From Department of Medical Epidemiology and Biostatistics
Karolinska Institutet, Stockholm, Sweden

DEVELOPMENT AND APPLICATION OF COMPETING RISKS AND MULTI-STATE MODELS IN CANCER EPIDEMIOLOGY

Nikolaos Skourlis

Karolinska Institutet

Stockholm 2023

All previously published papers were reproduced with permission from the publisher.
Published by Karolinska Institutet.
Printed by Universitetsservice US-AB, 2023
© Nikolaos Skourlis, 2023
ISBN 978-91-8016-934-9
Cover illustration by Nikolaos Skourlis and the use of the self-developed RShiny application MSMplus described in Study II.

Development and application of competing risks and multi-state models in cancer epidemiology
THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Nikolaos Skourlis

The thesis will be defended in public at Atrium, Karolinska Institutet, Nobels väg 12B, on Friday March 17th, kl 09.00

Principal Supervisor:
Professor Paul C. Lambert
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics
and
University of Leicester, UK
Department of Population Health Sciences

Co-supervisor(s):
Docent Therese M-L. Andersson
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics

Dr. Michael J. Crowther
Red Door Analytics

Opponent:
Dr. Christopher Jackson
University of Cambridge, U.K
Medical Research Council Biostatistics Unit
School of Clinical Medicine

Examination Board:
Docent Orsini Nicola
Karolinska Institutet
Department of Global Public Health

Docent Anna Grimby Ekman
University of Gothenburg
School of Public Health and Community Medicine, Institute of Medicine

Docent Mattias Rantalainen
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics

Dedicated to:

My mother and my brother. Ευχαριστώ πολύ για την αγάπη και την υποστήριξη σας αυτά τα τέσσερα μεταβατικά χρόνια της ζωής μου. Δε ξέρω τι επιφυλάσσει το μέλλον, αλλά ξέρω ότι το ταξίδι θα αξίζει.

My friends from KI. We have created plenty of fun memories together! You helped me power through stress and difficulties and enjoy life in Sweden. A deep thanks to all of you!

POPULAR SCIENCE SUMMARY OF THE THESIS

The use of competing risks and multi-state models in survival analysis settings allows us to study complex disease settings and answer composite research questions, with useful applications in epidemiology. This thesis aims to explore these areas of survival analysis by a) quantifying the influence of the choice of timescale in a competing risks setting, b) exploring different research questions via multi-state models using registry-based repeated prescriptions of antidepressants, discussing the interpretations, traits, limitations of each structure and alternative modelling choices, c) communicating the structures and results of multi-state models via a self-developed, online, interactive web tool and d) evaluating recurrent multi-state modelling approaches in a setting of recurrent events under the presence of a terminal event, when the underlying data mechanism is that of a joint frailty model.

ABSTRACT

Competing risks and multi-state models allow us to study complex disease settings and answer composite research questions and should be used more widely in epidemiology. This thesis aims to explore the competing risks and multi-state models areas using flexible parametric survival models (FPSMs), studying several aspects, such as the choice of timescale, choice of multi-state structure, sharing information across transitions by imposing restrictions in the estimation of the parameters, as well as communicating the results of such models to a wider audience and evaluating the use of recurrent multi-state structures in the area of recurrent events when a terminal event is present.

In competing risks settings, a common timescale is normally used for all competing events. For example, in a setting where death due to colon cancer is the event of interest and death due to other causes serves as a competing event, time since diagnosis is frequently used as the timescale when modelling the hazard rates for both events. However, attained age has been proposed as a more natural timescale when modelling mortality rate that is not associated with the event of interest (colon cancer). In **Study I**, the aim was to assess how the choice of timescale for other cause mortality (time since diagnosis versus attained age) influence the estimated cumulative incidence functions (CIFs) and how several factors contribute to that influence (sample size, non-proportional hazards, shape of baseline other cause mortality rate, variance in age at diagnosis) via a simulation analysis, assuming that the mortality rate is a function of attained age. I found that the bias of the CIF estimates for colon cancer mortality is negligible under all the different approaches and all factor levels. The bias in the CIF estimates for other cause mortality is also low when using time since diagnosis as the timescale for both events, provided that we include age at diagnosis in the models with sufficient flexibility (splines). When a covariate has non-proportional hazards for other cause mortality on the attained age scale, using time since diagnosis as the timescale for other cause mortality may lead to a low but non-negligible bias, no matter how flexibly we model the hazard rate.

The structural complexity of a multi-state structure and the variety of the predicted measures over time for individuals with different covariate patterns may render the communication of the results complicated and difficult. This issues motivated me to develop an interactive web-tool in **Study II** that can be used from researchers to present their multi-state model results to audiences with a variety of interactive graphs that will render the results more communicable and intuitive. The name of the application is MSMplus and it was written using the package RShiny in R. Multi-state model results can easily be wrapped up and uploaded to the application using the `multistate` package in Stata and the `MSMplus` package in R.

When studying a disease process, different research questions may require different multi-state structures in order to be addressed, each structure with different interpretations of the estimated measures, advantages compared to the other structures as well as limitations. There are also a number of modelling choices to consider such as the timescale used for each transition, and sharing information across transitions by imposing specific restrictions in the estimation process. In **Study III**, we explore different research questions via the use of a range of multi-

state models of increasing complexity when dealing with registry-based repeated prescriptions of antidepressants, using the Breast Cancer Data Base Sweden 2.0 research database. I derive probability estimates that address different research questions regarding antidepressant use patterns, beginning with a single-event survival model, moving to a competing risks and a 3-state Illness-Death model, then a 4-state unidirectional and bidirectional model with a post-medication state. Finally, I fit a multi-state structure with recurrent pairs of medication cycles/ discontinuation period states, first with separately estimated transition intensity rates and then allowing sharing of information across transitions by imposing specific restrictions between the baseline transition intensity rates.

When we are interested in studying a recurrent event process in the presence of a terminal event, there is a variety of different frameworks and approaches, joint frailty models being a framework that is frequently used. A multi-state model with recurrent event states and an absorbing state representing the terminal event can also be used in this context. In **Study IV,** I am interested in evaluating via simulation the use of a multi-state model with recurrent states and a competing terminal absorbing state, with and without restrictions among the baseline transition intensity rates, when the underlying data generating mechanism follows a joint frailty model. I focus on the probabilities of death and of a new recurrent event across follow-up time given zero, one, two or three previous recurrences up to the first year of the follow-up, probability measures that can be targeted by both a joint frailty and a multi-state model. Then the bias and relative precision of the different modelling approaches are evaluated. Finally, I engage in a discussion of the similarities, the different assumptions and the focus of each framework.

LIST OF SCIENTIFIC PAPERS

I. **Nikolaos Skourlis**, Michael J. Crowther, Therese M.-L. Andersson, Paul C. Lambert. On the choice of timescale for other cause mortality in a competing risk setting using flexible parametric survival models. *Biometrical Journal 64.7 (2022): 1161-1177*

II. **Nikolaos Skourlis**, Michael J. Crowther, Therese M.-L. Andersson, Paul C. Lambert. Development of a dynamic interactive web tool to enhance understanding of multi-state model analyses: MSMplus. *BMC medical research methodology 21.1 (2021): 1-9.*

III. **Nikolaos Skourlis**, Michael J. Crowther, Therese M-L. Andersson, Donghao Lu, Mats Lambe, Paul C. Lambert. Exploring different research questions via complex multi-state models when using registry-based repeated prescriptions of antidepressants in women with breast cancer and a matched population comparison group.
Submitted Manuscript 2022

IV. **Nikolaos Skourlis**, Michael J. Crowther, Therese M.-L. Andersson, Paul C. Lambert. Evaluating the use of multi-state models when studying recurrent events in the presence of a terminal event.
Manuscript

CONTENTS

LIST OF ABBREVIATIONS

AG	Andersen-Gill
AIC	Akaike information criterion
AJ	Aalen-Johansen
ATC	Anatomical Therapeutic Chemical
BC	Breast Cancer
BCBaSe 2.0	Breast Cancer Data Base Sweden 2.0
BIC	Bayesian information criterion
CIF	Cumulative Incidence Function
CR	Competing risks
CSH	Cause-Specific Hazard
DDD	Defined daily dose
DGM	Data Generating Mechanism
DM	Distant Metastasis
DoH	Declaration of Helsinki
EBMT	European Blood and Marrow Transplant
FPSM	Flexible Parametric Survival Model
GDPR	General Data Protection Regulation
ICD	International Classification of Diseases
LISA	Longitudinal Integration Database for Health Insurance and Labour Market Studies
LR	Local Recurrence
MSM	Multi-state models
MIDAS	MIcro Data for Analysis of Social Insurance
PWP	Prentice-William-Peterson
RP	Relative Precision
SCR	Swedish Cancer Registry
WLW	Wei-Lin-Weissfeld
WMA	World Medical Association

1 INTRODUCTION

Competing risks and multi-state models are survival models that allow the study of one or more events/ states of interest, accounting for competing events that may influence the observation of the main events of interest and allowing the study of complex disease pathways, addressing composite research questions. When using these models in epidemiology, there are different modelling assumptions and choices to be made, such as the choice of timescale for each transition, the shared or separate estimation of parameters and the choice of the multi-state structure. Even after fitting a competing risks or a multi-state model, the communication of the results is not always straightforward due to the time dimension of the predictions and the multitude of measures that can be derived. In addition, for some cases of time-to-event data, alternative modelling frameworks can be applied. For example, when studying recurrent events in the presence of a terminal event, joint frailty models are a commonly used approach, but multi-state models can also be used.

In the present thesis, I focus on exploring and assessing the choice of timescale when implementing competing risks and multi-state models, the choice of multi-state structure and the choice of sharing information when applying multi-state models. These assessments are done either via simulation (**Study I**), or via development, implementation and discussion of different multi-state structures and sensitivity analyses of different modelling choices (**Study III**). I developed an R package and an online RShiny application called MSMplus, aimed to facilitate the communication of multi-state structures and estimated measures from the application of MSM, using interactive graphs (**Study II**). Finally, I implemented a recurrent multi-state structure to study recurrent events in the presence of a terminal event when the underlying data generating process is that of a joint frailty model and, deriving useful estimates and evaluating their bias while also discussing similarities and differences between the two frameworks (**Study IV**).

2 LITERATURE REVIEW-BACKGROUND

2.1 COMPETING RISKS AND MULTI-STATE MODELS IN CANCER EPIDEMIOLOGY

Cancer survival analysis is based on time from diagnosis of the cancer type under study until a pre-defined event, that event often being death due to that cancer. This standard analysis setting can be thought of as the transition from one state (alive) to another state (death due to cancer), and the hazard rate for the event as the transition intensity rate between the two states (see Figure 2.1a).

