduced that bias even further.

Throughout this paper, we focus on the problem of informative visiting process by assuming that the dropout process is not informative. First, we describe characteristics of the observation process and we define when it can be deemed informative in Section 2. Then, we introduce a joint model for the observation and longitudinal processes that can be easily extended within a multivariate generalised linear and nonlinear mixed-effects models framework (Crowther, 2017) in Section 3, and introduce the IIVW method in more detail in Section 4. We compare the performance of this model against other alternatives that have been introduced in the literature via Monte Carlo simulation in Section 5. Finally, we illustrate the use of the joint model using data from a pragmatic trial in chronic kidney disease (CKD) and discuss our conclusions in Sections 6 and 7, respectively.

2 | CHARACTERISTICS OF THE OBSERVATION PROCESS

An observation process can have regular or irregular visits. With regular visits, the jth visit time for the ith individual T_{ij} is the same for all individuals: $T_{ij} = t_j \; \forall \; i,j$, with $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, n_i$. Conversely, with irregular visits that is no longer true. With irregular visits, the observation process—denoted by the counting process $N_i(t)$—can be defined to be completely at random when visit times and outcomes are independent (Pullenayegum & Lim, 2016):

$$E[\Delta N_i(t) | \bar{Y}_i(\infty), \bar{X}_i(\infty)] = E[\Delta N_i(t)],$$

where $\Delta N_i(t) = N_i(t) - N_i(t^-)$, with t^- being the instant of time right before t. $\bar{Y}_i(\infty)$ and $\bar{X}_i(\infty)$ denote the values of outcome and covariates for any $t > 0$.

The observation process can be deemed informative when it is not completely at random, that is, when the condition above is not verified. In that case, it is possible to identify the following two scenarios.

- Observation process at random, when visiting at time t is independent of the outcome at time t given data recorded up to time t:

  $$E\left[\Delta N_i(t) | \bar{X}_i(t), \bar{N}_i(t^-), \bar{Y}_i^{\text{obs}}(t^-), Y_i(t)\right] = E\left[\Delta N_i(t) | \bar{X}_i^{\text{obs}}(t), \bar{N}_i(t^-), \bar{Y}_i^{\text{obs}}(t^-)\right],$$

  where $\bar{X}_i(t)$ and $\bar{X}_i^{\text{obs}}(t)$ denote the covariates history up to time t and its observed values, $\bar{N}_i(t^-)$ denotes the history of the observation process up to time t^-, and $\bar{Y}_i^{\text{obs}}(t^-)$ denotes the observed values of the outcome up to time t^-.
- Observation process not at random, where the definition of missing at random does not hold. That is, the scenario where visiting at time t is not independent of the outcome at time t, even after conditioning on data recorded up to time t:

  $$E\left[\Delta N_i(t) | \bar{X}_i(t), \bar{N}_i(t^-), \bar{Y}_i^{\text{obs}}(t^-), Y_i(t)\right] \neq E\left[\Delta N_i(t) | \bar{X}_i^{\text{obs}}(t), \bar{N}_i(t^-), \bar{Y}_i^{\text{obs}}(t^-)\right].$$

Gruger, Kay, and Schumacher (1991) illustrate four possible models that could be linked to the abovementioned scenarios.

1. The *examination at regular intervals* model, consisting of observation times that are predefined and equal for all patients. This scenario yields the so-called *balanced panel data*.
2. The *random sampling* model, consisting of a sampling scheme (e.g., an observation process) that is not predefined, but still independent of the disease history of the study subjects.
3. The *doctor's care* model, consisting of an observation process that depends on the characteristics of the patient at the moment of the current doctor's examination. For instance, a doctor could require stricter monitoring for subjects with more advanced disease status or with abnormal values of a biomarker.
4. The *patient self-selection* model, yielding observations that are triggered by the patients themselves. According to this model, patients may choose to visit their doctor when they feel unwell, or they may choose to skip a visit that was preplanned when they feel the treatment they are receiving is not beneficial to their health status. Unfortunately, the factors that cause patients to self-select themselves are generally unknown or not recorded.

Models (1) and (2) could be characterised as *observation completely at random*; model (3) could be characterised as *observation at random*; finally, model (4) could be characterised as *observation not at random*.

3 | A JOINT MODEL FOR THE OBSERVATION PROCESS AND A LONGITUDINAL OUTCOME

Let $D_{ij}(t) = I(T_{ij} = t)$ denote the presence of an observation at time t for the ith individual: At each $D_{ij}(t) = 1$, a new observation of the longitudinal outcome Y_{ij} is recorded. Let $\tilde{t}_{ij}$ be the gap time between the jth and $(j+1)$th measurement for the ith individual. Let $\tilde{d}_{ij}$ be the binary indicator variable that denotes whether the gap time $\tilde{t}_{ij}$ is observed (or not). In practice, gap times are always observed except when the observation process is censored at the end of follow-up, for example, the date when the data extraction occurs. Let z_{ij} be the covariate vector for the longitudinal outcome, and let w_i be the covariate vector for the observation process; z_{ij} and w_i do not necessarily overlap, and it is assumed that both could be extended to include time-dependent exogenous covariates (e.g., w_{ij}). We model the observation process and the repeated measures process using a joint longitudinal and survival model. Conditional on random effects u_i, the submodel for the time to each observation is a proportional hazards model with hazard for gap time $\tilde{t}_{ij}$:

$$r\left(\tilde{t}_{ij}|w_{ij}, u_i, \theta_t\right) = r_0(\tilde{t}_{ij})\exp(w_{ij}\beta + u_i), \tag{1}$$

where $\theta_t = \beta$. The submodel for the jth longitudinal observation for the ith individual is

$$(y_{ij}|D_{ij}(t) = 1, z_{ij}, u_i, v_i, \theta_y) = m_{ij} + \epsilon_{ij} = z_{ij}\alpha + \gamma u_i + v_i + \epsilon_{ij}, \tag{2}$$

where $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ and $\theta_y = \{\alpha, \gamma, \sigma_\epsilon^2\}$.

Equation (1) is a recurrent-events model for the observation process, with $r_0(\tilde{t}_{ij})$ any parametric or flexible parametric (Royston & Parmar, 2002) baseline hazard function (also referred to as baseline intensity—we use the terms hazard and intensity interchangeably throughout this manuscript). Equation (2) is a linear mixed model for the longitudinal outcome with a random intercept v_i. The two processes are linked together via the shared, individual-specific, random effect u_i. Including the γ parameter in the longitudinal model allows for an association between the two equations, association that will be estimated from data; when $\gamma = 0$, the two processes are independent of each other; that is, the observation process is not informative. Finally, we assume that the random effects follow a multivariate normal distribution with null mean vector and variance–covariance matrix $\Sigma_{u,v}$.

The model is fitted using maximum likelihood; the individual-specific contribution to the likelihood can be written as

$$\begin{aligned} L_i(\theta) &= \int p\left(\tilde{t}_{ij}, \tilde{d}_{ij}, y_{ij}, b_i; \theta\right) db_i \\ &= \int \prod_{j=1}^{n_i} p\left(\tilde{t}_{ij}, \tilde{d}_{ij}|b_i, \theta_t\right) p(y_{ij}|b_i, \theta_y) p(b_i|\theta_b) db_i, \end{aligned}$$

where $\theta = \{\theta_t, \theta_y, \theta_b\}$ is the overall parameters vector, $b_i = \{u_i, v_i\}$ is the vector of random effects,

$$p\left(\tilde{t}_{ij}, \tilde{d}_{ij}|b_i, \theta_t\right) = r\left(\tilde{t}_{ij}|w_{ij}, u_i, \theta_t\right)^{\tilde{d}_{ij}} \exp\left(-\int_0^{\tilde{t}_{ij}} r(s|w_{ij}, u_i, \theta_t) ds\right)$$

is the contribution to the likelihood of the time to the jth observation in individual i,

$$p(y_{ij}|b_i, \theta_y) = \left(2\pi\sigma_\epsilon^2\right)^{-1/2} \exp\left(-\frac{(y_{ij} - m_{ij})^2}{2\sigma_\epsilon^2}\right)$$

is the contribution of the jth longitudinal observation, and $p(b_i|\theta_b)$ is the density of the random effects. The likelihood does not have a closed form, as it is necessary to integrate out the distribution of the random effects; methods such as Gaussian quadrature and Monte Carlo integration can be used for that purpose (Pinheiro & Bates, 1995).

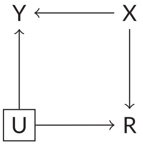


FIGURE 1 Simplified directed acyclic graph depicting a joint model for a longitudinal outcome and its observation process

A simplified DAG that illustrates how the joint model accounts for the correlation between a longitudinal outcome *Y* and its observation process *R* is included as Figure 1 (Liu, Zheng, & Kang, 2018); *X* represents covariates included in the model, and *U* represents the shared random effects. After adjusting for all covariates (e.g., confounders) *X*, the longitudinal outcome and the observation process are associated only through the shared *U*. However, when estimating the joint model, we assume a distribution for *U* (e.g., Gaussian) and we integrate it out of the marginal likelihood, blocking the path between *Y* and *R*. Therefore, for the joint model to be valid, the observation process has to be at least *at random*, according to the definition of Section 2.

This model is nested within a wide family of multivariate generalised linear and nonlinear mixed-effects models (Crowther, 2017). The model presented in this section can easily be extended to multiple random effects (potentially nested within each other), to different parametric and flexible parametric baseline hazard formulations for the recurrent-events model, and to include other outcomes (e.g., a dropout process, or a second longitudinal outcome); we focus on the model formulated in this section for simplicity. Finally, this joint model (and several extensions) can be easily fitted in Stata using the user-written command `merlin` (Crowther, 2018). We produce example code that is included in the Online Supplementary Material.

4 | INVERSE INTENSITY OF VISIT WEIGHTING

The bias induced by an informative observation process can be adjusted by using the IIVW method first proposed by Robins et al. (1995) as an extension of the IPW method (Cole & Hernán, 2008). This method was further developed by Bååžková and Lumley (2007), and there are a few examples of this method applied in practice (Bååžková, Brown, & John-Stewart, 2010; Van Ness, Allore, Fried, & Lin, 2009). The IIVW approach accommodates an informative observation process in a marginal regression model by weighting each observation by the inverse of the probability of each measurement to be recorded. This approach creates a pseudopopulation in which the observation process is static and can be ignored. The weights can be estimated by fitting a regression model including all covariates that inform the observation process and further stabilised to increase efficiency (Cole & Hernán, 2008). The weighting model could include current and past values of any covariate that may affect the visiting process; however, as with IPW, all covariates that might be related to the observation process should be included in the weighting model; otherwise, bias will incur.

The approach we illustrate follows from Van Ness et al. (2009). The model used to estimate weights is an Andersen-Gill recurrent-events model (Andersen & Gill, 1982) for the observation process, assuming a gap-time scale (as described in Section 3):

$$r\left(\tilde{t}_{ij}\right) = r_0\left(\tilde{t}_{ij}\right)\exp(z_i\eta),$$

where $\tilde{t}$ are gap times between consecutive observations, $r_i(\tilde{t})$ is the intensity of visit for individual i at gap time $\tilde{t}$, $r_0(\tilde{t})$ is the unspecified baseline intensity at gap time $\tilde{t}$, and z_i is a vector of coefficients that are assumed to accurately describe the observation process for individual i. η is a vector of regression coefficients that is estimated using the Cox partial likelihood method and a robust jack-knife estimator for the variance of the regression coefficients. The inverse intensity of visit weights are estimated by taking the inverse of the linear predictor $\exp(z_i\hat{\eta})$ at each time point, and further normalised by subtracting the mean inverse weight and adding the value 1 to each weight; the distribution of the weights is therefore centred on the value 1. Finally, two further adjustments are needed. First, because the last data entry for each individual represents the end of follow-up of the study, each weight is shifted by one time point. Second, given that each individual is observed at least once (i.e., at baseline), a weight of one is assigned to the first observation of each individual.

The marginal model for the longitudinal outcome is then fit using generalised estimating equations and including the normalised inverse intensity of visit weights as probability weights in the model. The model has the form

$$E(y_{ij}) = \alpha_0 + Z_i\alpha_1 + t_{ij}\alpha_2,$$

and can be fit using readily available statistical software. We use the Stata command `glm`.

5 | A MONTE CARLO SIMULATION STUDY

Aim

We design a simulation study aimed to assess the impact of ignoring the observation process in longitudinal mixed-effects models when the observation process is informative.

Data-generating mechanisms

We simulate data from the following joint model:

$$\begin{gathered} r(\tilde{t}) = r_0(\tilde{t})\exp(Z_i\beta + u_i) \\ y_{ij}|(D_{ij}(t) = 1) = \alpha_0 + Z_i\alpha_1 + t_{ij}\alpha_2 + \gamma u_i + v_i + \epsilon_{ij}. \end{gathered}$$

Z_i is a time-invariant covariate (for simplicity) representing a binary treatment, simulated from a Bernoulli random variable with probability 0.5: $Z_i \sim \text{Bern}(1, 0.5)$. The coefficient associated to the treatment variable is $\beta = 1$ for the observation process; $\alpha_1 = 1$ for the longitudinal process. The fixed intercept of the longitudinal model is $\alpha_0 = 0$, and the fixed effect of time is $\alpha_2 = 0.2$. The random effects u_i and v_i are simulated from a Normal random variable with null mean and variance $\sigma_u^2 = 1$ and $\sigma_v^2 = 0.5$, respectively. The residual error of the longitudinal model is assumed to follow a Normal distribution with null mean and variance $\sigma_\epsilon^2 = 1$. We assume independence between the random effects and the residual variance, and between random effects (i.e., $\Sigma_{u,v}$ is a diagonal matrix with $\text{diag}(\Sigma_{u,v}) = \{\sigma_u^2, \sigma_v^2\}$). We assume independent random effects for simplicity, but we show in the Online Supplementary Material how to fit a joint model with correlated random effects. The joint model with correlated random effects can be thought of as a reparameterisation of the joint model with independent random effects, where the association parameter γ is related to the correlation between the two random effects in the bivariate version. The baseline hazard from the recurrent visit process is assumed to follow a Weibull distribution with shape parameter $p = 1.05$; we vary the scale parameter λ and, therefore, the baseline intensity of the visiting process, with $\lambda = \{0.10, 0.30, 1.00\}$. These baseline intensities along with the value of β correspond to an expected median gap time between observations of 5.83 and 2.25 years for unexposed and exposed individuals if $\lambda = 0.10$, 2.05, and 0.79 years if $\lambda = 0.30$, and 0.65 and 0.25 years if $\lambda = 1.00$, respectively. Each observation time is simulated using the inversion method of Bender, Augustin, and Blettner (2005), assuming a gap-time scale (where the time index is reset to zero after the occurrence of each observation; the resulting recurrent-events model is then a semi-Markov model). We vary the association parameter γ between the two submodels, with $\gamma = \{0.00, 1.50\}$; we expect all models to perform similarly when $\gamma = 0$, that is, when the longitudinal process is independent of the observation process.

In addition to simulating data from the joint model above, we generate the observation process by drawing from a Gamma distribution. Specifically, we draw the observation times from a Gamma distribution with shape = 2.00 and scale:

$$\exp(-\psi \beta Z_i + \xi_i),$$

where ξ_i is simulated from a Normal distribution with null mean and variance $\sigma_\xi^2 = 0.1$. Z_i is the same binary treatment covariate as before, with the same associated parameter $\beta = 1$. The value of ψ defines the association between the observation, for example, when $\psi = 0$, the observation process is not informative; we set $\psi = \{0.00, 2.00\}$. We also simulate a scenario where the observation process depends on treatment and on previous values of the longitudinal outcome Y. In this setting, we draw observation times from a Gamma distribution with shape = 2.00 and scale

$$\exp(-\psi \beta Z_i + \omega y_{i,j-1} + \xi_i)$$

for the jth observation time of the ith individual, with $\psi = 2.00$ and $\omega = 0.20$. Finally, we simulate a scenario from a joint model to which we add regular (i.e., planned) visits every year, as suggested by Neuhaus et al. (2018). We simulate this scenario from the abovementioned joint model, and we set $\gamma = 3.00$ and $\lambda = 0.05$ to obtain an observation process that is sparse and strongly associated with the longitudinal outcome.

We simulate 200 study individuals under each data-generating mechanism and the recurrent observation process continues for each individual until the occurrence of administrative censoring, which we simulated from a Unif(5, 10) random variable.

We define the last gap time for each individual as the difference between the last observation and the censoring time.

Estimands

The main estimand of interest is the vector of regression coefficients $\alpha = \{\alpha_0, \alpha_1, \alpha_2\}$, with specific focus on the treatment effect α_1. In the Online Supplementary Material, we also report on the estimated association parameter γ and on the estimated variance of the random effects and the residual errors: σ_u^2, σ_v^2, and σ_ϵ^2.

Methods

We fit five competing models to each simulated data set:

1. Model A, the joint model described above (at the beginning of the "Data-generating mechanisms" section) and corresponding to the true data-generating mechanisms when simulating data from a joint model;

2. Model B, a linear mixed model including the number of visits (centred on the mean value) as a fixed effect in the model;
3. Model C, a linear mixed model including the cumulative number of visits as a fixed effect in the model;
4. Model D, a linear mixed model that disregards the observation process completely;
5. Model E, a marginal model fitted using generalised estimating equations and inverse intensity of visit weights.

Model A is fit using `merlin` (Crowther, 2018) and `gsem` in Stata. Model B follows from previous work by Goldstein, Bhavsar, Phelan, and Pencina (2016), where they demonstrate that, conditioning on the number of health care encounters, it is possible to remove bias due to an informative observation process (they denote this bias as “informed presence bias”). We therefore include the number of observations per individual, centred on the mean value, in a mixed-effects model for the longitudinal outcome:

$$y_{ij} = \alpha_0 + Z_i\alpha_1 + t_{ij}\alpha_2 + n_i^c\alpha_3 + \nu_i + \epsilon_{ij},$$

with ν_i a random intercept and n_i^c the number of observations for the ith individual. Model C is analogous to Model B, adjusting for the cumulative number of measurements up to time j instead, denoted as $\bar{n}_{it_j}$:

$$y_{ij} = \alpha_0 + Z_i\alpha_1 + t_{ij}\alpha_2 + \bar{n}_{it_j}\alpha_3 + \nu_i + \epsilon_{ij}.$$

Model D is analogous to Models B and C, assuming $\alpha_3 = 0$. Models B, C, and D are fit using the `mixed` command in Stata. Models A, B, C, D are fit assuming an independent structure for the variance–covariance matrix of the random effects. Finally, Model E is fitted following the two-stage procedure presented in Van Ness et al. (2009) and illustrated in Section 4.

Performance measures

We will assess average estimates and standard errors, empirical standard errors, bias, and coverage probability of $\hat{\alpha}_m$, with $m = \{0, 1, 2\}$. However, the main performance measures of interest are bias and coverage probability: the former quantifies whether an estimator targets the true value on average, whereas the latter represents the proportion of times that a confidence interval based on $\hat{\alpha}_{m,k}$ and $\hat{\text{SE}}(\hat{\alpha}_{m,k})$ contains the true value α_m, with k indexing each replication. We compute and report Monte Carlo standard errors to quantify the uncertainty in estimating bias and coverage (Morris, White, & Crowther, 2019). If we assume that $\text{Var}(\hat{\alpha}_m) \leq 0.1$ (or, equivalently, $\text{SE}(\hat{\alpha}_m) \leq 0.32$) and we require a Monte Carlo standard error for bias of 0.01 or lower, given that $\text{MCSE(Bias)} = \sqrt{\text{Var}(\hat{\alpha}_m)/K}$, we would require a number of replications K = 1,000. The assumed standard error is larger than the standard errors reported by Liu et al. (2008) for a model similar to Model A. The expected Monte Carlo standard error for coverage, assuming a worst-case scenario of coverage = 0.50, would be 0.02, which we deem acceptable. Therefore, we proceed by simulating 1,000 independent data sets for this simulation study.

Software

The simulation study is coded and run using Stata version 15, built-in functions (such as `mixed`, `glm`, `gsem`), and the user-written commands `survsim` (Crowther & Lambert, 2012) and `merlin` (Crowther, 2018); results of the simulation study are summarised using R (R Core Team, 2019) and the R package `rsimsum` (Gasparini, 2018). All the codes required to simulate data, fit each model, and produce summary tables and figures are publicly available on the GitHub page of the first author (https://github.com/ellessenne/infobsmcsim).

Results

We focus on results for the estimated treatment effect α_1, which are depicted in Figure 2. Tabulated values are included in the Online Supplementary Material, alongside results for the other regression coefficients α_0, α_2, estimated variances of the random effects, summaries for the association parameter γ, and convergence rates of each model under each data-generating mechanism.

Descriptive results Each simulated data set had 200 distinct individuals; summary descriptive statistics for each data-generating mechanism are included in Table 1. The median sample size per simulated data set varied between 666 and 4,482, the median number of measurements per individual varied between 2 and 13, and the median gap time between observations varied between 0.13 and 1.31 (years). In simulated scenarios from the joint model, as expected, a higher baseline intensity of visit process yielded more frequent measurements and a larger number of measurements overall; values were not affected by the association parameter.

Results for noninformative observation processes When the observation process was not informative, all models estimated regression coefficients with null to negligible bias. Coverage probability of the regression coefficients was also optimal, with slight undercoverage for the intercept term α_0 and the treatment effect α_1 for estimates originating from the IIVW model. Mean squared errors were similar across the range of scenarios with a noninformative observation process. Bias for the variance of the residual error term was null to negligible as well, with good coverage. Conversely, the variability of the random intercept ν was estimated with slight negative bias from all models, with subpar coverage (between 90% and 95 %); this is expected as we use maximum likelihood and not restricted maximum likelihood. Finally, the estimated variance of the random effect linking the two outcomes in the joint model was positively biased with coverage of approximately 75%; the magnitude of bias decreased as the baseline intensity λ increased.

Results for informative observation processes When generating data from a Γ distribution depending on treatment only, all models were able to estimate the regression coefficients with no bias, optimal coverage probability, and comparable mean squared errors. Conversely, in all other scenarios, the models performed quite differently. In the scenario with observation times simulated from a Γ distribution depending on treatment and previous values of the longitudinal outcome, all models but Model B (adjusting for the number of measurements) could estimate the treatment effect with null or minimal bias; Model B overestimated the treatment effect. The same pattern was observed for coverage of the treatment effect, with Model B undercovering, and for the mean squared errors. The effect of time was estimated with small bias and good coverage from all models, with Model E (IIVW model) performing slightly worst; mean squared errors were comparable. In scenarios simulated from a joint model, as expected, the joint model (Model A) performed best overall, with minimal to no bias, optimal coverage, and the lowest mean squared errors. Model C (adjusting for the cumulative number of measurements) and Model D (plain mixed model) overestimated the intercept term and underestimated the treatment effect whilst showing small bias when estimating the effect of time. Interestingly, both

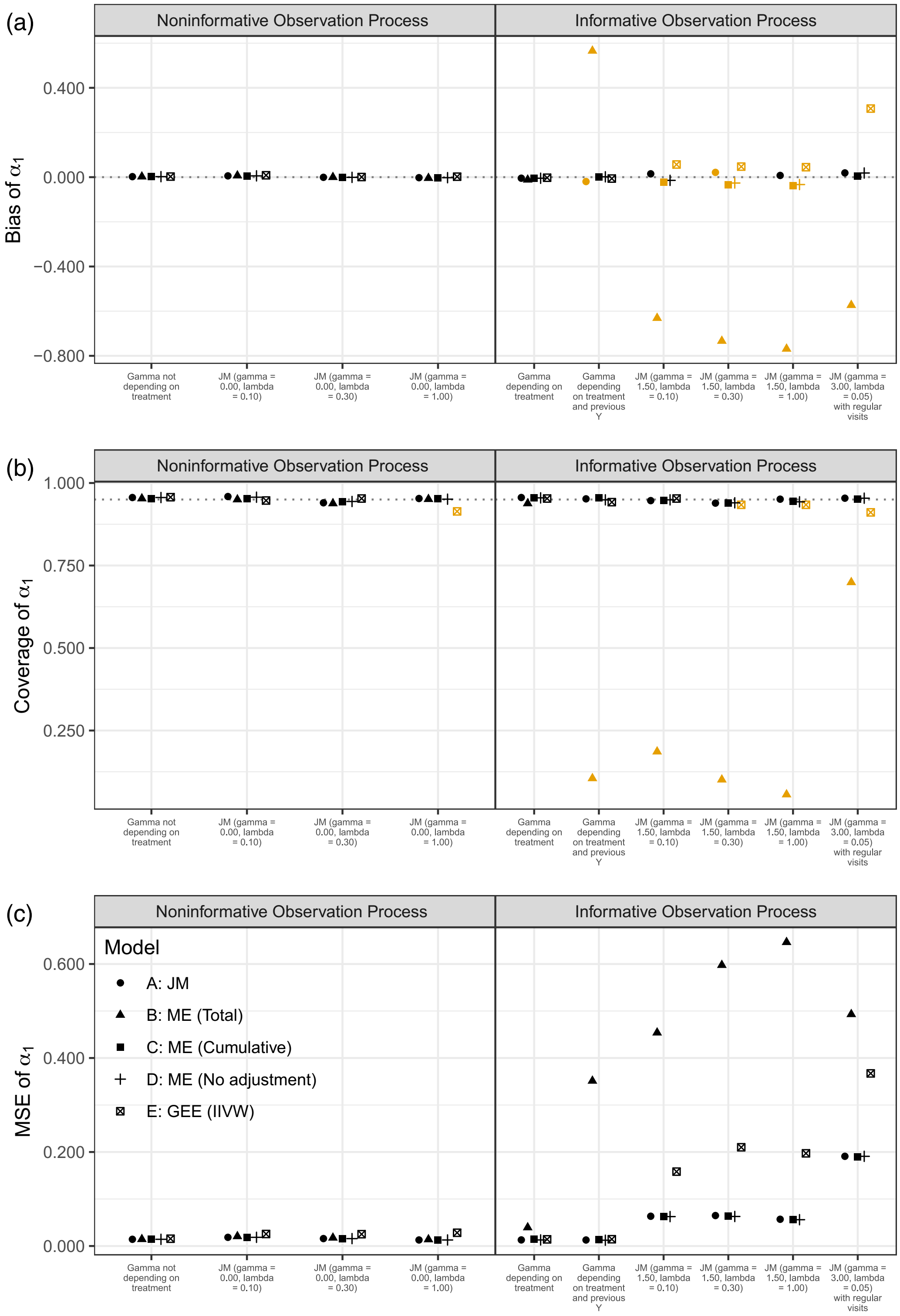


FIGURE 2 Bias (a), coverage (b), and mean squared error (c) of the estimated treatment effect α_1. The orange colour identifies scenarios where the summary statistics were significantly different than the target value (0 for bias, 95% for coverage) using Z-tests based on estimated Monte Carlo standard errors

TABLE 1 Summary characteristics of simulated data under each data-generating mechanism

Data-generating mechanism	Sample size	No. of measurements	Gap time
Γ distribution not depending on treatment	938 (918–957)	4 (3–6)	1.31 (0.74–2.17)
JM ($\gamma = 0.00, \lambda = 0.10$)	666 (634–705)	2 (1–4)	0.91 (0.33–2.12)
JM ($\gamma = 0.00, \lambda = 0.30$)	1,564 (1,475–1,667)	5 (2–9)	0.37 (0.13–0.94)
JM ($\gamma = 0.00, \lambda = 1.00$)	4,489 (4,188–4,815)	13 (6–27)	0.13 (0.04–0.33)
Γ distribution depending on treatment	3,444 (3,296–3,606)	11 (4–28)	0.23 (0.12–0.41)
Γ distribution depending on Y treatment and previous	2,564 (2,457–2,670)	9 (4–20)	0.31 (0.16–0.60)
JM ($\gamma = 1.50, \lambda = 0.10$)	669 (637–707)	2 (1–4)	0.90 (0.33–2.11)
JM ($\gamma = 1.50, \lambda = 0.30$)	1,556 (1,461–1,654)	5 (2–9)	0.37 (0.13–0.94)
JM ($\gamma = 1.50, \lambda = 1.00$)	4,482 (4,218–4,794)	13 (6–26)	0.13 (0.04–0.33)
JM ($\gamma = 3.00, \lambda = 0.05$) with regular visits	1,842 (1,818–1,867)	9 (7–10)	1.00 (1.00–1.00)

Note. Values are median with interquartile interval.

models showed that the bias when estimating the effect of time decreased as the baseline intensity λ increased: as expected, including more measurements allows to better estimate the effect of time. Model B performed worst when estimating the effect of treatment, with large negative bias. It also yielded biased intercept and effect of time; however, as with Models C and D, bias for the estimate of time decreased as more measurements were available. Finally, Model E slightly overestimated the effect of treatment. Model E showed increasing bias when estimating the intercept as the visiting process was denser, whilst (analogously as with Models B, C, and D) showing less biased estimates of the effect of time as the baseline intensity increased. All models with the largest biases showed also poor coverage and the largest standard errors. Overall, in settings simulated from a joint model, Model B and Model E performed worse and showed the largest biases. In the scenario simulated from a joint model with a sparse observation process and regular yearly visits, the joint model (Model A) and the plain mixed model (Model D) performed best, managing to recover the true values of all regression coefficients with no bias, and optimal coverage probabilities and mean squared errors. Model B managed to estimate the effect of time with small bias, but largely overestimated the intercept and underestimated the treatment effect. Model C managed to estimate the intercept and the treatment effect with small or no bias, but severely underestimated the effect of time. Coverage and mean squared errors followed the same pattern.