However, disease processes are often complex and many disease related events can occur on the path from diagnosis to death. In addition, not all cancer patients die from their cancer. For example, many patients diagnosed with cancer are of old age and are therefore at risk of dying from a number of causes other than their diagnosed cancer. Multi-state models is a framework that extends standard survival models by including more than one transition and more than two states (1) with competing risks models being essentially a special case of multi-state models where all states are absorbing (terminal) states (e.g deaths due to different causes). Multi-state models enable the detailed analysis and understanding not only of the overall hazard for the event of interest but also of the disease process' history via the estimation of multiple, clinically significant measures.

In the area of cancer epidemiology, competing risk models have been extensively applied in settings where the event that is relevant to the cancer of interest, for example, cancer incidence or mortality, may not be observed due to the presence of competing events (2–11). Multi-state models have also been applied, albeit much less compared to competing risk models, when studying several types of cancer such as breast cancer (10,12–18), lung cancer (19,20), colon cancer (21), pancreatic cancer (22) and prostate cancer (23). Most of their applications have been applied to epidemiological data but applications in clinical trials data are increasingly present in the literature (24–27). Applications of multi-state models on cancer screening also exist (28–31).

2.2 COMPETING RISKS

Competing risks occur when an individual can experience one or more terminal/ absorbing outcomes which 'compete' with the outcome of interest (32) and may prevent it from being observed. For example, an individual diagnosed with cancer can die from various other causes than the diagnosed cancer (see Figure 2.1b for example). When competing risks are present a typical survival analysis for a single terminal event does not give the probability that the individual will actually experience (or not experience/survive) the event of interest (33). The probability of experiencing the event of interest as well as the competing events is targeted by the cause-specific cumulative incidence function (CIF) measure. In a competing risks analysis setting, the cause-specific CIF for each event can be estimated a) non-parametrically (34), b) semi-parametrically, via cause-specific Cox hazard models (35) or sub-distribution hazard models (36) and parametrically via cause-specific hazard (CSH) models (37). The sub-distribution hazard modelling approach directly models the CIFs while the cause-specific hazard modelling approaches do so indirectly.

2.2.1 **Cause-specific hazard models**

If we denote k as each competing event with $k \in (1, \dots K)$, and time T as the time until that event, then, given covariates of interest $\boldsymbol{Z}$, the cause-specific hazard function can be defined as:

$$h_k(t|\boldsymbol{Z_i}) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t, Event = k | T \geq t, \boldsymbol{Z_i})}{\Delta t} \quad (1)$$

2.2.2 **CIF definition**

The cause-specific survival functions for the i^{th} individual for the competing event k can be expressed as a function of the cause-specific hazard:

$$S_k(t|\boldsymbol{Z_i}) = \exp\left(- \int_0^t h_k(u|\boldsymbol{Z_i})\ du \right) \quad (2)$$

In a competing risk setting with K competing events, the probability of having a particular event k by time t, the cause-specific CIF is a function of all K cause-specific hazard rates.

The definition of the cause-specific CIF_k for the k^{th} event is given by:

$$CIF_k(t|\boldsymbol{Z_i}) = \int_0^t \left(\prod_{k=1}^{K} S_k(u|\boldsymbol{Z_i}) \right)\ h_k(u|\boldsymbol{Z_i})\ du \quad (3)$$

It is important to note that $\prod_{k=1}^{K} S_k(u|\boldsymbol{Z_i})$, which is essentially the all-cause survival function, is a function of the cause-specific hazard functions for all the competing events and that $CIF_k(t|\boldsymbol{Z_i})$ is a non-linear function of all cause-specific hazard functions.

2.2.3 **CIF estimation**

In the present thesis, I focus on the estimation of the cause-specific hazard rates via flexible parametric survival models (FPSMs), developed by Royston and Parmar (38), later extended by Lambert and Royston (39), that can flexibly model the effect of (ln)time for the log baseline cumulative hazard $\ln[\widehat{H_k}(t|\boldsymbol{Z_i})]$. More details about FPSM are provided in section 4.2.1.

Estimates on the hazard scale can easily be obtained by calculating the derivative of the exponential of $\ln[\widehat{H_k}(t|\boldsymbol{Z_i})]$:

$$\widehat{h_k}(t|\boldsymbol{Z_i}) = \frac{d\exp\left\{\ln\left[\widehat{H_k}(t|\boldsymbol{Z_i})\right]\right\}}{dt} = \frac{d\widehat{H_k}(t|\boldsymbol{Z_i})}{dt} \quad (4)$$

Estimates of the survival probabilities $\widehat{S_k}(t|\boldsymbol{Z_i})$ can also be derived via calculating the exponential of the minus exponential of $\ln[\widehat{H_k}(t|\boldsymbol{Z_i})]$:

$$\widehat{S_k}(t|\boldsymbol{Z_i}) = \exp\{-\exp(\ln[\widehat{H_k}(t|\boldsymbol{Z_i})])\} = \exp\{-\widehat{H_k}(t|\boldsymbol{Z_i})\} \quad (5)$$

Finally, the $\widehat{h_k}(t|\boldsymbol{Z_i})$ and $\widehat{S_k}(t|\boldsymbol{Z_i})$ estimates can be plugged in the CIF formula (equation 3) to derive the CIF estimates. Gaussian quadrature is used to numerically approximate the integral of equation 3 with the plugged-in survival and hazard estimates while the delta method is used in order to derive confidence intervals for the CIFs (40).

2.3 MULTI-STATE MODELS

Multi-state models are more general than competing risks models and can consider complex pathways between initial and absorbing states, frequently including intermediate/transient states. A typical multi-state setting is the so called "Illness-Death model" where an individual can go to the absorbing state (death) either directly or after passing through an intermediate state (e.g cancer recurrence for individuals previously treated for cancer). Such an example of multi-state setting is Figure 2.1c. Multi-state models are used in a variety of epidemiological settings, enabling the study of individuals through different disease states. Studying acute (41), chronic disease progression (24), or recurrent events such as repeated hospitalizations (42) are typical examples of multi-state model use. When studying such processes, multi-state models are used both in the epidemiological research as well as in health economics, in order to portray accurately, with sufficient complexity the real-world issue under study and provide useful and meaningful predictions.

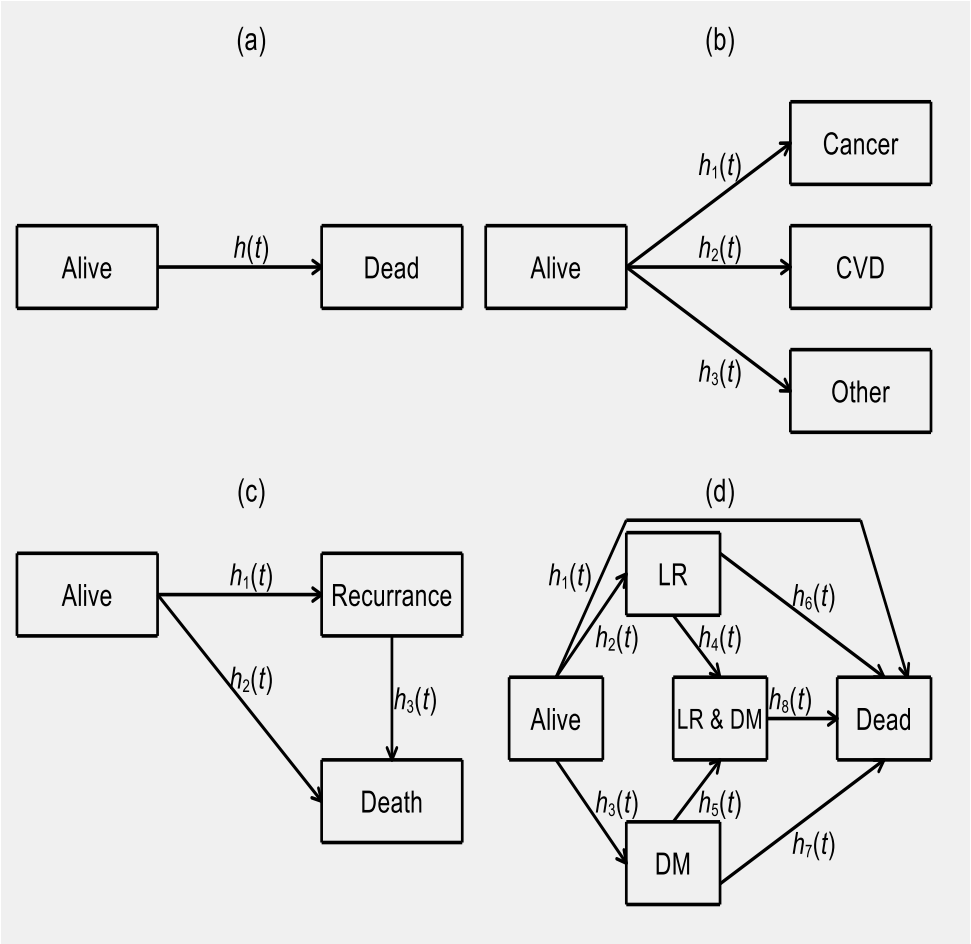
2.3.1 Examples of multi-state structures

The estimates of primary interest during the statistical analysis of a multi-state process are the transition/state occupation probabilities and the transition-specific covariate hazard ratios, with transition intensity (hazard) rates also being very important in understanding the process under study. A variety of other measures that facilitate a more in-depth understanding of the multi-state process can also be estimated depending on the models used and their assumptions such as restricted expected length of stay in a state, probability of ever visiting a state, expected number of visits in each state and more (43,44). Figure 2.1 shows four multi-state model examples from the perspective of a population diagnosed with cancer, with increasing complexity. Figure 2.1a shows a standard survival analysis, where it is only possible to move from the initial "alive" state to the "death" (from any cause) state. Figure 2.1b depicts a competing risks setting with death due to cancer being the event of interest but other competing events exist as well (cardiovascular disease, other causes). In this type of setting, an individual is at risk of death from a number of different causes, but it is only possible to experience one of them; the 3 death states are known as absorbing states as it is not possible to leave them. The final two settings (Figure 2.1c and Figure 2.1d) have transient-intermediate states. Figure 2.1c shows a 3-state Illness-Death model where an individual treated for cancer is at risk of recurrence and may experience death before or after recurrence. Figure 2.1d shows a more complex example of a multi-state setting where an individual is at risk of local recurrence (LR) or distant metastasis (DM), but also at risk of death. This is a complex setting, as there are

multiple potential transitions and different disease routes for the individual to take (different disease histories). In this case of Figure 2.1d, there are eight transition intensity rates (hazard functions) denoted by $h_k(t)$, that need to be modelled. With population registry data, because of large sample sizes, it is possible to model the transition intensity rates (jointly or separately) and thus develop an improved understanding of how the risk factors under study are associated with the whole disease process.

When using multi-state models there are various assumptions that need to be considered. Under the Markov assumption we assume that the transition intensity rates depend on the history of the process only through the current state. This assumption can be relaxed through a semi-Markov assumption where transition intensity rates are a function of time since entering the current state (45). These assumptions can be further relaxed by incorporating more than one time-scale for certain transitions (46) and by using different time-scales for different transitions (**Study III**).