Results for the association parameter γ The estimating procedure worked well when the two submodels were not associated. For instance, there was no bias, coverage probabilities were optimal, and mean squared errors were small, irrespectively of the baseline intensity of visit λ. Conversely, when the submodels were associated ($\gamma = 1.50$), the estimated association parameter was slightly negatively biased (−0.11 to −0.06), with suboptimal coverage (75% to 83%). Mean squared error decreased when the baseline intensity of visit increased. Finally, the scenario simulated from a joint model with a strong association parameter $\gamma = 3.00$ and regular visits showed the worst performance, with large negative bias (−3.7289), poor coverage, and large mean squared error. Including regular visits caused γ to shrink toward the null, with a median estimate of −0.7289.

Convergence rates Convergence rates for all models included in this comparison were generally good. All models showed a perfect convergence rate of 100% except the joint model, which showed a lower convergence rate of 96% and 99% in two simulated scenarios, both with an informative observation process. However, the remaining scenarios showed a perfect convergence rate for the joint model as well.

6 | APPLICATION

We fit the models included in this comparison to data obtained from the *Primary–Secondary Care Partnership to Prevent Adverse Outcomes in Chronic Kidney Disease* (PSP-CKD) study (ClinicalTrials.gov Identifier: NCT01688141; Major et al., 2019). PSP-CKD is a cluster-randomised controlled pragmatic trial of enhanced CKD care against usual primary care management. From the Nene Clinical Commissioning Group, Northamptonshire, UK, 49 primary care practices were randomised to either enhanced care or usual care; informed consent was provided at the practice level. Adult individuals with CKD were identified from each practice by using a research version of the web-based CKD management and audit tool IMPAKT (available at http://www.impakt.org.uk/); all data were anonymised prior to removal from the primary care practice. Individuals were included if a recorded estimated glomerular filtration rate (eGFR) below 60 ml/min/1.73 m^2 was found during 5 years before the date of randomisation; eGFR was estimated using the Modification of Diet in Renal Disease (MDRD) equation (Levey et al.,).

We extracted baseline data (collected retrospectively at the date of randomisation and up to 5 years prior) from the PSP-CKD study consisting of all longitudinal eGFR measurements recorded during routine visits to the practices prior to randomisation; we also extracted the gender of each participant. This resulted in 239,468 eGFR measurements for 36,527 individuals, of which 14,268 (39%) were males and the remaining 22,259 (61%) were females. The median gap time between observations was 0.35 years (129 days), with interquartile interval of 0.11 – 0.74 years (39 – 272 days). We aim to evaluate whether the longitudinal eGFR trajectory before randomisation to treatment differs between males and females.

We start by evaluating whether the visiting process could be informative. First, we computed Spearman's rank correlation between gap time and gender: ($\rho = 0.01$). The correlation coefficient was significantly different than zero. Second, we fitted a linear mixed model for gap time versus gender with a random intercept and a random gender effect, and we found a significant association, as females had an 8.56-day-longer gap time (95% CI: 5.58 – 11.54). Finally, fitting the Andersen-Gill model for the observation process as described in Section 4 with gender as the only covariate included in the model yielded a hazard ratio of 0.9589 (with 95% C.I.: 0.9398 – 0.9783) for females compared with males. In conclusion, we found the gap time to be associated with gender; hence, we deem the visiting process to likely be informative.

We fit the models included in the comparison, with gender as the binary exposure variable. The joint model included gender as the only covariate in the observation process submodel, and so did the recurrent-events model utilised to fit weights for the IIVW model.

The estimated coefficients for the longitudinal trajectory from each model are presented in Figure 3. The marginal model estimated an intercept and gender effect significantly different than the other four models: Specifically, the estimated intercept from the marginal model was approximately two units lower, and the effect of gender was approximately seven times higher and statistically significant, compared to a nonstatistically significant effect of gender estimated by the remaining models. The estimated effect of time was similar between all models (approximately -0.70 per unit of time), with the exception of the mixed model adjusting for the cumulative number of measurements as a time-varying covariate (estimated effect of approximately −0.60). The interaction between gender and time was similarly estimated by all models, ranging between 0.4679 and 0.5158, and was statistically significant. This showed that females had a slower decline in renal function over time compared to men. The estimated coefficient for the observation process from the joint model shows a reduced risk of having a measured value for females compared to males (approximately 6%, hazard ratio of 0.9417 with 95% CI: 0.9245 – 0.9589). This value, jointly with the estimated value of the association parameter γ (-3.8018, 95% CI: -3.9943 to -3.6092), seem to confirm that the observation process is informed by gender.

Overall, all models estimated a similar longitudinal trajectory (Figure 4), with the IIVW model being the exception. We saw in the results of our simulations in Section 5 that the IIVW model yielded biased results for the exposure and the intercept of the longitudinal model under a variety of scenarios, and we observe this difference in our applied setting as well. Interestingly, all other models performed similarly, even the mixed model adjusting for the total number of measurements; our simulations showed that the effect of a binary exposure was estimated with bias, but we did not saw this difference in practice.

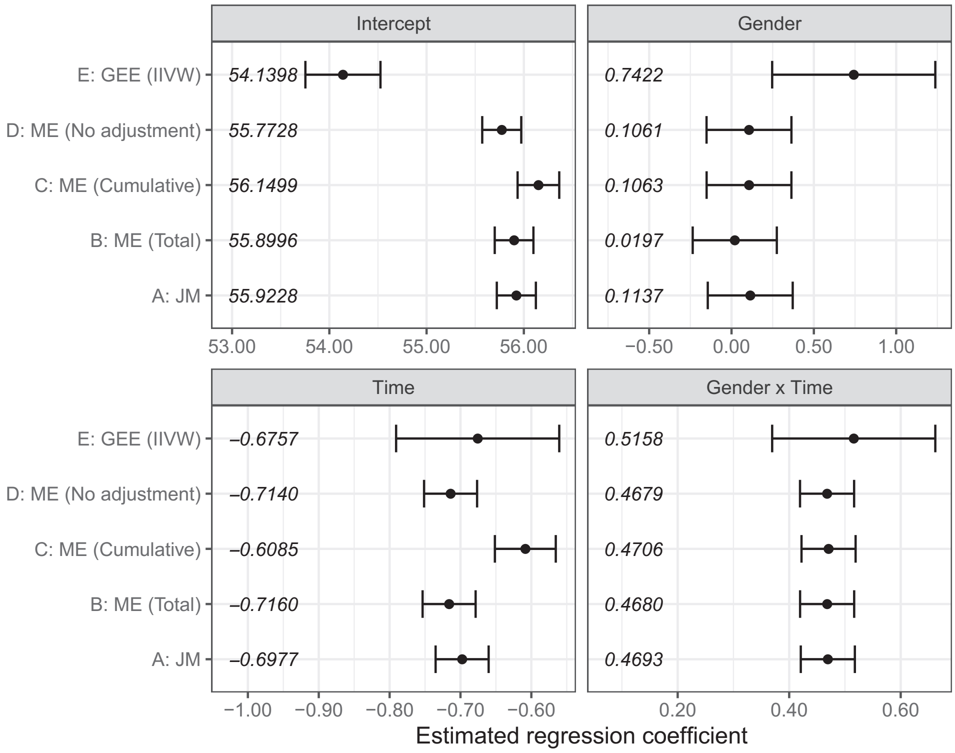


FIGURE 3 Forest plot with estimated coefficients for the longitudinal component, models fit to the application data from the PSP-CKD study. Each estimated coefficient is included as text placed on the leftmost side of each subplot. PSP-CKD = Primary–Secondary Care Partnership to Prevent Adverse Outcomes in Chronic Kidney Disease

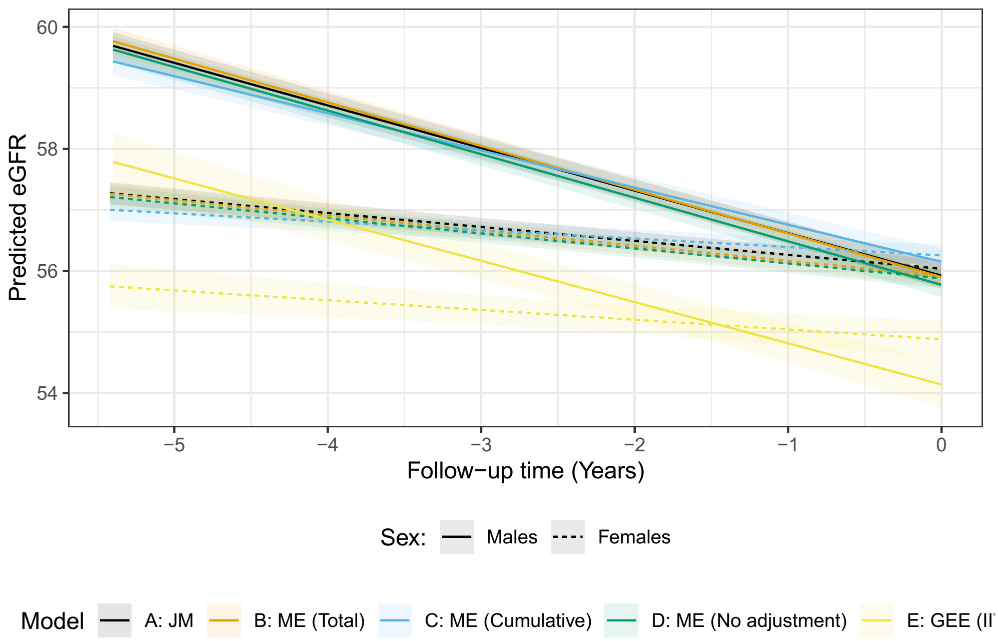


FIGURE 4 Predicted longitudinal trajectories from the models fit to the application data from the PSP-CKD study. The solid lines represent estimated trajectories for males, whereas the dashed lines represent trajectories for females. Colours identify the model. PSP-CKD = Primary–Secondary Care Partnership to Prevent Adverse Outcomes in Chronic Kidney Disease; eGFR = estimated glomerular filtration rate

7 | DISCUSSION

In this article, we formalise the problem of informative visiting process within a framework of multivariate generalised linear and nonlinear mixed-effects models, including causal considerations. Via Monte Carlo simulation, we illustrate (1) how ignoring an informative visiting process leads to biased estimates of the regression coefficient of a longitudinal model and (2) we compare some of the methods that have been proposed in the literature to account for it. To the best of our knowledge, there is only one comparison currently in the literature (Neuhaus et al., 2018), albeit they include different models in their comparison and simulate an informative observation process differently by first generating a grid of potential observation times and then relating the probability of being observed to a given functional form of current (or lagged) covariates. They also do not include a joint model analogous to the model introduced in our manuscript in Section 3 in their comparison.

As expected, the joint model that accounts for the informative observation process by modelling it via a recurrent-events survival model performed best. Interestingly, the mixed-effects model that disregarded completely the observation process performed worse than the joint model, but outperformed other methods; the inflation in the variance of the random intercept of the plain mixed model seemed to capture part (if not most) of the variability due to the observation process, although this result needs to be thoroughly tested in more complex scenarios (e.g., with random effects of time, etc.). The mixed models adjusting for the total number of measurements or the cumulative number of measurements (as a time-varying covariate) performed worst, and we would not recommend their usage in practice in these settings; this finding contrasts the findings of Goldstein et al. (2016), although their settings were quite different than ours. Further to that, they acknowledged the potential for collider bias (due to conditioning on a collider, the number of measurements) when the phenotyping algorithm for determining the exposure has high sensitivity; indeed, in our settings, the sensitivity is perfect as there is no misspecification of the exposure. An additional possible explanation could be that, in our settings, the model adjusting for the total number of measurements is in fact conditioning on the future, as the total number of observations is not determined at the beginning of the study. This may be explaining the poor performance of this method in the settings of our simulations. The performance of the marginal model fitted using generalised estimating equations and inverse intensity of visit weights laid between the plain mixed model and the remaining mixed models; furthermore, its performance seemed to improve when the observation pattern became denser, except for the intercept term α_0. This pattern was generally observed throughout all scenarios and models, as the performance seemed to increase with more frequent observation patterns; this finding is consistent with Hernán et al. (2009). The results of our simulations are consistent with those of Neuhaus et al. (2018): The IIVW approach showed bias in all the settings of their simulation where the observation process was informative, even when adding regular visits to the study. To compute the weights of the IIVW approach, applied researchers need to correctly specify the model for the visit process, a challenging task, especially when not all the information required to fit the correctly specified model is observed (or known). We also observed that the IIVW model performed quite differently than the other methods in our applied example, although the observed difference does not seem to be clinically relevant.

Most importantly, our simulations show that, under the null, all the approaches compared in this study produce unbiased estimates of the regression coefficients, the implication being that overmodelling the observation process does not seem to introduce bias in the analysis. In settings where it is not clear whether the observation process is informative or not, fitting the joint model would provide applied researchers with a method for estimating (and testing) the association between the two outcomes: This could be especially useful, for example, as a sensitivity analysis of standard mixed-effects models.