Figure 2.1. Examples of multi-state models taken from the perspective of a population diagnosed with cancer. For (d) LR = Local Recurrence, DM = Distant metastasis.



2.3.2 Transition intensity rates

Consider a stochastic process $Y(t)$ $(t \geq 0)$ with a finite space of states $\Omega = 1, \ldots, L$ and a history of the process defined at time s as $\boldsymbol{H_{s-}} = Y(u); 0 \leq u \leq s$. Then, according to the multi-state structure that corresponds to that process, a transition rate matrix is defined, a matrix of all possible transition intensity rates between states (See Section 2.3.3). Let a, b be states of

the state space Ω. For process $Y(t)$, with K potential transitions ($k = 1,.,K$), each transition uniquely links a state a with another state b $(a \rightarrow b)$. Under the assumption that the stochastic process for the next state depends only on the current state (Markov assumption), the definition of the transition intensity rate does not depend on the history of the process $\boldsymbol{H_{s-}}$ before state a and thus can be defined as:

$$h_{ab}(t) = \lim_{\delta t \to 0} \frac{P(Y(t+\delta t) = b | Y(t) = a)}{\delta t} \tag{6}$$

We can relax the Markov assumption by assuming that the times to each next state b depend only on the present state and the time since entry of that state (semi-Markov assumption) (43), treating the transition intensity rates as functions of time since entering state a and/or assume a state arrival extended (semi-)Markov model where the transition intensity rates rely on the time that state a was entered as a covariate in the model (1).

2.3.3 **Transition rate matrix**

Let's also define the transition intensity/rate matrix $\boldsymbol{Q}$ as an array of the instantaneous rates h_{ab} (for $a \neq b$) between states, with diagonal elements being equal to $h_{aa} = -\sum_{b \neq a} h_{ab}$. Let's take the 3-state Illness-Death model of Figure 2.1c as an example of a continuous multi-state structure/ process, with state space $\Omega = \{1,2,3\}$ (Alive=State 1, Recurrence= State 2, Death= State 3). The transition intensity matrix for this multi-state structure can be depicted as:

$$\boldsymbol{Q}(t) = \begin{bmatrix} -(h_{12}(t) + h_{13}(t)) & h_{12}(t) & h_{13}(t) \\ 0 & -h_{23}(t) & h_{23}(t) \\ 0 & 0 & 0 \end{bmatrix} \tag{7}$$

with each row specifying the number of the starting state of a transition and each column specifying the ending state of a transition. Transitions that cannot happen have a transition intensity rate of zero over time (e.g $h_{21}(t) = 0$).

2.3.4 **Transition probabilities**

Similarly to the competing risks setting and the definition of the cause-specific CIFs, the transition probabilities can be expressed as non-linear functions of transition intensity rates and survival functions. The probability of being in state b at time t given being in state a at time s, can be defined as:

$$P(Y(t) = b | Y(s) = a, \boldsymbol{H_{s-}}) \tag{8}$$

where the term $\boldsymbol{H_{s-}}$ can be dropped under the Markov assumption**.** Staying in the example of the 3-state Illness-Death model of Figure 2.1c, we will define the conditional cause-specific survival functions between states, as these feed into the equations for deriving the transition probabilities. The conditional cause-specific survival from transitioning between State 1 and State 2, given being in State 1 at time s, $S_{12}(t|s)$, can be defined as:

$$S_{12}(t|s) = \frac{S_{12}(t)}{S_{12}(s)} = \frac{\exp\left(-\int_0^t h_{12}(u)\ du\right)}{\exp\left(-\int_0^s h_{12}(u)\ du\right)} = \exp\left(-\int_s^t h_{12}(u)\ du\right) \quad (9)$$

where $t \geq s$, noting that t is time since the origin. The cause-specific survival functions from State 1 to State 3, $S_{13}(t|s)$ and from State 2 to State 3, $S_{23}(t|s)$ are defined similarly.

The probability of staying in a state can be thought of as the probability of surviving from experiencing any of the possible states an individual can potentially transition to, and thus can be expressed as a function of survival probabilities. The probabilities for the 3-state Illness-Death multi-state structure can be written as:

$$P_{11}(s,t) = P(Y(t) = 1|Y(s) = 1) = S_{12}(t|s)\, S_{13}(t|s) = \exp\left(-\int_s^t h_{12}(u) + h_{13}(u)\ du\right) \quad (10)$$

$$P_{22}(s,t) = P(Y(t) = 2|Y(s) = 2) = S_{23}(t|s) = \exp\left(-\int_s^t h_{23}(u)\ du\right) \quad (11)$$

$$P_{33}(s,t) = P(Y(t) = 3|Y(s) = 3) = 1 \quad (12)$$

In order to define the probability of being in state b at time t given being in state a at time s, with $a \neq b$, we need to consider the probability of entering state b at any intermediate time $s < u \leq t$ and then remain at state b until time t. In our 3-state Illness-Death model setting, for $a = 1$ and $b = 2$, the transition probability can be defined as:

$$P_{12}(s,t) = P(Y(t) = 2|Y(s) = 1) = \int_s^t P_{11}(s,u)h_{12}(u)P_{22}(u,t)\, du = \int_s^t S_{12}(u|s)S_{13}(u|s)h_{12}(u)S_{23}(u|s)\, du = \int_s^t \exp\left(-\int_s^u h_{12}(w) + h_{13}(w)\ dw\right)h_{12}(u)\ \exp\left(-\int_u^t h_{23}(w)\ dw\right)du \quad (13)$$

with three components, i) the probability of remaining in State 1 (surviving from transitioning to State 2 or State 3) until time u, ii) the instantaneous risk of transitioning to State 2 at time u and iii) the probability of remaining in State 2 until time t or, put otherwise, the probability of "surviving" the transition to State 3, integrated over all possible times u.

The probability of reaching the absorbing state, in this case State 3, by time t given being in State 1 (Initial state) at time s, is equal to the probability of not managing to stay in State 1 and not managing to stay in State 2, up until time t:

$$P_{13}(s,t) = P(Y(t) = 3|Y(s) = 1) = 1 - P_{11}(s,t) - P_{12}(s,t) \tag{14}$$

Correspondingly, the probability of reaching the absorbing state, State 3, by time t given being in the intermediate state, State 2, at time s, is the probability of not managing to stay in State 2 up until time t:

$$P_{23}(s,t) = P(Y(t) = 3|Y(s) = 2) = 1 - S_{23}(t|s) = 1 - P_{22}(s,t) \tag{15}$$

The transition probabilities $P_{22}(s,t)$ and $P_{23}(s,t)$ can be derived similarly under a semi-Markov assumption, with the hazard and survival functions being functions of time since entering State 2, $h_{23}(t-r), S_{23}(t-r)$, where r the time of entering State 2.

2.3.5 Expected length of stay

Aside from the aforementioned measures (transition probabilities and transition intensity rates) there are other clinically significant measures such as restricted expected length of stay (or length of stay for simplicity) in state b during the time period from s to t, conditional on the patient being in state a (non-absorbing) at time s, is defined as

$$e_{ab}(s,t) = \int_s^t P(Y(u) = b|Y(s) = a)\ du \tag{16}$$

which defines the amount of time spent in state b, starting in state a at time s, up until time t. If $t = \infty$, and state $a = b$ is a healthy state and all possible next states are deaths, then equation 16 represents life expectancy.

2.3.6 Probability of ever visiting a state

Another measure of clinical importance is the probability of ever visiting a state b by time τ given being at state a at time s, that is the cumulative probability of ever entering state b over the time period $[s,\tau]$ and can be defined as:

$$v_{ab}(s,\tau) = P(Y(\forall t \in [s,\tau]) = b|\ Y(s) = a) \tag{17}$$

where symbol ∀ signifies "for any".

2.3.7 **Rationale for using competing risks and multi-state models**

Firstly, it is important to clarify that fitting a standard survival model is not wrong. In the examples in Figure 2.1, a research question can be answered by modelling just one of the transitions between states. For example, in Figure 2.1d that depicts a complex multi-state setting, the researcher can study how an individual's traits and the characteristics of the diagnosed tumor are associated with the rates of distant metastasis in breast cancer patients (47). However, the separate analysis of each transition fails to reveal the associations between the different types of events (48). Therefore, while studying a single transition via standard survival analysis allows comparison of rates, it fails to estimate real-world probabilities of being in a certain disease state. In order to estimate probabilities of being in each disease state, as well as the rest of the aforementioned measures, modeling of all relevant events is required. A well-known showcase example of the previous argument is the fact that a cause-specific survival curve does not estimate the probability of dying from the particular cause, as this probability (CIF) depends also on the mortality rate due to other causes as was shown in equation 3 and an appropriate analysis will take the competing risk into account (49). The same issue extends beyond the competing risk setting to the multi-state setting as well, when interest lies in understanding the impact of a risk factor over the whole process of the disease (1).

3 RESEARCH AIMS

The current research focusses on a number of extensions in the areas of competing risks and multi-state models using flexible parametric survival models (FPSMs). These include consideration of different time scales for different transitions, sharing of information across transitions by imposing restrictions in the parameter estimation process, application and interpretation of a range of multi-state structures of increasing complexity and deriving complex (but useful) predictions from such models. The research projects also investigate the use of multi-state models in comparison with other modelling frameworks and explore ways to efficiently communicate structures and estimation results from multi-state models by developing an interactive RShiny application (MSMplus). The methods are applied to national cancer registration data (50) and a more detailed linked breast cancer data base, BCBaSe 2.0 (51), which links to numerous other registers including the Prescribed Drug Register (52).

The specific aims of the current research are:

- To incorporate different time scales (attained age and time since diagnosis) when modelling competing risks using flexible parametric survival models and assess the influence of the choice of timescale on the estimation of the cause-specific CIFs for different levels of selected factors (**Study I**).
- To present and compare the structures and estimation results of multi-state models in a novel, flexible and interactive way via developing an online interactive web tool (**Study II**).
- To explore different research questions via multi-state models of varying complexity when using registry-based repeated prescriptions of antidepressants in women with breast cancer and a matched population comparison group and discuss different modelling choices (**Study III**).
- To use and evaluate a multi-state model framework as an alternative to joint frailty models when studying reccurrent events with a presence of a terminal event (**Study IV**).

4 MATERIALS AND METHODS

4.1 DATA SOURCES

4.1.1 Swedish Cancer Registry

In **Study I**, I use data on colon cancer diagnosis from the Swedish Cancer Register as a motivational example of applying competing risks using either a common timescale or different timescales for the different causes of death (colon cancer and other causes). The Swedish Cancer Register is a nationwide, population-based register, established in 1958 (50). All health care practitioners and laboratories are compelled by law to report any new cancer diagnosis so registration is close to complete. The register includes demographic information (age, gender, place of residence), tumour-specific information (site, stage and histological type), information about diagnosis (date, reporting hospital and more) as well as follow-up date (date and cause of death, or date of migration). In **Study I,** I limit the sample to all adults individuals that were diagnosed with colon cancer between 2005 and 2017 (n=53,630).

4.1.2 Swedish Cause of Death Register

Information on date and cause of death regarding the individuals with colon cancer diagnosis of **Study I** was retrieved from the Cause of Death Register (53). Statistics on cause of death started being reported in 1749 by decision of the Swedish parliament. Different governmental organizations bore the responsibility of the register between 1831 and 1911, including as causes of death maternal death or plague. From 1911 to 1993, Statistics Sweden was responsible for the register, including all causes of death. From 1994 to present, the register is maintained by the Swedish National Board of Health and Welfare and is updated on an annual basis. It includes demographic information, underlying cause of death (ICD-6 to ICD-10), date and place of death, autopsy, surgery within a month before death and more. In **Study I**, I use the underlying cause of death information to classify an individual as having died from colon cancer, from other causes or still being alive at the end of the follow-up period. I use the register version with data up to the end of 2017. I excluded individuals for which the underlying colon cancer was found during an autopsy.

Record linkage between the Swedish Cancer Register and the Swedish Cause of Death Register for **Study I** is straightforward via using the unique civic registration number assigned to all Swedish citizens. The linkage between the two registers has been conducted in a pre-analysis stage, and the data has been pseudonymized.

4.1.3 Breast Cancer Data Base Sweden 2.0

Breast Cancer Data Base Sweden 2.0 (BCBaSe 2.0) is a register-based research resource with data on an unselected cohort of women and men diagnosed with breast cancer (BC) in Sweden (51). It consists of individuals diagnosed with BC between 1992 and 2012 (n = 68,450) that are age and gender matched with breast cancer free controls with a ratio of approximately 1:5 (n =343,200), in three Swedish Health Care regions. The mean age at inclusion was 61.8 years (range 19-102) and the cohort has been followed up until 31 December 2013. BCBaSe 2.0

includes information from regional and national BC quality registers which is then linked to national demographic and health care population-based registers (Swedish Cancer Register (50), Cause of Death Register (53), National Patient Register (54), Register of total population (55), Prescribed Drug Register (52), Multi-Generation Register (56), Longitudinal Integration Database for Health Insurance and Labour Market Studies or LISA (57) and MIcro Data for Analysis of Social Insurance or MiDAS (58)). Thus, there is information on tumor characteristics and cancer treatment, but also information on socioeconomic variables, comorbidity and use of prescribed drugs. In **Study III**, I use mainly data from the Prescribed Drug Register in order to explore different research questions regarding antidepressant medication use patterns among a sub-sample of women diagnosed with invasive breast cancer and healthy-matched women via the use of multi-state models.

4.1.4 Prescribed Drug Register

The Prescribed Drug Register (52) was initiated in 2005 and contains full data about medication dispensed at pharmacies in Sweden since 01/07/2005, with unique patient identifiers for all dispensed prescriptions in Sweden. The data is collected by the National Corporation of Swedish Pharmacies. The register includes, among other information, demographic information, product-specific information (Anatomical Therapeutic Chemical (ATC) code, name, pack size, recommended daily dose), prescription-specific information (quantity and number of packages, date of prescription, date of purchase, prescribed dosage), as well as cost-specific and prescriber-specific information. Each row of the register database corresponds to one dispensation at a pharmacy. In **Study III,** I am interested in exploring different research questions via multi-state models using registry-based repeated prescriptions of antidepressants, so I selected the rows/ prescribed medication based on the Anatomical Therapeutic Chemical (ATC) classification system, using the code N06A.

4.1.5 European Blood and Marrow Transplant registry example

The European Blood and Marrow Transplant (EBMT) registry (59) is the backbone of the EBMT's research activities. An example dataset consisting of 2.204 individuals who have received bone marrow transplantation is freely available via library `msm` in R (ebmt3). It is a typical example dataset used for the application of an Illness-Death multi-state model as the individual, may experience platelet recovery or relapse/death. The relapse/death event may take place with or without previous platelet recovery. This dataset is used in **Study II**, as an example for the Rshiny application I develop called MSMplus, an online interactive tool for communication of multi-state model results to scientific audiences.

4.1.6 Readmission data

This is a freely available dataset that is used as an example dataset within the `frailtypack` package in R (60) and is used in **Study IV** to compare the predictions from the application of the joint frailty models with the predictions from the various multi-state modelling approaches. It includes information about repeated hospitalizations among 403 colorectal cancer patients with death as a terminal event and the origin of this data is a study conducted by Gonzalez et al (61).

4.1.7 Ethical Considerations

Any research that involves human participants should follow strict ethical principles so that it does not endanger their physical or mental health or integrity. The Declaration of Helsinki (DoH) is the World Medical Association's (WMA) best-known policy statement (62). It contains a set of principles that set a boundary between the potential profit for society and the interests/well-being of the individual patients who are part of clinical trials, recognizes the existence of vulnerable groups (children, inmates, soldiers and more) and promotes the right of the individual to make informed decisions. The first version was adopted in 1964 and has been amended seven times since, most recently at the General Assembly in October 2013. Even though the declaration itself is not legally binding, many countries have passed its principles as national laws. Specifically in Sweden, the Ethical Review Act (2003:460), the Public Access and Secrecy Act (2009:300) and the personal Data Act (1998:204) incorporated the DoH principles in national legislation. Then, under the EU directive in order to harmonize relevant personal Data Act legislation across Europe, the General Data Protection Regulation (GDPR) was implemented in 2018 (63). The ethical review act provides details about the information that should be communicated to the participant, the requirement of consent, the right of withdrawal, cases of research where consent is not required, due to a mental disorder or an weakened state of being, as well as the distribution of responsibility for checking and granting research among ethical review boards and departments across the country. In cases when no active participation is required from the participants, the Ethical review authority can waive the requirement of informed consent.

In the present thesis, two types of datasets are used: a) toy example datasets that are freely available on web and can be used as motivational examples for the application of a technique (52,59) and b) datasets originating from the Swedish Cancer Register (50) and BCBaSe 2.0 (51) that are based on the population-based registers mentioned in Section 4.1.3. The first class of datasets do not require any kind of ethical approval in order to be used, while the second-class datasets require permission from the ethical review authority in order to be dispatched from the register holders to the institution that will serve as a host for the researcher to analyze them. The Regional Ethics Review board in Stockholm granted ethical approval for the use of the Swedish cancer registry data for the aims included in **Study I** of this thesis (2006/914-31/3, 2008/1469-32, 2009/634-32, 2010/1928-32). It also granted ethical approval for the use of the BCBaSe 2.0 register-based data source for the aims included in **Study III** of this thesis (2013/1272-31/4). The data has been pseudonymised, encrypted and properly stored at servers of the Department of Epidemiology and Biostatistics (MEB). Access is given only to researchers involved in the data management and analysis of the data regarding a project with aim that is included in the granted ethical approvals.

4.2 FLEXIBLE PARAMETRIC MODELS

Royston & Parmar (38) developed a class of flexible parametric models which were later extended by Lambert & Royston (39). This type of models allow both for right censoring and left truncation and use restricted cubic spline functions $s()$ to flexibly model the effect of the logarithm of time $s(\ln(t)\,|\boldsymbol{\gamma}, \boldsymbol{m_0})$ for the log baseline cumulative hazard, with M_0 knots, $\boldsymbol{m_0} = (m_{0_1}, m_{0_2}, \dots, m_{0_{M_0}})$ is the knots vector with associated parameter vector $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \dots, \gamma_{M_0-1})$.

The restriction of the restricted cubic spline function is that it is linear before the first knot and after the last knot. In order to fit a restricted cubic spline function with M_0 knots, $M_0 - 1$ variables must be created so that the cubic spline function can be defined as:

$$s(\ln(t)\,|\boldsymbol{\gamma}, \boldsymbol{m_0}) = \gamma_0 + \gamma_1 z_1 + \cdots + \gamma_{M_0-1} z_{M_0-1} \tag{18}$$

Where variables $z_j, j \in \{1, \ldots, M_0 - 1\}$ are known as the basis functions and can be derived as:

$$\begin{gathered} z_1 = \ln(t), \\ z_j = \left(\ln(t) - m_{0_j}\right)_+^3 - \frac{m_{0_{M_0}} - m_{0_j}}{m_{0_{M_0}} - m_{0_1}} \left(\ln(t) - m_{0_1}\right)_+^3 \\ - \left(1 - \frac{m_{0_{M_0}} - m_{0_j}}{m_{0_{M_0}} - m_{0_1}}\right) \left(\ln(t) - m_{0_{M_0}}\right)_+^3, \\ j = 2, \ldots, M_0 - 1 \end{gathered} \tag{19}$$

This class of models enable proportional hazards, but can easily be extended to time-varying effects (non-proportional hazards) by including interaction terms between the covariates and the timescale. A flexible parametric proportional hazards model on the log cumulative hazard scale $\ln H(t)$, with time since diagnosis as the timescale t, and covariate vector $\boldsymbol{Z}$ will be:

$$\ln\left(H(t|\boldsymbol{\gamma}, \boldsymbol{m_0}, \boldsymbol{Z_i})\right) = \ln(H_0(t)) + \boldsymbol{\beta_Z Z_i} = s(\ln(t)\,|\boldsymbol{\gamma}, \boldsymbol{m_0}) + \boldsymbol{\beta_Z Z_i} \tag{20}$$

while a flexible parametric non-proportional hazards model with time-dependent effects for the covariate vector $\boldsymbol{Z}$ will be:

$$\ln[H(t|\boldsymbol{\gamma}, \boldsymbol{m_0}, \boldsymbol{\delta_m}, \boldsymbol{m_j}, \boldsymbol{Z_i})] = s(\ln(t)\,|\boldsymbol{\gamma}, \boldsymbol{m_0}) + \boldsymbol{\beta_Z Z_i} + \sum_{j=1}^{D} s\left(\ln(t)\,\middle|\boldsymbol{\delta_m}, \boldsymbol{m_j}\right) \boldsymbol{Z_i} \tag{21}$$

with D the number of time dependent effects, $\boldsymbol{m_j}$, the knots for the j^{th} time-dependent effect with parameters $\boldsymbol{\delta_m}$.

The survival and the hazard functions can then be derived as:

$$S(t|\boldsymbol{Z_i}) = \exp\{-\exp(\ln[H(t|\boldsymbol{\gamma}, \boldsymbol{m_0}, \boldsymbol{Z_i})])\} \tag{22}$$

$$h(t|\boldsymbol{Z_i}) = \frac{d\exp(\ln[H(t|\boldsymbol{\gamma}, \boldsymbol{m_0}, \boldsymbol{Z_i})])}{dt} = \frac{d(s(\ln(t)|\boldsymbol{\gamma}, \boldsymbol{m_0}))}{dt} \exp(\boldsymbol{\beta_Z Z_i}) \tag{23}$$

The parameter estimates are derived via maximum likelihood estimation. The contribution of the i^{th} individual to the log-likelihood of a right censored FPSM is

$$\ln L_i = d_i \ln[h(t_i|\boldsymbol{\gamma}, \boldsymbol{m_0}, \boldsymbol{\beta_Z})] - H(t_i|\boldsymbol{\gamma}, \boldsymbol{m_0}, \boldsymbol{\beta_Z}) \quad (24)$$

See Royston and Parmar (38) for more details.