The joint model for the observation process and a longitudinal outcome that we described in Section 3 can be further extended. For instance, additional random effects could be introduced in the model to account for, say, heterogeneity in the trajectory of the longitudinal outcome over time. The functional form of the effect of time (both fixed and random) could also be generalised by using fractional polynomials or splines; the longitudinal trajectories need to be modelled appropriately and best fit could be assessed via information criteria such as the Akaike information criterion and Bayesian information criterion. In fact, in the applied example of Section 6, we assumed a linear effect of time on eGFR for simplicity; in actual applied projects, one should assess whether the final model is correctly specified. One could also extend the model to account for time-varying treatments, in both the observation process and longitudinal outcome submodels. That would however require further investigations to assess the performance of the joint model in those settings.

We assumed the treatment to be constant over time for simplicity, but in real-life settings, individuals are likely to start and drop treatment when deemed necessary by their treating physician. We assumed the baseline hazard of the recurrent-events model for the observation process to follow a Weibull distribution: This assumption could be further relaxed, and one could assume any parametric function, or even use flexible, spline-based formulations (e.g., Royston & Parmar, 2002). Additionally, for diseases with a high mortality rate, a terminal event that truncates observation of the longitudinal process is likely to be informative in the sense that it likely correlates with disease severity. That is, dropout is likely to be informative as the tendency to drop out after the occurrence of a terminal event is related to the current level of the longitudinally recorded biomarker. The proposed model could be easily extended to include a third equation with a time-to-event submodel for the dropout process, as in Liu et al. (2008). All of these extensions can be fit within the general framework of Crowther (2017) using the Stata command `merlin`. Finally, we could explore the association structure between the two submodels. For instance, we could reverse the association structure and include γ in the observation submodel: In that setting, assuming a positive association, higher values of the longitudinal process would lead to a more frequent visiting process (and vice-versa in the setting of negative association). The observation process could also depend on lagged values of the longitudinal outcome or of the exposure; this would relax the semi-Markov assumption in some of our data-generating mechanisms. More biologically (and clinically), plausible association structures (such as the current value, current slope, cumulative effect parametrisations) could also be investigated; more details are in Rizopoulos (2012).

In conclusion, it is important to account for the visiting process when analysing health care utilisation data, and we showed that ignoring it leads to biased estimates. Given the wide range of applied settings in which this could be relevant, the review of Farzanfar et al. (2017) points toward a lack of awareness of the problem and the lack of readily available, user-friendly software to fit more complex joint models; throughout this paper, we outlined a framework in which `merlin` could be easily used to fit complex joint model and help to reduce this translational gap. We provide example code using Stata in the Online Supplementary Material.

ACKNOWLEDGMENTS

The authors would like to thank all primary care practices and Nene Clinical Commissioning Group for participating in the PSP-CKD study. The PSP-CKD study was funded by the National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care (CLAHRC) East Midlands. Ongoing support for the study is funded by NIHR CLAHRC East Midlands and Kidney Research UK (Grant TF2/2015). JKB is supported by the MRC Unit Programme MC_UU_00002/5. MJC is partially supported by a MRC New Investigator Research Grant (MR/P015433/1).

ORCID

Alessandro Gasparini https://orcid.org/0000-0002-8319-7624

REFERENCES

Andersen, P. K., & Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *The Annals of Statistics*, *10*, 1100–1120.

Benchimol, E. I., Smeeth, L., Guttmann, A., Harron, K., Moher, D., Petersen, I., … RECORD Working Committee. (2015). The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement, *12*, *PLOS MEDICINE*, e1001885.

Bender, R., Augustin, T., & Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*, *24*, 1713–1723.

Bůžková, P., Brown, E. R., & John-Stewart, G. C. (2010). Longitudinal data analysis for generalized linear models under participant-driven informative follow-up: An application in maternal health epidemiology. *American Journal of Epidemiology*, *171*, 189–197.

Bůžková, P., & Lumley, T. (2007). Longitudinal data analysis for generalized linear models with follow-up dependent on outcome-related variables. *The Canadian Journal of Statistics*, *35*, 485–500.

Cole, S. R., & Hernán, M. A. (2008). Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, *168*, 656–664.

Crowther, M. J. (2017). Extended multivariate generalised linear and non-linear mixed effects models. arXiv preprint arXiv:1710.02223. https://arxiv.org/abs/1710.02223

Crowther, M. J. (2018). merlin-a unified modelling framework for data analysis and methods development in Stata. arXiv preprint arXiv:1806.01615. https://arxiv.org/abs/1806.01615

Crowther, M. J., & Lambert, P. C. (2012). Simulating complex survival data. *The Stata Journal*, *12*, 674–687.

Denaxas, S. C., George, J., Herrett, E., Shah, A. D., Kalra, D., Hingorani, A., … Hemingway, H. (2012). Data resource profile: Cardiovascular disease research using linked bespoke studies and electronic health records (CALIBER). *International Journal of Epidemiology*, *41*, 1625–1638.

Farzanfar, D., Abumuamar, A., Kim, J., Sirotich, E., Wang, Y., & Pullenayegum, E. M. (2017). Longitudinal studies that use data collected as part of usual care risk reporting biased results: A systematic review. *BMC Medical Research Methodology*, *17*, 133.

Gasparini, A. (2018). rsimsum: Summarise results from Monte Carlo simulation studies. *The Journal of Open Source Software*, *3*, 739.

Goldstein, B. A., Bhavsar, N. A., Phelan, M., & Pencina, M. J. (2016). Controlling for informed presence bias due to the number of health encounters in an electronic health record. *American Journal of Epidemiology*, *184*, 847–855.

Gruger, J., Kay, R., & Schumacher, M. (1991). The validity of inferences based on incomplete observations in disease state models. *Biometrics*, *47*, 595–605.

Hemmelgarn, B. R., Clement, F., Manns, B. J., Klarenbach, S., James, M. T., Ravani, P., … Tonelli, M. (2009). Overview of the Alberta kidney disease network. *BMC Nephrology*, *10*, 30.

Hernán, M. A., Hernández-Díaz, S., & Robins, J. M. (2004). A structural approach to selection bias. *Epidemiology*, *15*, 615–625.

Hernán, M. A., McAdams, M., McGrath, N., Lanoy, E., & Costagliola, D. (2009). Observation plans in longitudinal studies with time-varying treatments. *Statistical Methods in Medical Research*, *18*, 27–52.

Kurland, B. F., Johnson, L. L., Egleston, B. L., & Diehr, P. H. (2009). Longitudinal data with follow-up truncated by death: Match the analysis method to research aims. *Statistical Science*, *24*, 211.

Levey, A. S., Bosch, J. P., Lewis, J. B., Greene, T., Rogers, N., & Roth, D. (1999). A more accurate method to estimate glomerular filtration rate from serum creatinine: A new prediction equation. *Annals of Internal Medicine*, *130*, 461–470.

Liu, L., Huang, X., & O'Quigley, J. (2008). Analysis of longitudinal data in the presence of informative observational times and a dependent terminal event, with application to medical cost data. *Biometrics*, *64*, 950–958.

Liu, L., Zheng, C., & Kang, J. (2018). Exploring causality mechanism in the joint analysis of longitudinal and survival data. *Statistics in Medicine*, *37*, 3733–3744.

Major, R. W., Brown, C., Shepherd, D., Rogers, S., Pickering, W., Warwick, G. L., … Brunskill, N. J. (2019). The primary-secondary care partnership to improve outcomes in chronic kidney disease (PSP-CKD) study: A cluster randomized trial in primary care. *Journal of the American Society of Nephrology*, *30*, 1261–1270.

McCulloch, C. E., Neuhaus, J. M., & Olin, R. L. (2016). Biased and unbiased estimation in longitudinal studies with informative visit processes. *Biometrics*, *72*, 1315–1324.

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*, 2074–2102.

Neuhaus, J. M., McCulloch, C. E., & Boylan, R. D. (2018). Analysis of longitudinal data from outcome-dependent visit processes: Failure of proposed methods in realistic settings and potential improvements. *Statistics in Medicine*, *37*, 4457–4471. https://doi.org/10.1002

Pinheiro, J. C., & Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, *4*, 12–35.

Pullenayegum, E. M., & Lim, L. S. (2016). Longitudinal data subject to irregular observation: A review of methods with a focus on visit processes, assumptions, and study design. *Statistical Methods in Medical Research*, *25*, 2992–3014.

R Core Team (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data: With applications in R*. Boca Raton, FL: Chapman and Hall/CRC.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, *90*, 106–121.

Royston, P., & Parmar, M. K. (2002). Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine*, *21*, 2175–2197.

Runesson, B., Gasparini, A., Qureshi, A. R., Norin, O., Evans, M., Barany, P., … Carrero, J. J. (2016). The Stockholm CREAtinine Measurements (SCREAM) project: Protocol overview and regional representativeness. *Clinical Kidney Journal*, *9*, 119–127.

Tanuseputro, P., Wodchis, W. P., Fowler, R., Walker, P., Bai, Y. Q., Bronskill, S. E., & Manuel, D. (2015). The health care cost of dying: A population-based retrospective cohort study of the last year of life in Ontario, Canada, *PLOS ONE*, *10*, e0121759.

Van Ness, P. H., Allore, H. G., Fried, T. R., & Lin, H. (2009). Inverse intensity weighting in generalized linear models as an option for analyzing longitudinal data with triggered observations. *American Journal of Epidemiology*, *171*, 105–112.

Wu, M. C., & Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, *44*, 175–188.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Gasparini A, Abrams KR, Barrett JK, et al. Mixed-effects models for health care longitudinal data with an informative visiting process: A Monte Carlo simulation study. *Statistica Neerlandica*. 2019;1–19. https://doi.org/10.1111/stan.12188

G Supplementary Results: Monte Carlo Simulation on Modelling the Observation Process

This Appendix contains supplementary results from the Monte Carlo simulation on modelling the observation process when analysing longitudinal data, introduced in Section 7.5.

MSEs for the regression coefficients are included in Figure G.1.

Bias, coverage probabilities, and MSEs for all variance components are included in Figures G.2, G.3, and G.4, respectively.

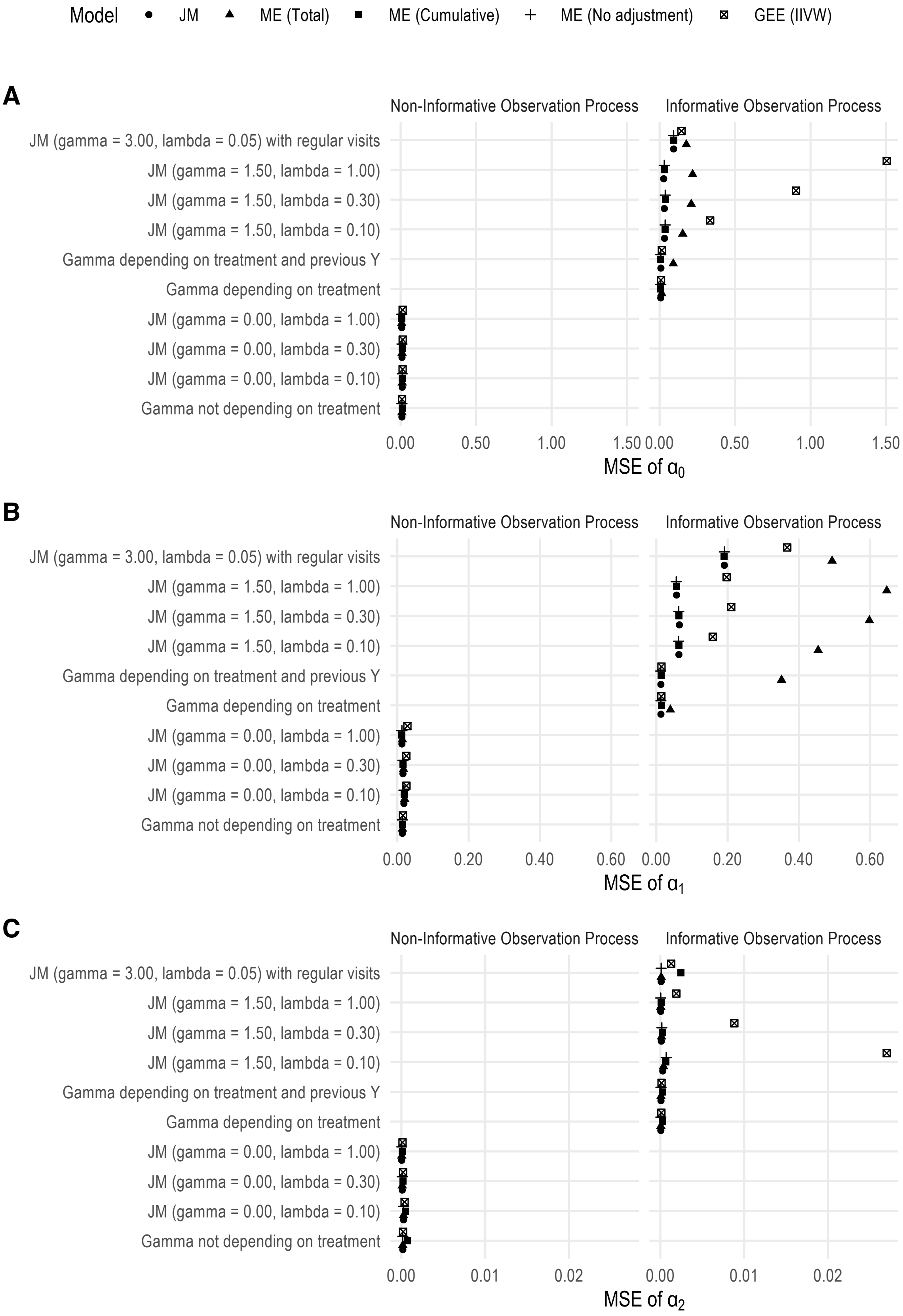


FIGURE G.1: MSEs of regression coefficients, Monte Carlo simulation study on modelling the observation process