4.3 PARAMETRIC APPROACHES IN MULTI-STATE MODELS

4.3.1 Earlier parametric approaches

Omar et al (64) developed an Illness–Death model assuming all transition intensity rates to be Weibull hazard functions. In 2011, Jackson (44) presented, via package `msm` in R, a (piecewise-) exponential multi-state model under a Markov assumption, with piecewise constant transition intensity rates, that can capture complex hazard functions. In 2015, Titman (65) developed a multi-state model using B-splines when modelling the baseline transition intensity rates (non-homogeneous) under the Markov assumption and derived the transition probabilities via the development of an equation solver. Krol and Saint Pierre (66) developed a multi-state model under a semi-Markov assumption allowing for transition-specific intensity rates but only offer the options of exponential, Weibull and exponential-Weibull. Blaser et al (67) developed a more flexible parametric multi-state model, with transition-specific intensity rates, each of which can be an arbitrarily specified hazard function, using piecewise approximation when simulating survival times to derive predictions under a semi-Markov approach.

4.3.2 Flexible parametric survival models

Flexible parametric multi-state models, either under the Markov or the semi-Markov assumption, with transition-specific definition of the hazard distribution can be fit via packages `flexrsurv` (68) and `rstpm2`(69) in R and via the `multistate` package in Stata (43). In addition, the parameter estimates can be shared or separate across transitions while some or all transitions can incorporate time-dependent effects. Adopting a flexible parametric approach offers a number of advantages, the most important being the flexible modelling of time dependent effects or in other words, non-proportional hazards, in a much easier and more straightforward way, thus providing greater model flexibility and thus the potential for more accurate predictions. These advantages have already been demonstrated in a competing risk setting (37,40,70,71) and practical applications of these methods can be found in the relevant literature (72–74). A natural extension to the above is the use of flexible parametric approaches to a more complex multi-state framework that focuses on studying the entire disease history (43). A parametric approach makes the extensions feasible, leading to more complex, but at the same time more realistic models, thus improving our understanding of disease process under study.

4.3.3 Estimation of transition/state occupancy probabilities

Cause-specific parametric models can be used to estimate the cause-specific log cumulative hazard and the cause-specific survival and hazard functions. These can then be plugged-in to the relevant equations to estimate the cause-specific CIFs in the competing risk setting

(equation 3 of Section 2.2) and the transition probabilities in the multi-state setting (equations 8-13 of Section 2.3). There are alternative approaches for estimating the transition probabilities. Consider for example a multi-state process $Y(t)$ $(t \geq 0)$ with a finite space of states $\Omega = 1, \ldots, L$. The transition probabilities can be derived via a parametric adaption of the Aalen-Johansen estimator (equation 25), which is the solution to the Kolmogorov equation (equation 27) under the Markov assumption, by plugging in the cumulative hazard estimates derived from a parametric hazard model:

$$\widehat{\boldsymbol{P}}(s,t) = \prod_{s<u\leq t} \left(\boldsymbol{I} + d\widehat{\boldsymbol{H}}(u)\right) \tag{25}$$

With $\boldsymbol{I}$ being the $L \times L$ identity matrix, and $\mathrm{d}\widehat{\boldsymbol{H}}$ being the $L \times L$ matrix of increments in the cumulative hazards $\widehat{\boldsymbol{H}}$ over all transition combinations. The transition probability for a specific $a \to b$ transition is $(\hat{P}(Y(t) = b|Y(s) = a))$ is equal to the (a, b) element of the $\widehat{\boldsymbol{P}}(s,t)$ matrix. The Aalen-Johansen estimator is usually used for non-parametric estimation. However, if the cumulative hazards are predicted at small time intervals then the parametric approach can be used.

The overall probability of being in each state l over time $P_l(t) = P(Y(t) = l)$ can be derived by multiplying the transition probability towards that state since time zero $P(Y(t) = l|Y(0) = 1)$ with the vector of probabilities of being in each state at time 0, with the l^{th} element being $\pi_l(0) = P(Y(0) = l)$. The estimated state occupation probability matrix can therefore be derived as:

$$\widehat{\boldsymbol{P}}(t) = \widehat{\boldsymbol{\pi}}(0) \prod_{s<u\leq t} \left(\boldsymbol{I} + d\widehat{\boldsymbol{H}}(u)\right) \tag{26}$$

From equation 26 it follows that, if all individuals start from the same, initial state at time 0, then $\pi_1(0) = P(Y(0) = 1) = 1$, in which case the state occupancy probability matrix $\boldsymbol{P}(t)$ coincides with the transition probability matrix $\boldsymbol{P}(s,t)$. In the current work, in all multi-state model examples, all individuals start from the same, initial state, that is why the terms transition probability and state occupancy probability are often used interchangeably.

Transition probabilities can also be derived by numerically solving the Kolmogorov forward equation (75):

$$\frac{d\boldsymbol{P}(s,t)}{dt} = \boldsymbol{P}(s,t) * \boldsymbol{Q}(t) \tag{27}$$

as done by Titman (65), with $\boldsymbol{P}(s,s) = \boldsymbol{I}$ identity matrix being the initial condition.

As it will be discussed in detail in the Section 4.4, transition probabilities as well as other measures can also be derived under simulation-based approaches that essentially simulate a high number of individuals and potential trajectories across the states based on the transition

matrix and the hazard rate estimates from the implementation of the cause-specific hazard models.

4.4 SIMULATION FOR DERIVING PREDICTIONS

Some of the clinically useful measures in the multi-state setting require complex numerical integration to be estimated. In order to bypass the integration issue, a simulation approach can be used. Also, under a semi-Markov assumption, these measures can more easily be obtained via simulation. However, simulating disease histories in multi-state models is a complex procedure and it has been argued that this is one reason why many previous modelling approaches were quite simplistic (46). With increases in the computing power, following the evolution of computer technology, it is now possible to use simulation-based methods to calculate a variety of useful predictions. Parametric models render this procedure even easier as they provide all the parameters needed to construct the data generating mechanism used for the simulation, rather than Cox models where the baseline hazard for each transition is not directly estimated (42). During the simulation approach, the parametric models are fitted and then, using the estimated model parameters, a large number n of individuals are simulated through the model, thus making the predictions a simple process of counting and averaging rather than complex nested numerical integration. The simulation can be repeated m times, using random draws from a multivariate normal distribution of the estimated parameters, with mean $\widehat{\boldsymbol{\beta}}$ and the associated variance-covariance matrix $\boldsymbol{V}$ in order to also derive confidence intervals. A general survival simulation framework can be used (76) that is highly flexible to obtain clinically useful measures such as the probability of being in each disease state as a function of time and the total length of stay in each state (77). In addition, restricted life expectancy can easily be derived as a function of the predicted length of stay in each state. Contrasts of a measure within each state and between individuals with different covariate patterns can also be estimated, such as differences or ratios, also accompanied by confidence intervals. Another metric for contrast is the probability that an exposed individual has a shorter time spent in a certain state when compared to an unexposed individual. This is an example of a measure that is very difficult to be analytically calculated, but has a very simple and straightforward derivation under the simulation approach.

In **Study III**, I use `predictms` command in Stata (43), a simulation-based approach for deriving measure estimates for multi-state models making Markov and semi-Markov assumptions. In **Study IV**, I derive the true values of the probabilities for death under each scenario by simulating a large population sample based on the parameters set for each scenario. I also explore deriving predictions of the same probabilities from a joint frailty model via simulation apart from the analytical approach.

There are different simulation-based approaches that can be used for multi-state models, such as the latent failure times approach (78) and the simulation design by Beyersmann *et al* (79). The latent failure times approach is essentially a series of subsequent competing risks simulations, simulating an event time for each competing event via its cause-specific hazard function, then keeping the minimum of the simulated times, and treating the event with that time as the observed event. In Beyersmann *et al*, the cause-specific hazard functions for all competing events at each step of the multi-state process are summed. Then, based on the total

hazard function, times to event are simulated. For each simulated time, the type of event is given as a result of a binomial (for two competing events) or a multinomial (for more than two competing events) experiment, with a probability for each competing event equal to the cause-specific hazard rate value at the simulated time over the summed hazard function value at the same time.

Below we describe a latent failure times simulation approach in order to transition probability estimates as well as probability estimates of ever visiting a state and estimated restricted expected length of stay in different states, after applying a multi-state model assuming a semi-Markov process.

4.4.1 **Step 1: Fitting the model**

The first step in this process is to fit the selected multi-state model in order to derive the transition intensity rate estimates. Let's assume the 3-state Illness- Death model of Figure 2.1c, with transition intensity rates h_{12}, h_{13}, and h_{23} and covariate vector $\boldsymbol{Z}$ for all transitions. Under the semi-Markov assumption, the transition intensity rates are assumed to be functions of time since entering the current state. This means that the intensities h_{12} and h_{13} are still functions of the total time, with $h_{12}(t|\boldsymbol{Z})$ and $h_{13}(t|\boldsymbol{Z})$, while h_{23} is a function of time $t' = t - t_1$, with t_1 being the time of entering State 2, resulting in $h_{23}(t - t_1|t \geq t_1, \boldsymbol{Z})$. Fitting a multi-state model with separate parametric estimation of each transition intensity rate, for example via FPSMs, will derive the estimated transition intensity rate functions. It is also possible to fit multiple transition intensity rate models simultaneously via a stacked multi-state model, permitting joint parameter estimation and enabling information sharing across transitions (See Section 6).

4.4.2 **Step 2: Simulating individual trajectories**

Based on the multi-state structure, survival times can be simulated for all competing states an individual is at risk for while being in the current state, simulating their individual trajectory across states and time. For an individual who is at the starting state (State 1) of the 3-state Illness-Death model, the competing states are the intermediate event (State 2) and the terminal event (State 3). A survival time is simulated for this individual along with an event indicator for both competing states, (t_1, δ_1) for State 2 and (t_2, δ_2) for State 3, while also setting a maximum follow-up time T_c after which the observations are censored.

Table 4.1. Simulated survival times and indicators from the initial state (State 1) to the intermediate event (State 2) and the terminal/ absorbing event (State 3)

	Starting state to Intermediate event	Starting state to Terminal event
Survival time	$t_1 \leq T_c$	$t_2 \leq T_c$
Indicator	δ_1 =0/1	δ_1 =0/1

If $t_1, t_2 = T_c$, then the individual stays at State 1 across the follow-up time till the censoring. If $t_2 < t_1, T_c$, then the individual transitions to the absorbing State 3 at time t_2. If $t_1 < t_2, T_c$, then the individual transitions to the intermediate state (State 2). There, a survival time t_3 in the intermediate state is simulated with a maximum of $T_c - t_1$ based on the estimated transition intensity rate for the current transition $\widehat{h_{23}}(t - t_1 | t \geq t_1, \boldsymbol{Z})$. If $t_3 = T_c - t_1$ then $\delta_3 = 0$, the individual remains at State 2 is until the censoring at time $T_c - t_1$. If $t_3 < T_c - t_1$ then $\delta_3 = 1$, and the individual transitions to the terminal state at time t_3.

Step 2 is repeated for a large number of pseudo-individuals. We have to note here, that the process is done for individuals with a specific covariate pattern $Z = z_1$ but then the process can enter a loop over many different covariate patterns.

4.4.3 **Step 3: Deriving predictions**

After simulating trajectories over a population of a specific covariate pattern, with every simulated individual starting from State 1 at time 0, the transition probabilities can easily be estimated by counting how many individuals populate each state for each time point of prediction, where time refers to the time since entering the starting state of the simulated individual, in this case, State 1.

Let us assume a table of 10 simulated individuals, all starting in State 1 at time 0, with their corresponding simulated times of reaching State 2 and State 3 and a maximum follow-up time of 5 years.

Table 4.2. Example of simulated individuals and times of entering each state of the 3-State Illness- Death model. A missing value (.) signifies that the specific state was not entered by the individual during the maximum follow-up time. The computed variables of length of stay of each individual in each state by time equal to 2 years after the start of follow-up are also given.

Id	Time until State 2	Time until State 3	Maximum follow-up time	Length of stay in State 1 by t=2	Length of stay in State 2 by t=2	Length of stay in State 3 by t=2
1	1.5	3	5	1.5	0.5	0
2	.	2.5	5	2	0	0
3	4	.	5	2	0	0
4	.	.	5	2	0	0
5	1.5	4	5	1.5	0.5	0
6	3	5	5	2	0	0
7	0.5	1.5	5	0.5	1	0.5
8	.	0.5	5	0.5	0	1.5
9	.	.	5	2	0	0
10	1	3	5	1	1	0

For a certain time point of prediction t_{pred}, and for the j^{th}state of the multi-state structure, the estimated transition probabilities from State 1 to all states of the 3-state Illness-Death

structure will be equal to the number of individuals populating each state at t_{pred}, $N_j(t_{pred})$, divided by the total population starting at the initial state (State 1) at time 0, $N_1(0)$:

$$\hat{P}_{1j}(0, t_{pred}) = \hat{P}(Y(t_{pred}) = j | Y(0) = 1) = {}^{N_j(t_{pred})}\big/_{N_1(0)} \quad (28)$$

For $t_{pred} = 2$, the $\hat{P}_{11}(0,2)$ transition probability (probability of still being in State 1 at time 2 years after the start of the follow-up, is 0.5 as, out of the $N_1(0) = 10$ simulated individuals, $N_1(2) = 5$ of them (id= 2, 3, 4, 6, 9) are still in State 1 for $t_{pred} = 2$. Similarly, the $\hat{P}_{12}(0,2)$ transition probability (probability of being in State 2 at time 2 years after the start of the follow-up, is 0.3 as, out of the $N_1(0) = 10$ simulated individuals, $N_2(2) = 3$ of them (id=1, 5, 10) have entered State 2 and are still in that state by time $t_{pred} = 2$. Finally, the $\hat{P}_{13}(0,2)$ transition probability (probability having entered the absorbing State 3 by time 2 years after the start of the follow-up, is 0.2 as, out of the $N_1(0) = 10$ simulated individuals, $N_3(2) = 2$ of them (id=7, 8) has entered State 3 by time $t_{pred} = 2$.

The probability of ever visiting the j^{th}state by the prediction time t_{pred}, given being in State 1 at time 0, can be easily estimated by dividing the number of simulated individuals who have ever experienced state j by time t_{pred}, $M_j(t_{pred})$, divided by the total population starting at the initial state at time 0, $N_1(0)$:

$$\hat{v}_{1j}(0, t_{pred}) = \hat{P}(Y(\forall t \in [0, t_{pred}]) = j | Y(0) = 1) = \frac{M_j(t_{pred})}{N_1(0)} \quad (29)$$

The $\hat{v}_{12}(0,2)$ estimated probability of ever visiting State 2 by time $t_{pred} = 2$, is 0.4 as, out of the $N_1(0) = 10$ simulated individuals, $M_2(2) = 4$ of them (id=1, 5, 7, 10) have ever experienced entering State 2 by $t_{pred} = 2$. Similarly, the $\hat{v}_{13}(0,2)$ estimated probability of ever visiting State 3 by time $t_{pred} = 2$, is 0.2 as, out of the $N_1(0) = 10$ simulated individuals, $M_3(2) = 2$ of them (id=7, 8) has experienced State 3 by $t_{pred} = 2$. We should note that, given being in a specific state (State 1), the probability of ever visiting a terminal state, in our case State 3, is equal to the transition probability to that state, $P_{13}(0, t_{pred}) = v_{13}(0, t_{pred})$. Also, even if it is self-evident, the probability of ever visiting the State 1, given that all individuals start from there, is always 1, $v_{11} = 1$.

The restricted expected length of stay at state j up to a specific time point of prediction t_{pred} given being in State 1 at time 0, $e_{1j}(0, t_{pred})$ can also be easily estimated, by summing the length stay of each individual in State j up to time t_{pred}, over individuals and then divide it by the total population starting at the initial state at time 0, $N_1(0)$:

$$\widehat{e_{1j}}(0, t_{pred}) = \frac{\sum_{i=1}^{N_1(0)} e_{1j,i}(0, t_{pred})}{N_1(0)} \quad (30)$$

The estimated restricted expected length of stay in State 1, $\widehat{e_{11}}$, can be derived by calculating the sum of times that the $N_1(0)$ stayed in State 1, up to time $t_{pred} = 2$, and then divide it by $N_1(0)$. The length of time that each individual spends in each state can be easily computed, with the individual length of stay in State 1, State 2 and State 3 until time $t_{pred} = 2$, namely $e_{11,i}(0,2)$, $e_{12,i}(0,2)$ and $e_{13,i}(0,2)$ given in Table 4.2. Then:

$$\widehat{e_{11}}(0,2) = \frac{\sum_{i=1}^{10} e_{11,i}(0,2)}{N_1(0)} = \frac{1.5 + 2 + 2 + 2 + 1.5 + 2 + 0.5 + 0.5 + 2 + 1}{10} = \frac{15}{10} = 1.5 \quad (31)$$

For State 2 and State 3, we can derive both the restricted expected length of stay up until time t_{pred} among all individuals (equations 32, 33) or among the individuals that have ever experienced State 2 during the 5 years of the follow-up period $M_2(5) = 6$ (id=1, 3, 5, 6, 7, 10) (equations 34, 35).

$$\widehat{e_{12}}(0,2) = \frac{\sum_{i=1}^{10} e_{12,i}(0,2)}{N_1(0)} = \frac{0.5 + 0 + 0 + 0 + 0.5 + 0 + 1.5 + 0 + 0 + 1}{10} = 0.35 \quad (32)$$

$$\widehat{e_{13}}(0,2) = \frac{\sum_{i=1}^{N_1(0)} e_{13,i}(0,2)}{N_1(0)} = \frac{0 + 0 + 0 + 0 + 0 + 0 + 0.5 + 1.5 + 0 + 0}{6} = \frac{2}{6} = 0.33 \quad (33)$$

$$Cond.\,\widehat{e_{12}}(0,2) = \frac{\sum_{i=1}^{M_2(5)} e_{12,i}(0,2)}{M_2(5)} = \frac{0.5 + 0 + 0.5 + 0 + 1 + 1}{6} = 0.5 \quad (34)$$

$$Cond.\,\widehat{e_{13}}(0,2) = \frac{\sum_{i=1}^{M_2(5)} e_{13,i}(0,2)}{M_2(5)} = \frac{0 + 0 + 0 + 0 + 0.5 + 0}{6} = 0.083 \quad (35)$$

If we are interested in estimates regarding transitions that have a starting state other than the initial state of the structure (State 1), for example the transition probability from State 2 to State 3, $\hat{P}_{23}(s, t_{pred})$, we can simulate a population (survival times) via transition intensity rate models with left truncation at time s, with every individual starting at State 2. Then, we just have to count how many individuals have transitioned from State 2 to State 3 from time s until time t_{pred}.

4.5 SIMULATION FOR EVALUATION OF STATISTICAL METHODS

Simulation techniques are used in **Study I** and **Study IV** of this thesis. In **Study I**, different modelling approaches are used to estimate the cause-specific CIF for colon cancer and other cause mortality when the hazard for other cause mortality is a function of attained age. In **Study**

IV, multi-state models are used in order to estimate probabilities of event recurrences and death given that the underlying recurrent events and death processes are based on a joint frailty model underlying mechanism and are related via a common frailty term u_i. In both cases, I try to assess the performance of the proposed approaches based on certain performance measures. I briefly describe the process followed during a simulation study and the performance measures that can be derived in order to assess the use of each approach.

4.5.1 Aim of simulation- Estimands

Firstly, it is important to set the aim of the simulation. In the current work, there are measures/quantities that need to be estimated with low bias, appropriate coverage and high precision. In **Study I**, the estimands are the cause-specific CIF for death due to colon cancer (CIF_1) and CIF for other cause mortality (CIF_2) across time t in years after the colon cancer diagnosis. In **Study IV**, the estimands of interest are the probabilities of a new recurrent event or death given one, two, or three past recurrent events within 1 year after the start of the follow-up.

4.5.2 Data generating mechanism- True values

For each scenario created, we have to select values for a series of parameters. These values are treated as the true values of the parameters for the hypothetical population of each scenario. In case of composite estimands, these true values of the parameters are used to derive the "true values" of the estimands to be studied. In the case of **Study I**, I derive the true values of the cause-specific CIFs based on the parameters set for the scenario-specific population (variance of simulated age at diagnosis variable) and the hazard functions of each competing event (parameters for baseline hazard, covariate effects). Through integration of a composite function of "true" hazard functions, I derive the "true" cause-specific CIF values across time. In **Study IV**, it is not possible to analytically derive the true values of the measures of interest, the probabilities of a new recurrence or death up to time t. Thus, for **Study IV**, I simulate a large population (7 million individuals) under different scenarios of the variance of the gamma distribution of the frailties and the alpha parameter the signifies the association between the recurrent event and the death process. The size of the population should be big enough so that the Monte-Carlo error across different simulated populations is very small. Treating the generated population under each scenario as the underlying population of interest, the probabilities of a new recurrence or death up to time t can be derived via simple frequencies over time.