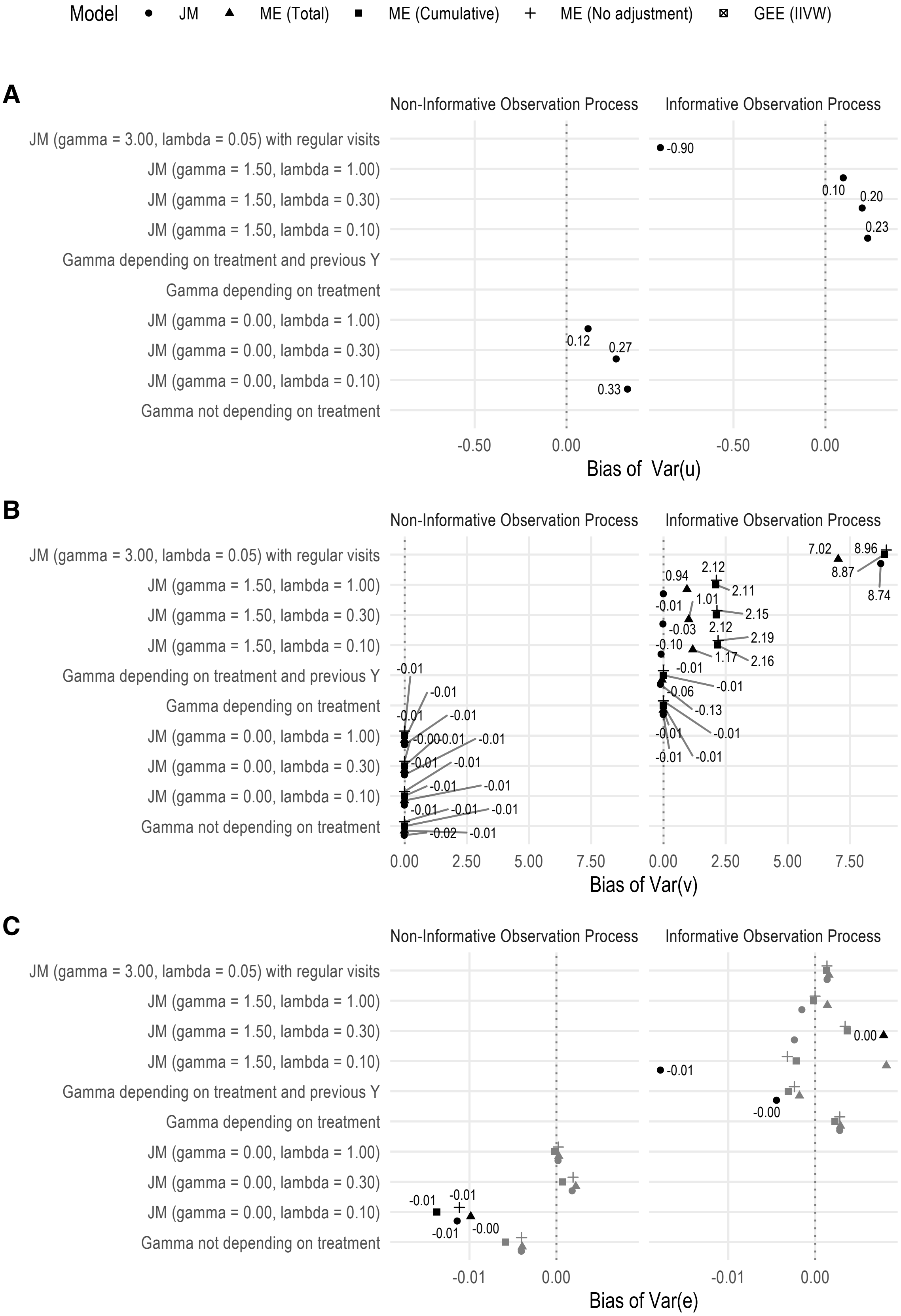


Figure G.2: Bias of variance components, Monte Carlo simulation study on modelling the observation process. Labelled values (with points in black) are statistically significant values, determined via Z tests based on Monte Carlo standard errors

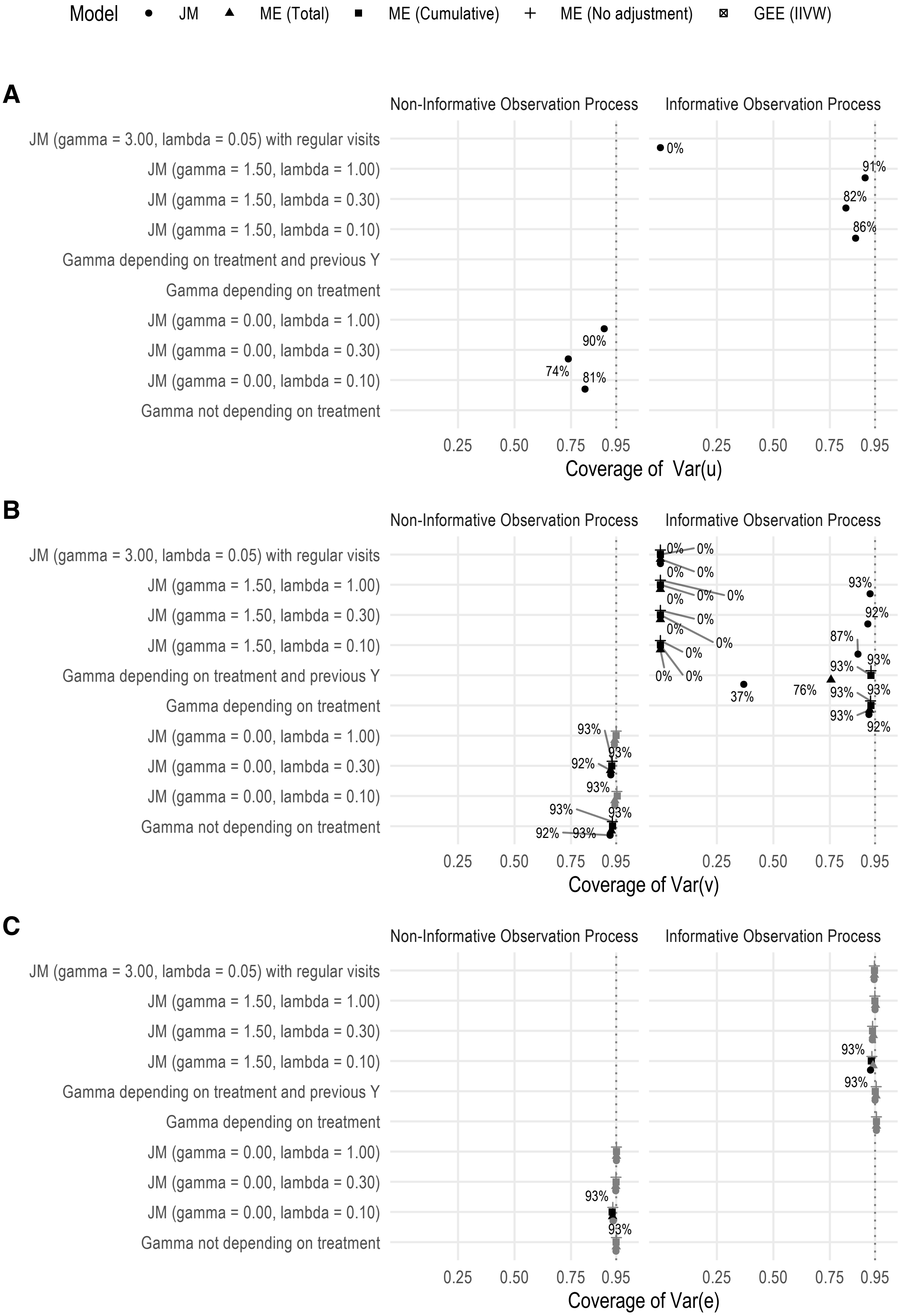


Figure G.3: Coverage probability of variance components, Monte Carlo simulation study on modelling the observation process. Labelled values (with points in black) are statistically significant values, determined via Z tests based on Monte Carlo standard errors

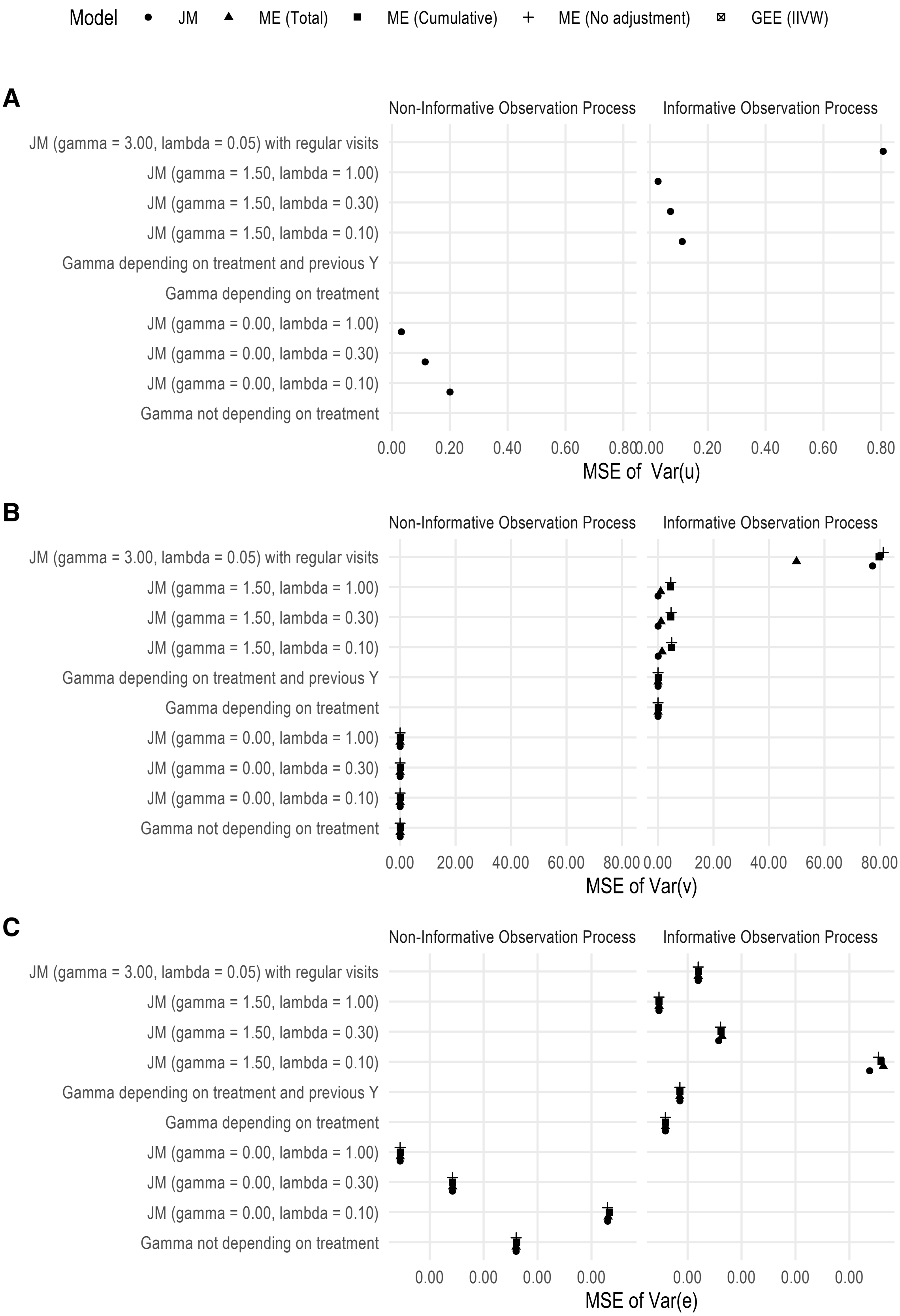


Figure G.4: MSEs of variance components, Monte Carlo simulation study on modelling the observation process

Bibliography

[1] Spiros C Denaxas, Julie George, Emily Herrett, Anoop D Shah, Dipak Kalra, Aroon D Hingorani, Mika Kivimaki, Adam D Timmis, Liam Smeeth, and Harry Hemingway. Data resource profile: cardiovascular disease research using linked bespoke studies and electronic health records (CALIBER). *International Journal of Epidemiology*, 41:1625–1638, 2012.

[2] Björn Runesson, Alessandro Gasparini, Abdul Rashid Qureshi, Olof Norin, Marie Evans, Peter Barany, Björn Wettermark, Carl Gustaf Elinder, and Juan Jesús Carrero. The Stockholm CREAtinine Measurements (SCREAM) project: protocol overview and regional representativeness. *Clinical Kidney Journal*, 9(1):119–127, 2015.

[3] Clinical Practice Research Datalink (CPRD): UK data driving real-world evidence. `https://www.cprd.com`. Accessed: 2019-03-21.

[4] Derk C F Klatte, Alessandro Gasparini, Hong Xu, Pietro de Deco, Marco Trevisan, Anna L V Johansson, Björn Wettermark, Johan Ärnlöv, Cynthia J Janmaat, Bengt Lindholm, Friedo W Dekker, Josef Coresh, Morgan E Grams, and Juan J Carrero. Association between proton pump inhibitor use and risk of progression of chronic kidney disease. *Gastroenterology*, 153(3):702–710, 2017.

[5] Sara L Thomas, Caroline Minassian, Vijeya Ganesan, Sinéad M Langan, and Liam Smeeth. Chickenpox and risk of stroke: a self-controlled case series analysis. *Clinical Infectious Diseases*, 58(1):61–68, 2014.

[6] How EHRs facilicate clinical research. `http://www.appliedclinicaltrialsonline.com/how-ehrs-facilitate-clinical-research`. Accessed: 2019-03-21.

[7] eMERGE Network. `https://emerge.mc.vanderbilt.edu`. Accessed: 2019-03-21.

[8] Eric I Benchimol, Liam Smeeth, Astrid Guttmann, Katie Harron, David Moher, Irene Petersen, Henrik T Sørensen, Erik von Elm, Sinéad M Langan, and RECORD Working Committee. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. *PLOS Medicine*, 12(10):1–22, 2015.

[9] Riordan Riordan, Chrysanthi Papoutsi, Julie E Reed, Cicely Marston, Derek Bell, and Azeem Majeed. Patient and public attitudes towards informed consent models and levels of awareness of electronic health records in the UK. *International Journal of Medical Informatics*, 84(4):237–247, 2015.

[10] Spiros C Denaxas and Katherine I Morley. Big biomedical data and cardiovascular disease research: opportunities and challenges. *European Heart Journal - Quality of Care and Clinical Outcomes*, 1(1):9–16, 2015.

[11] Joan A Casey, Brian S Schwartz, Walter F Stewart, and Nancy E Adler. Using electronic health records for population health research: A review of methods and applications. *Annual Review of Public Health*, 37(1):61–81, 2016.