4.5.3 Data simulation- Estimates

We can then simulate multiple random datasets based on the parameters of the DGM under each scenario. A seed number is used in order to be able to replicate the same set of simulated datasets no matter how many times the simulation is rerun. The number and the size of the simulated datasets differ in different simulation studies and depend on the aims of the simulation, the performance measures, and the desired level of the Monte-Carlo error. In this step, each modelling approach is applied to all simulated datasets, deriving estimates of the

target estimand for each scenario. For **Study I**, those are the CIF_1 and CIF_2 estimates across time that are derived under approaches that use different timescales when modelling the other cause mortality rate or different level of complexity in how they include the effect of age at diagnosis in the other cause mortality rate model.

4.5.4 **Performance measures**

Depending on the aim of the simulation study, different performance measures may be derived. In the case of **Study I**, bias, relative precision and coverage were the performance measures of interest while the Monte-Carlo standard error of the bias estimate was also presented as an estimate of the simulation uncertainty. After deriving the estimates $\hat{\theta}_i$ (i is the simulated sample index), $i = 1, \dots, n_{sim}$, we can combine them with the true values of our estimands θ and estimate the performance measures of interest.

Bias estimate

$$Bias = \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} (\hat{\theta}_i - \theta) \tag{36}$$

Monte-Carlo standard error of bias estimate

$$Monte - Carlo\ S.E_{bias} = \sqrt{\frac{1}{n_{sim}(n_{sim}-1)} \sum_{i=1}^{n_{sim}} (\hat{\theta}_i - \bar{\theta})^2} \tag{37}$$

Empirical Standard error of estimates

$$Empirical\ S.E = \sqrt{\frac{1}{n_{sim}-1} \sum_{i=1}^{n_{sim}} (\hat{\theta}_i - \bar{\theta})^2} \tag{38}$$

Relative increase in precision when comparing approach B with approach A estimate

$$Relative\ precision = 100 \left[\left(\frac{\widehat{EmpSE}_A}{\widehat{EmpSE}_B} \right)^2 - 1 \right] \tag{39}$$

Coverage

$$Coverage = \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} 1(\hat{\theta}_{lower,i} < \theta < \hat{\theta}_{upper,i}) \tag{40}$$

There are more performance measures that can be derived such as mean square error (MSE), power, average model standard errors and others. More details about the definitions of the

performance measures but also regarding the overall simulation procedure are given by Morris, White and Crowther (80).

4.6 TIMESCALES IN SINGLE-EVENT SURVIVAL ANALYSIS

4.6.1 Choice of timescale and truncation

The random variable of most interest in a survival analysis setting is the time-to-event. However, depending on the type of event, the suggested time of origin may vary and thus, so does the timescale. For example, suppose we study time to death due to cancer. In that case, we know that the cancer mortality rate is mainly a function of time since diagnosis, so that should be the timescale of choice when modelling that rate. Instead, if we study other outcomes, such as mortality, coronary heart disease and stroke, it feels more natural to consider attained age (time since birth) as the timescale rather than time-on-study as the hazard rate for the event tends to change more as a function of age than as a function of time-on-study (81).

This choice of timescale, depending on the event of interest and the time of origin, may require accounting for left-truncation, meaning the fact that we observe only those individuals that live long enough to be diagnosed, and thus, observed (82). Therefore, as most studies have data on the patients only after a point in time, for example, after the start of the study, we should take into account the left truncation if we use attained age as the timescale of the hazard rate model.

In this case, the hazard rates and the survival functions should take account of the left-truncation. A hazard function with left truncation on the attained age timescale a can be defined as:

$$h\left(a\middle|\boldsymbol{Z_i}, a_{0_i}\right) = \lim_{\Delta a \to 0} \frac{P(a \leq A < a + \Delta a | A \geq a, A \geq a_{0_i}, \boldsymbol{Z_i})}{\Delta a} \tag{41}$$

with a_0 the age at diagnosis variable and $\boldsymbol{Z_i}$ the vector of covariates. Then the survival function conditional on the i^{th} individual surviving at least until the age at diagnosis a_{0_i} is

$$S\left(a\middle|\boldsymbol{Z_i}, a_{0_i}\right) = \frac{S(a|\boldsymbol{Z_i})}{S\left(a_{0_i}\middle|\boldsymbol{Z_i}\right)} = \exp\left(- \int_{a_{0_i}}^{a} h\left(u\middle|\boldsymbol{Z_i}, a_{0_i}\right) du \right) \tag{42}$$

Equations 41 and 42 are also used in **Study I**, where the primary event is death due to colon cancer and attained age is used as the timescale for modelling other cause mortality (competing event) accounting for left truncation on age at diagnosis.

4.6.2 Multiple timescales

In this thesis, I only consider one timescale when applying hazard rate models. As shown in the results of **Study I** (See Section 5.1, Figures 5.1 and 5.2b), when the hazard rate is a function of attained age, modelling the hazard rate separately for the two components of attained age, time since diagnosis (as main timescale) and age at diagnosis (covariate in the model), can

introduce some degree of bias. In reality, the hazard rate for an event may be simultaneously a function of multiple timescales. In the case of colon cancer, even though colon cancer mortality among cancer patients is mainly driven by time since diagnosis, we expect it to change as a function of attained age as well. In such cases, the most common approach is to use as the main timescale the timescale across which the hazard rate presents the most variability – for example time since diagnosis- while any other timescale can be taken into account indirectly by including it as a time-fixed covariate in the model, with the additional option of including interactions between the covariate and the main timescale.

Two or more timescales can be simultaneously modelled if hazard models can be fit with the data split into short intervals of time across the relevant timescales. Then, a Cox or a Poisson generalised linear model with the time-intervals included as categories (83) or a continuous function through the intervals (e.g splines). Fitting a Cox model means that the timescale chosen as the baseline hazard timescale will not be parametrically estimated which hinders the process of modelling multiple timescales (84). In addition, whether Cox or Poisson model is used, the time-split of the data is essentially imposing the assumption of piecewise constant hazard rates within each time-interval, plus it can make the estimation computationally intensive. Instead, a FPSM can be fit modelling the log baseline hazard rate as a function of multiple timescales, with each timescale being included in the model as a continuous function of a reference timescale, with its effect modelled with spline functions (85). This way, an arbitrary number of timescales can be incorporated, as well as interactions between the timescales and time varying effects of other covariates with each timescale. A proportional hazards model with two timescales (t_1, t_2) can be expressed on the log hazard scale as:

$$\ln[h(t_1, t_2 | \boldsymbol{\gamma_p}, \boldsymbol{m_p}, \boldsymbol{\gamma_s}, \boldsymbol{m_s}, \boldsymbol{\beta}, \boldsymbol{Z_i})] = p_0(t_1 | \boldsymbol{\gamma_p}, \boldsymbol{m_p}) + s_0(t_2 | \boldsymbol{\gamma_s}, \boldsymbol{m_s}) + \boldsymbol{\beta^T Z_i} \quad (43)$$

with $\boldsymbol{m_p}$ and $\boldsymbol{\gamma_p}$ the knots vector and associated parameter vector for the spline function p_0 of the first timescale t_1 and $\boldsymbol{m_s}$ and $\boldsymbol{\gamma_s}$ the knots vector and associated parameter vector for the spline function s_0 of the second timescale t_2. By rewriting one of the timescales as a function of the other, for example $t_2 = t_1 + c,$ we can use FPSMs. For more detail, see Batyrbekova et al (85).

4.7 TIMESCALES IN COMPETING RISKS SETTINGS

In **Study I**, there is a competing risk setting with two competing events, death due to colon cancer and death due to other causes, where we consider time since diagnosis to be the natural choice of timescale for death due to colon cancer and attained age to be the natural choice for other cause mortality. Under that assumption, the colon cancer mortality rate can be expressed as $h_1^{time}\left(t \middle| \boldsymbol{Z_i}, a_{0_i}\right)$ as it is the rate for the first competing event $(k = 1)$ and is a function of time since diagnosis $(time)$, while the other cause mortality rate can be expressed as $h_2^{age}\left(a \middle| \boldsymbol{Z_i}, a_{0_i}\right)$ as it is the rate for the second competing event $(k = 2)$ and is a function of attained age (age). On the other hand, if we adopt time since diagnosis as the timescale for

both colon cancer and other cause mortality, then the mortality rate for other cause mortality can be expressed as $h_2^{time}\left(t\middle|\boldsymbol{Z_i}, a_{0_i}\right)$ as it is the rate for the second competing event ($k = 2$) but is now assumed to be a function of time since diagnosis ($time$). The survival functions corresponding to the aforementioned hazard functions are:

$$S_1^{time}\left(t\middle|\boldsymbol{Z_i}, a_{0_i}\right) = \exp\left(-\int_0^t h_1^{time}\left(u\middle|\boldsymbol{Z_i}, a_{0_i}\right) du\right), \tag{44}$$

$$S_2^{time}\left(t\middle|\boldsymbol{Z_i}, a_{0_i}\right) = \exp\left(-\int_0^t h_2^{time}\left(u\middle|\boldsymbol{Z_i}, a_{0_i}\right) du\right) \text{ and} \tag{45}$$

$$S_2^{age}\left(a\middle|\boldsymbol{Z_i}, a_{0_i}\right) = \frac{S_2^{age}(a|\boldsymbol{Z_i})}{S_2^{age}\left(a_{0_i}\middle|\boldsymbol{Z_i}\right)} = \exp\left(-\int_{a_{0_i}}^{a} h_2^{age}\left(u\middle|\boldsymbol{Z_i}, a_{0_i}\right)du\right) \tag{46}$$

The other cause mortality rate as a function of attained age can also be expressed in terms of time since diagnosis t following equations 2 and 4 of the **Study I** manuscript we have:

$$h_2^{age}\left(a\middle|\boldsymbol{Z_i}, a_{0_i}\right) = h_2^{age}\left(a_{0_i} + t\middle|\boldsymbol{Z_i}, a_{0_i}\right) \text{ and} \tag{47}$$

$$\begin{aligned} S_2^{age}\left(a\middle|\boldsymbol{Z_i}, a_{0_i}\right) &= \exp\left(-\int_{a_{0_i}}^{a} h_2^{age}\left(u\middle|\boldsymbol{Z_i}, a_{0_i}\right)du\right) \\ &= \exp\left(-\int_0^t h_2^{age}\left(a_{0_i} + w\middle|\boldsymbol{Z_i}, a_{0_i}\right)dw\right) = S_2^{age}\left(a_{0_i} + t\middle|\boldsymbol{Z_i}, a_{0_i}\right) \end{aligned} \tag{48}$$

Therefore, both the cumulative incidence function for colon cancer (CIF_1) and other cause mortality (CIF_2) can be derived by assuming either time since diagnosis or attained age as the timescale for other cause mortality.