[12] Martin R Cowie, Juuso I Blomster, Lesley H Curtis, Sylvie Duclaux, Ian Ford, Fleur Fritz, Samantha Goldman, Salim Janmohamed, Jörg Kreuzer, Mark Leenay, Alexander Michel, Seleen Ong, Jill P Pell, Mary Ross Southworth, Wendy Gattis Stough, Martin Thoenes, Faiez Zannad, and Andrew Zalewski. Electronic health records to facilitate clinical research. *Clinical Research in Cardiology*, 106(1):1–9, 2017.

[13] Wen-Wai Yim, Amanda J Wheeler, Catherine Curtin, Todd H Wagner, and Tina Hernandez-Boussard. Secondary use of electronic medical records for clinical research: challenges and opportunities. *Convergent Science Physical Oncology*, 4(1):014001, 2018.

[14] David G Kleinbaum and Mitchel Klein. *Survival analysis: A self-learning text*. Statistics for Biology and Health. Springer-Verlag New York, 3rd edition, 2012.

[15] Yin Bun Cheung, Fei Gao, and Kei Siong Khoo. Age at diagnosis and the choice of survival analysis methods in cancer epidemiology. *Journal of Clinical Epidemiology*, 56(1):38–43, 2003.

[16] Anne C M Thiébaut and Jacques Bénichou. Choice of time-scale in Cox's model analysis of epidemiologic cohort data: a simulation study. *Statistics in Medicine*, 23(24):3803–3820, 2004.

[17] David Collett. *Modelling survival data in medical research.* Chapman & Hall / CRC Texts in Statistical Science. Chapman & Hall / CRC, 3rd edition, 2014.

[18] EL Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

[19] Major Greenwood. The natural duration of cancer. *Reports on Public Health and Medical Subjects*, 33:1–26, 1926.

[20] Nathan Mantel. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, 50(3):163–170, 1966.

[21] David R Cox and David Oakes. *Analysis of survival data.* Chapman & Hall / CRC Monographs on Statistics and Applied Probability. Chapman & Hall / CRC, 1984.

[22] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

[23] John D Kalbfleisch and Ross L Prentice. Marginal likelihoods based on cox's regression and life model. *Biometrika*, 60(2):267–278, 1973.

[24] Patrick Royston and Mahesh K B Parmar. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine*, 21(15):2175–2197, 2002.

[25] Sylvain Durrleman and Richard Simon. Flexible regression models with cubic splines. *Statistics in Medicine*, 8(5):551–561.

[26] Hirotogu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[27] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[28] Mark J Rutherford, Michael J Crowther, and Paul C Lambert. The use of restricted cubic splines to approximate complex hazard functions in the analysis of time-to-event data: a simulation study. *Journal of Statistical Computation and Simulation*, 85(4):777–793, 2015.

[29] Elisavet Syriopoulou, Sarwar I Mozumder, Mark J Rutherford, and Paul C Lambert. Robustness of individual and marginal model-based estimates: a sensitivity analysis of flexible parametric models. *Cancer Epidemiology*, 58:17–24, 2019.

[30] Patrick Royston and Paul C Lambert. *Flexible parametric survival analysis using Stata: beyond the Cox model.* Stata Press, 2011.

[31] Peter J Diggle, Patrick Heagerty, Kung-Yee Liang, and Scott L Zeger. *Analysis of longitudinal data.* Oxford Statistical Science Series. Oxford University Press, 2$^{\text{nd}}$ edition, 2013.

[32] Paul S Albert. Longitudinal data analysis (repeated measures) in clinical trials. *Statistics in Medicine*, 18(13):1707–1732, 1999.

[33] Joseph J Locascio and Alireza Atri. An overview of longitudinal data analysis methods for neurological research. *Dementia and Geriatric Cognitive Disorders Extra*, 1(1):330–357, 2011.

[34] Kung-Yee Liang and Scott L Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

[35] Scott L Zeger and Kung-Yee Liang. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42(1):121–130, 1986.

[36] Robert W M Wedderburn. Quasi-likelihood functions, generalized linear models, and the Gauss—Newton method. *Biometrika*, 61(3):439–447, 1974.

[37] Wei Pan. Akaike's information criterion in generalized estimating equations. *Biometrics*, 57(1):120–125, 2001.

[38] María Carmen Pardo and Rosa Alonso. Working correlation structure selection in GEE analysis. *Statistical Papers*, pages 1–21, 2017.

[39] David A Harville. Maximum likelihood approaches to variance component

estimation and to related problems. *Journal of the American Statistical Association*, 72(358):320–338, 1977.

[40] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39(1):1–38, 1977.

[41] Iain L MacDonald. Numerical maximisation of likelihood: a neglected alternative to EM? *International Statistical Review*, 82(2):296–308, 2014.

[42] Tanner Sorensen, Sven Hohenstein, and Shravan Vasishth. Bayesian linear mixed models using Stan: A tutorial for psychologists, linguists, and cognitive scientists. *The Quantitative Methods for Psychology*, 12(3):175–200, 2016.

[43] Geert Molenberghs and Geert Verbeke. *Models for Discrete Longitudinal Data*. Springer-Verlag, 2005.

[44] Joseph C Gardiner, Zhehui Luo, and Lee Anne Roman. Fixed effects, random effects and GEE: what are the differences? *Statistics in Medicine*, 28(2):221–239, 2009.

[45] Alan E Hubbard, Jennifer Ahern, Nancy L Fleischer, Mark Van der Laan, Sheri A Lippman, Nicholas Jewell, Tim Bruckner, and William A Satariano. To GEE or not to GEE: comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. *Epidemiology*, 21(4):467–474, 2010.

[46] Kidney Disease: Improving Global Outcomes (KDIGO) CKD Work Group. KDIGO 2012 clinical practice guideline for the evaluation and management of chronic kidney disease. *Kidney International, Suppl.*, 3(1):1–150, 2013.

[47] Josef Coresh. Update on the burden of CKD. *Journal of the American Society of Nephrology*, 28(4):1020–1022, 2017.

[48] Rupert W Major, Celia Brown, David Shepherd, Stephen Rogers, Warren Pickering, Graham L Warwick, Shaun Barber, Nuzhat B Ashra, Tom Morris, and Nigel J Brunskill. The primary-secondary care partnership to improve outcomes in

chronic kidney disease (PSP-CKD) study: A cluster randomized trial in primary care. *Journal of the American Society of Nephrology*, 30(7):1261–1270, 2019.

[49] IMproving Patient care and Awareness of Kidney disease progression Together. `http://www.impakt.org.uk`. Accessed: 2019-03-01.

[50] Andrew S Levey, Juan P Bosch, Julia Breyer Lewis, Tom Greene, Nancy Rogers, David Roth, and The Modification of Diet in Renal Disease Study Group. A more accurate method to estimate glomerular filtration rate from serum creatinine: A new prediction equation. *Annals of Internal Medicine*, 130(6):461–470, 1999.

[51] Michael Schemper and Terry L Smith. A note on quantifying follow-up in studies of failure time. *Contemporary Clinical Trials*, 17(4):343–346, 1996.

[52] Derek C Angus, Walter T Linde-Zwirble, Jeffrey Lidicker, Gilles Clermont, Joseph Carcillo, and Michael R Pinsky. Epidemiology of severe sepsis in the united states: Analysis of incidence, outcome, and associated costs of care. *Critical Care Medicine*, 29(7):1303–1310, 2001.

[53] Djillali Annane, Philippe Aegerter, Marie Claude Jars-Guincestre, and Bertrand Guidet. Current epidemiology of septic shock. *American Journal of Respiratory and Critical Care Medicine*, 168(2):165–172, 2003.

[54] A randomised controlled trial of VAsopressin versus norepinephrine in Septic Shock. `http://www.isrctn.com/ISRCTN94845869`. Accessed: 2018-10-26.

[55] James A Russell, Keith R Walley, Joel Singer, Anthony C Gordon, Paul C Hébert, D James Cooper, Cheryl L Holmes, Sangeeta Mehta, John T Granton, Michelle M Storms, Deborah J Cook, Jeffrey J Presneill, and Dieter Ayers. Vasopressin versus norepinephrine infusion in patients with septic shock. *New England Journal of Medicine*, 358(9):877–887, 2008.

[56] JL Vincent, Rui Moreno, Jukka Takala, Sheila Willatts, Arnaldo De Mendonça, Hajo Bruining, CK Reinhart, Peter M Suter, and LG Thijs. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. *Intensive Care Medicine*, 22(7):707–710, 1996.

[57] Harm-Jan de Grooth, Irma L Geenen, Armand R Girbes, Jean-Louis Vincent,

Jean-Jacques Parienti, and Heleen M Oudemans-van Straaten. SOFA and mortality endpoints in randomized controlled trials: a systematic review and meta-regression analysis. *Critical Care*, 21(1):21–38, 2017.

[58] Tim P Morris, Ian White, and Michael J Crowther. Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, pages 1–29, 2019.

[59] Scopus: The largest database of peer-reviewed literature. `https://www.scopus.com`. Accessed: 2019-03-27.

[60] Elizabeth Koehler, Elizabeth Brown, and Sebastien J P A Haneuse. On the assessment of Monte Carlo error in simulation-based statistical analyses. *The American Statistician*, 63(2):155–162, 2009.

[61] Brennan C Kahan. Bias in randomised factorial trials. *Statistics in Medicine*, 32(26):4540–4549, 2013.

[62] Harlan Campbell and Charmaine B Dean. The consequences of proportional hazards based model selection. *Statistics in Medicine*, 33(6):1042–1056, 2014.

[63] Ralf Bender, Thomas Augustin, and Maria Blettner. Generating survival times to simulate cox proportional hazards models. *Statistics in Medicine*, 24(11):1713–1723, 2005.

[64] Michael J Crowther and Paul C Lambert. Simulating biologically plausible complex survival data. *Statistics in Medicine*, 32(23):4118–4134, 2013.

[65] Geoff J McLachlan and DC McGiffin. On the role of finite mixture models in survival analysis. *Statistical Methods in Medical Research*, 3(3):211–226, 1994.

[66] Paul C Lambert, Paul W Dickman, Claire L Weston, and John R Thompson. Estimating the cure fraction in population-based cancer studies by using finite mixture models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(1):35–55, 2010.

[67] Richard P Brent. *Algorithms for minimization without derivatives.* Prentice-Hall, 1973.

[68] Michael J Crowther and Paul C Lambert. Simulating complex survival data. *Stata Journal*, 12(4):674–687, 2012.

[69] Sam Brilleman. *simsurv: Simulate Survival Data*, 2019. R package version 0.2.3; Accessed: 2019-05-01.

[70] Ian R White. simsum: Analyses of simulation studies including monte carlo error. *The Stata Journal*, 10(3):369–385, 2010.

[71] Hadley Wickham. Tidy data. *Journal of Statistical Software*, 59(10), 2014.

[72] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York, 2016.

[73] Gerta Rücker and Guido Schwarzer. Presenting simulation results in a nested loop plot. *BMC Medical Research Methodology*, 14(1), 2014.

[74] Christine Laine, Steven N Goodman, Michael E Griswold, and Harold C Sox. Reproducible research: Moving toward research the public can really trust. *Annals of Internal Medicine*, 146(6):450–453, 2007.

[75] Kenneth F Schulz, Douglas G Altman, David Moher, and for the CONSORT Group. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *PLOS Medicine*, 7(3):1–7, 2010.

[76] Erik von Elm, Douglas G Altman, Matthias Egger, Stuart J Pocock, Peter Gøtzsch, Jan P Vandenbroucke, and for the STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: guidelines for reporting observational studies. *PLOS Medicine*, 4(10):1–5, 2007.

[77] Roger D Peng. Reproducible research in computational science. 334(6060):1226–1227, 2011.

[78] Davic C Hoaglin and David F Andrews. The reporting of computation-based results in statistics. *The American Statistician*, 29(3):122–126, 1975.

[79] Walter W Hauck and Sharon Anderson. A survey regarding the reporting of simulation studies. *The American Statistician*, 38(3):214–216, 1984.

[80] Ignacio Díaz-Emparanza. Is a small Monte Carlo analysis a good analysis? *Statistical Papers*, 43(4):567–577, 2002.

[81] Andrea Burton, Douglas G Altman, Patrick Royston, and Roger L Holder.

The design of simulation studies in medical statistics. *Statistics in Medicine*, 25(24):4279–4292, 2006.

[82] Mike K Smith and Andrea Marshall. Importance of protocols for simulation studies in clinical drug development. *Statistical Methods in Medical Research*, 20(6):613–622, 2011.

[83] David Spiegelhalter, Mike Pearson, and Ian Short. Visualizing uncertainty about the future. *Science*, 333(6048):1393–1400, 2011.

[84] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D^3: Data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, 2011.

[85] Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie, and Jonathan McPherson. *shiny: Web Application Framework for R*, 2019. R package version 1.3.2; accessed: 2019-05-03.

[86] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019.

[87] Winston Chang and Barbara Borges Ribeiro. *shinydashboard: Create Dashboards with 'Shiny'*, 2018. R package version 0.7.1; accessed: 2019-05-06.

[88] AdminLTE Control Panel Template. `https://adminlte.io/`. Accessed: 2019-05-06.

[89] Bootstrap. `https://getbootstrap.com`. Accessed: 2019-05-06.

[90] Ian R White, Patrick Royston, and Angela M Wood. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine*, 30(4):377–399, 2011.

[91] Nicholas J Tierney and Dianne H Cook. Expanding tidy data principles to facilitate missing data exploration, visualization and assessment of imputations. *arXiv e-prints*, page arXiv:1809.02264, 2018.

[92] Simon Garnier. *viridis: Default Color Maps from 'matplotlib'*, 2018. R package version 0.5.1; Accessed: 2019-07-27.

[93] Alessandro Gasparini, Mark S Clements, Keith R Abrams, and Michael J Crowther.

Impact of model misspecification in shared frailty survival models. *Statistics in Medicine*, 38(23):4474–4502, 2019.

[94] Per K Andersen and Richard D Gill. Cox's regression model for counting processes: A large sample study. *The Annals of Statistics*, 10(4):1100–1120, 1982.

[95] Ross L Prentice, BJ Williams, and AV Peterson. On the regression analysis of multivariate failure time data. *Biometrika*, 68(2):373–379, 1981.

[96] Terry M Therneau and Patricia M Grambsch. *Modeling survival data: extending the Cox model.* Springer-Verlag New York, 2000.

[97] James W Vaupel, Kenneth G Manton, and Eric Stallard. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16(3):439–454, 1979.

[98] Tony Lancaster. Econometric methods for the duration of unemployment. *Econometrica*, 47(4):939–956, 1979.

[99] Philip Hougaard. A class of multivanate failure time distributions. *Biometrika*, 73(3):671–678, 1986.

[100] Roberto G Gutierrez. Parametric frailty and shared frailty survival models. *The Stata Journal*, 2(1):22 – 44, 2002.

[101] Philip Hougaard. Frailty models for survival data. *Lifetime Data Analysis*, 1(3):255–273, 1995.

[102] Virginie Rondeau, Laurent Filleul, and Pierre Joly. Nested frailty models using maximum penalized likelihood estimation. *Statistics in Medicine*, 25(23):4036–4052, 2006.

[103] Virginie Rondeau, Stefan Michiels, Benoit Liquet, and Jean-Pierre Pignon. Investigating trial and treatment heterogeneity in an individual patient data meta-analysis of survival data by means of the penalized maximum likelihood approach. *Statistics in Medicine*, 27(11):1894–1910, 2008.

[104] Virginie Rondeau, Simone Mathoulin-Pelissier, Héléne Jacqmin-Gadda, Véronique Brouste, and Pierre Soubeyran. Joint frailty models for recurring events and death

using maximum penalized likelihood estimation: application on cancer events. *Biostatistics*, 8(4):708–721, 2006.

[105] Virginie Rondeau, Jean-Pierre Pignon, and Stefan Michiels. A joint model for the dependence between clustered times to tumour progression and deaths: A meta-analysis of chemotherapy in head and neck cancer. *Statistical Methods in Medical Research*, 24(6):711–729, 2015.

[106] Yassin Mazroui, Simone Mathoulin-Pelissier, Pierre Soubeyran, and Virginie Rondeau. General joint frailty model for recurrent event data with a dependent terminal event: Application to follicular lymphoma data. *Statistics in Medicine*, 31(11-12):1162–1176, 2012.

[107] Il Do Ha, Richard Sylvester, Catherine Legrand, and Gilbert MacKenzie. Frailty modelling for survival data from multi-centre clinical trials. *Statistics in Medicine*, 30(17):2144–2159, 2011.

[108] Michael J Crowther, Maxime P Look, and Richard D Riley. Multilevel mixed effects parametric survival models using adaptive Gauss-Hermite quadrature with application to recurrent events and individual participant data meta-analysis. *Statistics in Medicine*, 33(22):3844–3858, 2014.

[109] Samuli Ripatti and Juni Palmgren. Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics*, 56(4):1016–1022, 2000.

[110] Terry M Therneau, Patricia M Grambsch, and V Shane Pankratz. Penalized survival models and frailty. *Journal of Computational and Graphical Statistics*, 12(1):156–175, 2003.

[111] Philip Hougaard. Life table methods for heterogeneous populations: distributions describing the heterogeneity. *Biometrika*, 71(1):75–83, 1984.

[112] Philip Hougaard. Survival models for heterogeneous populations derived from stable distributions. *Biometrika*, 73(2):387–396, 1986.

[113] John C Naylor and Adrian F M Smith. Applications of a method for the efficient computation of posterior distributions. *Applied Statistics*, 31(3):214–225, 1982.

[114] Qing Liu and Donald A Pierce. A note on Gauss-Hermite quadrature. *Biometrika*, 81(3):624–629, 1994.

[115] Francis Tuerlinckx, Frank Rijmen, Geert Verbeke, and Paul Boeck. Statistical inference in generalized linear mixed models: A review. *British Journal of Mathematical and Statistical Psychology*, 59(2):225–255, 2006.

[116] Peter Jäckel. A note on multivariate Gauss-Hermite quadrature. 2005.

[117] José C Pinheiro and Douglas M Bates. Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, 4(1):12–35, 1995.

[118] Andrew Pickles and Robert Crouchley. A comparison of frailty models for multivariate survival data. *Statistics in Medicine*, 14(13):1447–1461, 1995.

[119] David V Glidden and Eric Vittinghoff. Modelling clustered survival data from multicentre clinical trials. *Statistics in Medicine*, 23(3):369–388, 2004.

[120] Katherine J Lee and Simon G Thompson. Flexible parametric models for random-effects distributions. *Statistics in Medicine*, 27(3):418–434, 2008.

[121] Xing-Rong Liu, Yudi Pawitan, and Mark S Clements. Generalized survival models for correlated time-to-event data. *Statistics in Medicine*, 36(29):4743–4762, 2017.

[122] Luc Duchateau, Paul Janssen, Patrick Lindsey, Catherine Legrand, Rosemary Nguti, and Richard Sylvester. The shared frailty model and the power for heterogeneity tests in multicenter trials. *Computational Statistics & Data Analysis*, 40(3):603–620, 2002.

[123] Il Do Ha and Youngjo Lee. Estimating frailty models via Poisson hierarchical generalized linear models. *Journal of Computational and Graphical Statistics*, 12(3):663–681, 2003.

[124] Chris Elbers and Geert Ridder. True and spurious duration dependence: The identifiability of the proportional hazard model. *The Review of Economic Studies*, 49(3):403–409, 1982.

[125] Theodor Adrian Bălan. *Advances in frailty models*. PhD thesis, Universiteit Leiden, 09 2018.

[126] Therese ML Andersson, Paul W Dickman, Sandra Eloranta, Mats Lambe, and Paul C Lambert. Estimating the loss in expectation of life due to cancer using flexible parametric survival models. *Statistics in medicine*, 32(30):5286–5300, 2013.

[127] Marlies Noordzij, Merel van Diepen, Fergus C Caskey, and Kitty J Jager. Relative risk versus absolute risk: one cannot be interpreted without the other. *Nephrology Dialysis Transplantation*, 32(suppl_2):ii13–ii18, 2017.

[128] Patrick Royston and Mahesh K B Parmar. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Medical Research Methodology*, 13(1), 2013.

[129] Hidetosi Takahasi and Masatake Mori. Double exponential formulas for numerical integration. *Publications of the Research Institute for Mathematical Sciences*, 9(3):721–741, 1973.

[130] Hans W Borchers. *pracma: Practical Numerical Math Functions*, 2019. R package version 2.2.5; Accessed: 2019-05-22.

[131] David H Bailey. Tanh-sinh high-precision quadrature. Technical report, 2006.

[132] Mark Clements and Xing-Rong Liu. *rstpm2: Generalized Survival Models*, 2019. R package version 1.4.5; Accessed: 2019-03-28.

[133] Katharina Hirsch and Andreas Wienke. Software for semiparametric shared gamma and log-normal frailty models: an overview. *Computer Methods and Programs in Biomedicine*, 107(3):582–597, 2012.

[134] John Monaco, Malka Gorfine, and Li Hsu. General semiparametric shared frailty model: Estimation and simulation with `frailtysurv`. *Journal of Statistical Software*, 86(4):1–42, 2018.

[135] Il Do Ha, Maengseok Noh, and Youngjo Lee. `frailtyHL`: A package for fitting frailty models with H-likelihood. *The R Journal*, 4(2):28–37, 2012.

[136] Marco Munda, Federico Rotolo, and Catherine Legrand. parfm: Parametric frailty models in R. *Journal of Statistical Software*, 51(11), 2012.

[137] Xing-Rong Liu, Yudi Pawitan, and Mark Clements. Parametric and

penalized generalized survival models. *Statistical Methods in Medical Research*, 27(5):1531–1546, 2018.

[138] Virginie Rondeau, Yassin Mazroui, and Juan R Gonzalez. frailtypack: An R package for the analysis of correlated survival data with frailty models using penalized likelihood estimation or parametrical estimation. *Journal of Statistical Software*, 47, 2012.

[139] Pierre Joly, Daniel Commenges, and Luc Letenneur. A penalized likelihood approach for arbitrarily censored and truncated data: Application to age-specific incidence of dementia. *Biometrics*, 54(1):185–194, 1998.

[140] Bradley Efron. Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika*, 68(3):589, 1981.

[141] Trevor J Hastie and Daryl Pregibon. Generalized linear models. In *Statistical Models in S*, pages 195–247. Routledge, 1992.

[142] Florin Vaida and Suzette Blanchard. Conditional Akaike information for mixed-effects models. *Biometrika*, 92(2):351–370, 2005.

[143] Hua Liang, Hulin Wu, and Guohua Zou. A note on conditional AIC for linear mixed-effects models. *Biometrika*, 95(3):773–778, 2008.

[144] Paola Rebora, Agus Salim, and Marie Reilly. bshazard: A flexible tool for nonparametric smoothing of the hazard function. *The R Journal*, 6(2):114–122, 2014.

[145] Hannah Bower, Michael J Crowther, Mark J Rutherford, Therese M L Andersson, Mark S Clements, Xing-Rong Liu, Paul W Dickman, and Paul C Lambert. Capturing simple and complex time-dependent effects using flexible parametric survival models: A simulation study. *Communications in Statistics - Simulation and Computation*, pages 1–17, 2019.

[146] Philip Hougaard. *Analysis of multivariate survival data.* Springer New York, 2000.

[147] Leila DAF Amorim and Jianwen Cai. Modelling recurrent events: a tutorial for analysis in epidemiology. *International Journal of Epidemiology*, 44(1):324–333, 2015.

[148] Willem Kuyken, Fiona C Warren, Rod S Taylor, Ben Whalley, Catherine Crane, Guido Boldolfi, Rachel Hayes, Marloes Huijbers, Helen Ma, Susanne Schweizer, Zindel Segal, Anne Speckens, John D Teasdale, Kees Van Heeringen, Mark Williams, Sarah Byford, Richard Byng, Tim Dalgleish, et al. Efficacy of mindfulness-based cognitive therapy in prevention of depressive relapse: An individual patient data meta-analysis from randomized trials. *JAMA Psychiatry*, 73(6):565–574, 2016.

[149] Bruno R Chrcanovic, Jenö Kisch, Tomas Albrektsson, and Ann Wennerberg. Bruxism and dental implant failures: a multilevel mixed effects parametric survival analysis approach. *Journal of Oral Rehabilitation*, 43(11):813–823, 2016.

[150] Ulrik Pedersen-Bjergaard, Stig Pramming, Simon R Heller, Tara M Wallace, Åse K Rasmussen, Hanne V Jørgensen, David R Matthews, Philip Hougaard, and Birger Thorsteinsson. Severe hypoglycaemia in 1076 adult patients with type 1 diabetes: influence of risk markers and selection. *Diabetes/Metabolism Research and Reviews*, 20(6):479–486, 2004.

[151] Giuseppe Gargiulo, Stephan Windecker, Bruno R da Costa, Fausto Feres, Myeong-Ki Hong, Martine Gilard, Hyo-Soo Kim, Antonio Colombo, Deepak L Bhatt, Byeong-Keuk Kim, Marie-Claude Morice, Kyung Woo Park, Alaide Chieffo, Tullio Palmerini, Gregg W Stone, and Marco Valgimigli. Short term versus long term dual antiplatelet therapy after implantation of drug eluting stent in patients with or without diabetes: systematic review and meta-analysis of individual participant data from randomised trials. *BMJ*, 355, 2016.

[152] Srikant Devaraj and Pankaj C Patel. Attributing responsibility: hospitals account for 20% of variance in acute myocardial infarction patient mortality. *Journal for Healthcare Quality*, 38(1):52–61, 2016.

[153] Nicola Saino, Maria Romano, Roberto Ambrosini, Diego Rubolini, Giuseppe Boncoraglio, Manuela Caprioli, and Andrea Romano. Longevity and lifetime reproductive success of barn swallow offspring are predicted by their hatching date and phenotypic quality. *Journal of Animal Ecology*, 81(5):1004–1012, 2012.

[154] Nancy Reid. A conversation with Sir David Cox. *Statistical Science*, 9(3):439–455,

1994.

[155] Christopher P Nelson, Paul C Lambert, Iain B Squire, and David R Jones. Flexible parametric models for relative survival, with application in coronary heart disease. *Statistics in Medicine*, 26(30):5486–5498, 2007.

[156] Shayle R Searle, George Casella, and Charles E McCulloch. *Variance Components*. John Wiley & Sons, Inc., 1992.

[157] Anastasios A Tsiatis, Victor DeGruttola, and Michael S Wulfsohn. Modeling the relationship of survival to longitudinal data measured with error. applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association*, 90(429):27–37, 1995.

[158] Joseph W Hogan and Nan M Laird. Increasing efficiency from censored survival data by using random effects to model longitudinal covariates. *Statistical Methods in Medical Research*, 7(1):28–48, 1998.

[159] Dimitris Rizopoulos, Laura A Hatfield, Bradley P Carlin, and Johanna J M Takkenberg. Combining dynamic predictions from joint models for longitudinal and time-to-event data using bayesian model averaging. *Journal of the American Statistical Association*, 109(508):1385–1397, 2014.

[160] Peter J Diggle and Michael G Kenward. Informative drop-out in longitudinal data analysis. *Applied Statistics*, 43(1):49, 1994.

[161] John J McArdle, Brent J Small, Lars Bäckman, and Laura Fratiglioni. Longitudinal models of growth and survival applied to the early detection of Alzheimer's disease. *Journal of Geriatric Psychiatry and Neurology*, 18(4):234–241, 2005.

[162] Yudi Pawitan and Steve Self. Modeling disease marker processes in AIDS. *Journal of the American Statistical Association*, 88(423):719–726, 1993.

[163] Michael S Wulfsohn and Anastasios A Tsiatis. A joint model for survival and longitudinal data measured with error. *Biometrics*, 53(1):330–339, 1997.

[164] Anastasios A Tsiatis and Marie Davidian. Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, 14:809–834, 2004.

[165] Robin Henderson, Peter J Diggle, and Angela Dobson. Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1(4):465–480, 2000.

[166] Peter Diggle. Dealing with missing values in longitudinal studies. In *Recent Advances in the Statistical Analysis of Medical Data*, pages 203–228. Wiley-BlackWell, 1998.

[167] Joseph G Ibrahim, Haitao Chu, and Liddy M Chen. Basic concepts and methods for joint models of longitudinal and survival data. *Journal of Clinical Oncology*, 28(16):2796–2801, 2010.

[168] Dimitris Rizopoulos. *Joint models for longitudinal and time-to-event data: with applications in R*. Biostatistics. Chapman & Hall / CRC, 2012.

[169] A Lawrence Gould, Mark Ernest Boye, Michael J Crowther, Joseph G Ibrahim, George Quartey, Sandrine Micallef, and Frederic Y Bois. Joint modeling of survival and longitudinal non-survival data: current methods and issues. report of the DIA bayesian joint modeling working group. *Statistics in Medicine*, 34(14):2181–2195, 2015.

[170] Michael J Sweeting and Simon G Thompson. Joint modelling of longitudinal and time-to-event data with application to predicting abdominal aortic aneurysm growth and rupture. *Biometrical Journal*, 53(5):750–763, 2011.

[171] Özgür Asar, James Ritchies, Philip A Kalra, and Peter J Diggle. Joint modelling of repeated measurement and time-to-event data: an introductory tutorial. *International Journal of Epidemiology*, 44(1):334–344, 2015.

[172] Montserrat Rué, Eleni-Rosalina Andrinopoulou, Danilo Alvares, Carmen Armero, Anabel Forte, and Lluis Blanch. Bayesian joint modeling of bivariate longitudinal and competing risks data: An application to study patient-ventilator asynchronies in critical care patients. *Biometrical Journal*, 59(6):1184–1203, 2017.

[173] Yi-Kuan Tseng, Fushing Hsieh, and Jane-Ling Wang. Joint modelling of accelerated failure time and longitudinal data. *Biometrika*, 92(3):587–603, 2005.

[174] Fushing Hsieh, Yi-Kuan Tseng, and Jane-Ling Wang. Joint modeling of survival

and longitudinal data: likelihood approach revisited. *Biometrics*, 62(4):1037–1043, 2006.

[175] Michael J Crowther, Keith R Abrams, and Paul C Lambert. Flexible parametric joint modelling of longitudinal and survival data. *Statistics in Medicine*, 31(30):4456–4471, 2012.

[176] Cécile Proust-Lima, Mbéry Séne, Jeremy MG Taylor, and Hélène Jacqmin-Gadda. Joint latent class models for longitudinal and time-to-event data: A review. *Statistical Methods in Medical Research*, 23(1):74–90, 2014.

[177] Paula R Williamson, Ruwanthi Kolamunnage-Dona, Pete Philipson, and Anthony G Marson. Joint modelling of longitudinal and competing risks data. *Statistics in Medicine*, 27(30):6426–6438, 2008.

[178] Graeme L Hickey, Pete Philipson, Andrea Jorgensen, and Ruwanthi Kolamunnage-Dona. Joint models of longitudinal and time-to-event data with more than one event time outcome: A review. *The International Journal of Biostatistics*, 14(1), 2018.

[179] Graeme L Hickey, Pete Philipson, Andrea Jorgensen, and Ruwanthi Kolamunnage-Dona. Joint modelling of time-to-event and multivariate longitudinal outcomes: recent developments and issues. *BMC Medical Research Methodology*, 16(1):117, 2016.

[180] Patrick Royston and Douglas G Altman. Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling. *Applied Statistics*, 43(3):429–467, 1994.

[181] Elizabeth R Brown, Joseph G Ibrahim, and Victor DeGruttola. A flexible B-spline model for multiple longitudinal biomarkers and survival. *Biometrics*, 61(1):64–73, 2005.

[182] Dimitris Rizopoulos and Pulak Ghosh. A bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Statistics in Medicine*, 30(12):1366–1380, 2011.

[183] Anastasios A Tsiatis and Marie Davidian. A semiparametric estimator for the

proportional hazards model with longitudinal covariates measured with error. *Biometrika*, 88(2):447–458, 2001.

[184] Paul C Abbott. Tricks of the trade: Legendre-Gauss quadrature. *Mathematica Journal*, 9(4):689–691, 2005.

[185] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[186] Daniel F Heitjian and Donald B Rubin. Ignorability and coarse data. *The Annals of Statistics*, 19(4):2244–2253, 1991.

[187] Matthew Powney, Paula R Williamson, Jamie Kirkham, and Ruwanthi Kolamunnage-Dona. A review of the handling of missing longitudinal outcome data in clinical trials. *Trials*, 15(1), 2014.

[188] Ruwanthi Kolamunnage-Dona, Colin Powell, and Paula R Williamson. Modelling variable dropout in randomised controlled trials with longitudinal outcomes: application to the MAGNETIC study. *Trials*, 17(1), 2016.

[189] Qiuju Li and Li Su. Accommodating informative dropout and death: a joint modelling approach for longitudinal and semicompeting risks data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(1):145–163, 2018.

[190] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121, 1995.

[191] Roderick J A Little. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88(421):125–134, 1993.

[192] Geert Molenberghs, Bart Michiels, Michael G Kenward, and Peter J Diggle. Monotone missing data and pattern-mixture models. *Statistica Neerlandica*, 52(2):153–161, 1998.

[193] Margaret C Wu and Raymond J Carroll. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, 44(1):175, 1988.

[194] Graeme L Hickey, Pete Philipson, Andrea Jorgensen, and Ruwanthi Kolamunnage-Dona. joineRML: a joint model and software package for

time-to-event and multivariate longitudinal outcomes. *BMC Medical Research Methodology*, 18(1):50, 2018.

[195] Graeme L Hickey, Pete Philipson, Andrea Jorgensen, and Ruwanthi Kolamunnage-Dona. *joineRML: Joint Modelling of Multivariate Longitudinal Data and Time-to-Event Outcomes*, 2018. R package version 0.4.2; Accessed: 2019-02-28.

[196] Haiqun Lin, Charles E McCulloch, and Susan T Mayne. Maximum likelihood estimation in the joint analysis of time-to-event and multiple longitudinal variables. *Statistics in Medicine*, 21(16):2369–2382, 2002.

[197] Danjie Zhang, Ming-Hui Chen, Joseph G Ibrahim, Mark E Boye, Ping Wang, and Wei Shen. Assessing model fit in joint models of longitudinal and survival data with applications to cancer clinical trials. *Statistics in Medicine*, 33(27):4715–4733, 2014.

[198] José Pinheiro, Douglas Bates, Saikat DebRoy, Deepayan Sarkar, and R Core Team. *nlme: Linear and Nonlinear Mixed Effects Models*, 2019. R package version 3.1-140; Accessed: 2019-06-03.

[199] Eleanor M Pullenayegum and Lily SH Lim. Longitudinal data subject to irregular observation: A review of methods with a focus on visit processes, assumptions, and study design. *Statistical Methods in Medical Research*, 25(6):2992–3014, 2016.

[200] Delaram Farzanfar, Asmaa Abumuamar, Jayoon Kim, Emily Sirotich, Yue Wang, and Eleanor M Pullenayegum. Longitudinal studies that use data collected as part of usual care risk reporting biased results: a systematic review. *BMC Medical Research Methodology*, 17(1):133, 2017.

[201] Alessandro Gasparini, Keith R Abrams, Jessica K Barrett, Rupert W Major, Michael J Sweeting, Nigel J Brunskill, and Michael J Crowther. Mixed-effects models for health care longitudinal data with an informative visiting process: A monte carlo simulation study. *Statistica Neerlandica*, pages 1–19, 2019.

[202] Jens Gruger, Richard Kay, and Martin Schumacher. The validity of inferences based on incomplete observations in disease state models. *Biometrics*, 47(2):595–605, 1991.

[203] Miguel A Hernán, Mara McAdams, Nuala McGrath, Emilie Lanoy, and Dominique Costagliola. Observation plans in longitudinal studies with time-varying treatments. *Statistical Methods in Medical Research*, 18(1):27–52, 2009.

[204] Miguel A Hernán, Sonia Hernández-Díaz, and James M Robins. A structural approach to selection bias. *Epidemiology*, 15(5):615–625, 2004.

[205] Haiqun Lin, Daniel O Scharfstein, and Robert A Rosenheck. Analysis of longitudinal data with irregular, outcome-dependent follow-up. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3):791–813, 2004.

[206] Petra Buzkova and Thomas Lumley. Longitudinal data analysis for generalized linear models with follow-up dependent on outcome-related variables. *Canadian Journal of Statistics*, 35(4):485–500, 2007.

[207] Stephen R Cole and Miguel A Hernán. Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, 168(6):656–664, 2008.

[208] Danyu Lin and Zhiliang Ying. Semiparametric and nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association*, 96(453):103–126, 2001.

[209] Danyu Lin and Zhiliang Ying. Semiparametric regression analysis of longitudinal data with informative drop-outs. *Biostatistics*, 4(3):385–398, 2003.

[210] Eleanor M Pullenayegum and Brian M Feldman. Doubly robust estimation, optimally truncated inverse-intensity weighting and increment-based methods for the analysis of irregularly observed longitudinal data. *Statistics in Medicine*, 32(6):1054–1072, 2012.

[211] Yu Liang, Wenbin Lu, and Zhiliang Ying. Joint modeling and analysis of longitudinal data with informative observation times. *Biometrics*, 65(2):377–384, 2008.

[212] Liuquan Sun, Xiaoyun Mu, Zhihua Sun, and Xingwei Tong. Semiparametric analysis of longitudinal data with informative observation times. *Acta Mathematicae Applicatae Sinica, English Series*, 27(1):29–42, 2011.

[213] Xinyuan Song, Xiaoyun Mu, and Liuquan Sun. Regression analysis of longitudinal data with time-dependent covariates and informative observation times. *Scandinavian Journal of Statistics*, 39(2):248–258, 2012.

[214] Liuquan Sun, Xinyuan Song, Jie Zhou, and Lei Liu. Joint analysis of longitudinal data with informative observation times and a dependent terminal event. *Journal of the American Statistical Association*, 107(498):688–700, 2012.

[215] Lei Liu, Xuelin Huang, and John O'Quigley. Analysis of longitudinal data in presence of informative observational times and a dependent terminal event, with application to medical cost data. *Biometrics*, 64:950–958, 2008.

[216] Michael J Crowther. Extended multivariate generalised linear and non-linear mixed effects models. *arXiv preprint arXiv:1710.02223*, 2017.

[217] Michael J Crowther. merlin - a unified modelling framework for data analysis and methods development in stata. *arXiv preprint arXiv:1806.01615*, 2018.

[218] Emma C Martin, Alessandro Gasparini, and Michael J Crowther. merlin: an R package for mixed effects regression for linear, nonlinear and user-defined models. 2019 - In preparation.

[219] John M Neuhaus, Charles E McCulloch, and Ross D Boylan. Analysis of longitudinal data from outcome-dependent visit processes: Failure of proposed methods in realistic settings and potential improvements. *Statistics in Medicine*, 37(29):4457–4471, 2018.

[220] Benjamin A Goldstein, Nrupen A Bhavsar, Matthew Phelan, and Michael J Pencina. Controlling for informed presence bias due to the number of health encounters in an electronic health record. *American Journal of Epidemiology*, 184(11):847–855, 2016.

[221] Jianguo Sun, Do-Hwan Park, Liuquan Sun, and Xingqiu Zhao. Semiparametric regression analysis of longitudinal data with informative observation times. *Journal of the American Statistical Association*, 100(471):882–889, 2005.

[222] Peter H Van Ness, Heather G Allore, Terri R Fried, and Haiqun Lin. Inverse intensity weighting in generalized linear models as an option for analyzing

longitudinal data with triggered observations. *American Journal of Epidemiology*, 171(1):105–112, 2009.

[223] Alessandro Gasparini. rsimsum: Summarise results from Monte Carlo simulation studies. *Journal of Open Source Software*, 3(26):739.

[224] Sara J Baart, Eric Boersma, and Dimitris Rizopoulos. Joint models for longitudinal and time-to-event data in a case-cohort design. *Statistics in Medicine*, 38(12):2269–2281, 2019.

[225] Ramnath Vaidyanathan, Yihui Xie, JJ Allaire, Joe Cheng, and Kenton Russell. *htmlwidgets: HTML Widgets for R*, 2018. R package version 1.3; Accessed: 2019-07-15.

[226] Carson Sievert. *plotly for R*, 2018. R package version 4.9.0; Accessed: 2019-07-15.

[227] Javier Luraschi and JJ Allaire. *r2d3: Interface to 'D3' Visualizations*, 2018. R package version 0.2.3; Accessed: 2019-07-15.

[228] Anders Skrondal. Design and analysis of Monte Carlo experiments: attacking the conventional wisdom. *Multivariate Behavioral Research*, 35(2):137–167, 2000.

[229] Matt D Stevenson, John E Brazier, Neill W Calvert, Martin Lloyd-Jones, Jeremy E Oakley, and John A Kanis. Description of an individual patient methodology for calculating the cost-effectiveness of treatments for osteoporosis in women. *Journal of the Operational Research Society*, 56(2):214–221, 2005.

[230] Mark Strong, Jeremy E Oakley, and Alan Brennan. Estimating multiparameter partial expected value of perfect information from a probabilistic sensitivity analysis sample. *Medical Decision Making*, 34(3):311–326, 2013.

[231] Harvey Goldstein, George Leckie, Christopher Charlton, Kate Tilling, and William J Browne. Multilevel growth curve models that incorporate a random coefficient model for the level 1 variance function. *Statistical Methods in Medical Research*, 27(11):3478–3491, 2018.

[232] Jessica K Barrett, Raphael Huille, Richard Parker, Yuichiro Yano, and Michael Griswold. Estimating the association between blood pressure variability and

cardiovascular disease: An application using the ARIC study. *Statistics in Medicine*, 38(10):1855–1868, 2019.

[233] Dimitris Rizopoulos, Jeremy M G Taylor, Joost Van Rosmalen, Ewout W Steyerberg, and Johanna J M Takkenberg. Personalized screening intervals for biomarkers using joint models for longitudinal and survival data. *Biostatistics*, 17(1):149–164, 2015.

[234] Anirudh Tomer, Daan Nieboer, Monique J Roobol, Ewout W Steyerberg, and Dimitris Rizopoulos. Personalized schedules for surveillance of low-risk prostate cancer patients. *Biometrics*, 75(1):153–162, 2018.

[235] Margarita Moreno-Betancur, John B Carlin, Samuel L Brilleman, Stephanie K Tanamas, Anna Peeters, and Rory Wolfe. Survival analysis with time-dependent covariates subject to missing data or measurement error: Multiple imputation for joint modeling (MIJM). *Biostatistics*, 19(4):479–496, 2017.

[236] Microsoft and Steve Weston. *foreach: Provides Foreach Looping Construct for R*, 2017. R package version 1.4.4; Accessed: 2019-03-28.