For example, if we assume time since diagnosis t as the timescale for both colon cancer and other cause mortality, the CIF for the k^{th} competing event with $k \in \{1,2\}$ can be expressed as:

$$CIF_k\left(t\middle|\boldsymbol{Z_i}, a_{0_i}\right) = \int_0^t S_1^{time}\left(u\middle|\boldsymbol{Z_i}, a_{0_i}\right) S_2^{time}\left(u\middle|\boldsymbol{Z_i}, a_{0_i}\right)\ h_k^{time}\left(u\middle|\boldsymbol{Z_i}, a_{0_i}\right) du \tag{49}$$

If we assume attained age a as the timescale for other cause mortality, the CIF for colon cancer mortality $k = 1$ can be expressed as a function of time since diagnosis as:

$$CIF_1\left(t\middle|\boldsymbol{Z_i}, a_{0_i}\right) = \int_0^t S_1^{time}\left(u\middle|\boldsymbol{Z_i}, a_{0_i}\right)\ S_2^{age}\left(a_{0_i} + u\middle|\boldsymbol{Z_i}, a_{0_i}\right) h_1^{time}\left(u\middle|\boldsymbol{Z_i}, a_{0_i}\right) du \tag{50}$$

while the CIF for other cause mortality ($k = 2$) can be expressed as a function of time since diagnosis:

$$CIF_2\left(a_{0_i} + t\middle|\boldsymbol{Z_i}, a_{0_i}\right) = \int_0^t S_1^{time}\left(u\middle|\boldsymbol{Z_i}, a_{0_i}\right) S_2^{age}\left(a_{0_i} + u\middle|\boldsymbol{Z_i}, a_{0_i}\right) h_2^{age}\left(a_{0_i} + u\middle|\boldsymbol{Z_i}, a_{0_i}\right)du \tag{51}$$

In **Study III**, in a multi-state setting, the hazard rates (transition intensity rates) of the transitions are modelled either as functions of one, common timescale (time since start of the study, time since entering a state) or each transition can be modelled on a different timescale (See more detail in Section 4.8.3).

4.8 TIMESCALES IN INTENSITY-BASED MULTI-STATE MODELS

4.8.1 Markov assumption- Total time- Clock forward

The Markov assumption is the assumption that the future state of the multi-state process depends only on the current state, and not on past states, or times that past states were reached. Under this assumption, the intensity rates of all transitions are functions of time t, the total time or time since the start of the multi-state process. Under this assumption, the transition probabilities can be derived analytically, using the approaches presented in Section 4.3.3. Most measures that are a function of transition probabilities can be derived by calculating that function (43,86). For example, in order to derive the restricted expected length of stay measure, we need to calculate the integral of the transition probabilities (or alternatively via simulation).

4.8.2 Semi-Markov - Time spent in current state - Clock reset

While the Markov assumption may be convenient, it is not always realistic to assume the transition intensity rates as being functions of the total time t of the process. Upon entering a certain state, the transition intensity rate towards the next state may depend much more on the time spent in the current state rather than the total time of the process. For example, a transition intensity rate from a medication discontinuation period towards a new antidepressant medication cycle is more likely to be a function of time since entering the discontinuation period rather than the time since the start of the follow-up. Under the semi-Markov assumption, the transition intensity rates are a function of time t_j since entering the current state j which is equal to the total time t of the process minus the time T_j of entering the j^{th} state of the multi-state structure, therefore $t_j = t - T_j$. Under this assumption, all the measures of interested can easily be predicted via the use of simulation-based approaches (43,68,87).

4.8.3 Different timescales for different transitions - Clock mix

Depending on the setting, one may assume it is more natural for specific transition intensity rates to be functions of total time, while for other transition intensity rates to be functions of time since entering the current state. For example, a transition intensity rate towards a death state is more likely to be a function of time since the start of the follow-up (e.g time since diagnosis) rather than time since entering the current antidepressant medication cycle. Similarly, as mentioned in Section 4.8.2, a transition intensity rate from the entering a medication discontinuation period towards a new medication cycle is more likely to be a function of time since entering the discontinuation period rather than the time since the start of the follow-up. In this case, we can model different transition intensity rates as functions of different timescales. Then, as in the case of the semi-Markov process, estimates of the transition probabilities and other measures of interest can be derived via simulating disease pathways for

a population-size number of individuals and then using the frequencies of individuals across states and time. In a sensitivity analysis of **Study III**, I tried to evaluate whether measures of interest are sensitive to different modelling assumptions regarding the timescales (clock forward, clock reset, clock mix). Under the clock mix approach, the transition probabilities and restricted expected length of stay measures are estimated based on transition intensity rates that are functions of total time for transitions towards the terminal state of death and functions of time since entering the current state for transitions towards intermediate states (medication cycles, discontinuation periods).

4.8.4 **Presenting conditional predictions**

When all individuals start from the same initial state, predictions that are conditional on a state other than the starting state of the process, should be derived given a left truncation time s greater than 0. For example, in the 3-state Illness-Death example of Figure 2.1c, it would be meaningful to present the probability of transitioning to the absorbing death state (State 3) by time t_{pred} given that you are at the intermediate state (State 2) at time s, $\widehat{P_{23}}(s, t_{pred}) = \hat{P}\big(Y(t_{pred}) = 3 \big| Y(s) = 2\big)$ only if $s > 0$. A prediction for $s = 0$ would not be sensible as it is not possible to be in State 2 at the start of the follow-up.

Under a semi-Markov multi-state model with time of left truncation s equal to time r of entering the state we want to condition on, for example State 2, the probability of transitioning to State 3 by time t_{pred} is $\widehat{P_{23,r}}(s, t_{pred}) = \widehat{P_{23}}(r, t_{pred})$. As all transition intensity rates starting from State 2 are, due to the semi-Markov assumption, functions of time since entering that state, $t_{pred} - r$, the predictions are not dependent on r itself, thus, $\widehat{P_{23}}(r, t_{pred})$= $\widehat{P_{23}}(r + x, t_{pred} + x)$. For $x = -r$, we will have $\widehat{P_{23}}(r, t_{pred})$= $\widehat{P_{23}}(r - r, t_{pred} - r)$, meaning that the predicted probability $\widehat{P_{23}}$ can also be reported on time since entering State 2.

4.8.5 **Multiple timescales**

Allowing each transition intensity rate to be a function of a different timescale is a more realistic approach compared to using strictly the Markov or semi-Markov assumptions. However, one can argue that the more realistic and flexible approach would be to allow each transition intensity rate to be a function of multiple timescales. Iacobelli *et al* (46) have suggested the use of multiple timescales via a parametric Poisson model with flexible baseline transition intensity rates for two timescales based on data split across the timescales, after the intermediate state is entered in a 3-state Illness-Death structure. This approach bypasses identifiability problems in simultaneously modelling the effect of time since start of the study t, time since entering the intermediate state t_j and time of entering intermediate state T_j, because of the linear relation of $T_j = t - t_j$, because T_j and t_j are not defined prior reaching the intermediate state. Transition probabilities can be then presented from the initial and the intermediate states for different entry times in the intermediate state, having flexibly modelled both time since start of follow-up and time since entering the intermediate state in the predictions. In this thesis, I am not using multiple timescales per transition, but I refer to this modelling choice for completeness.

4.9 RECURRENT EVENTS IN THE PRESENCE OF A TERMINAL EVENT

4.9.1 Joint frailty models

In most cases of survival analysis, we study the time to a specific event of interest and estimate measures such as survival probabilities and hazard ratios. In some clinical settings, an individual may experience the event of interest more than once (e.g repeated hospitalisations) and hence the focus of the research shifts towards the study of the number and the rate of recurrences. Recurrent event processes have been studied extensively using frailty models (88–91). As each individual may have his/her own unique underlying risk or frailty for a recurrence, there is heterogeneity between individuals in the model that cannot be captured by taking into account only the measured covariates, with frailty models attempting to quantify this heterogeneity.

When interest lies in studying recurrent events in the presence of a terminal event, different modelling frameworks and approaches have been considered (52–55). In such settings, it is common to assume that different individuals have different level of susceptibility/frailty, both for the recurrent and for the terminal event process that is left unexplained conditional on the observed covariates. It is also likely that these two processes present some level of association, either positive, or negative, inducing a degree of correlation between the frailty for the recurrent process and the frailty for the terminal process. In such settings, joint frailty models are a commonly used modelling approach as they directly model the variance of the frailties and the association between the processes.

As a function of the time since the start of the study, the joint frailty model for the recurrent event and the terminal process can be defined as:

$$\begin{cases} \lambda_{ij}^{R}(t|u_i) = u_i\lambda_{ij}^{R}(t) \ = u_i\lambda_0^{R}(t)\exp(\boldsymbol{\beta_1 Z_{ij}^{R}}) \\ \lambda_i^{D}(t|u_i) = {u_i}^{\alpha}\ \lambda_i^{D}(t) = {u_i}^{\alpha}\lambda_0^{D}(t)\exp(\boldsymbol{\beta_2 Z_i^{D}}) \end{cases} \quad (52)$$

With u_i the individual frailty term that is assumed to follow gamma distribution of mean equal to 1 and a variance equal to θ, α the term that allows different type of dependence between the two processes, $\lambda_{ij}^{R}(t|u_i)$, the recurrence rate conditional on the frailty of individual i for recurrence j, and $\lambda_i^{D}(t|u_i)$ the terminal event rate conditional on the frailty of individual i. The conditional hazard rates are derived by multiplying the frailty term with a hazard rate for recurrence $\lambda_{ij}^{R}(t)$ and a hazard rate for the terminal event $\lambda_i^{D}(t)$ for an individual with an average frailty ($u_i = 1$). The frailty term links the two processes and allows us to study both processes jointly/simultaneously. The α term in the terminal event process allows the frailty of each individual to be different between the recurrent event and the terminal event process and it allows us to draw conclusions about the dependence of the two processes. In **Study IV**, a more detailed description of Liu's joint frailty model and its predictions about recurrence and terminal event probabilities is given.

4.9.2 **Multi-state models**

4.9.2.1 Correspondence with other methods in a recurrent events setting

In recurrent event settings without considering a competing terminal event, various modelling approaches have been applied, including the Andersen- Gill (AG) model (96), the Prentice-William- Peterson (PWP) model (97), the Wei-Lin-Weissfeld (WLW) model (98), frailty models (88–91), and multi-state models (1,45,99). Amorim and Cai (100), and Rodrigo Villegas et al (101) offer useful reviews of these methods. The AG, PWP and WLW semi-parametric methods can be considered as specific cases of a multi-state model with recurrent event states and specific assumptions about the hazard rates/ transition intensity rates. The AG model can be thought of as a recurrent multi-state structure that imposes common baseline hazards for all recurrent events and common, proportional hazards for the covariate effects included in the model. The effects of past recurrent events on the hazard rate for recurrence can be modelled as a time-varying covariate in the transition rate models. The PWP model can be thought of as a recurrent multi-state model allowing baseline transition intensity rates to differ across transitions, with the PWP-total time approach corresponding to a multi-state model under a Markov assumption and the PWP-gap time approach corresponding to a multi-state model under a semi-Markov assumption, using time since last recurrent event as the timescale. In a WLW model anyone who has not yet experienced the j^{th} recurrent event up to time t is at the risk set for this event, even individuals that have not experienced the $j-1^{th}$recurrent event up to time t, corresponding to a recurrent multi-state model allowing different baseline transition intensity rates but not taking into account left truncation.

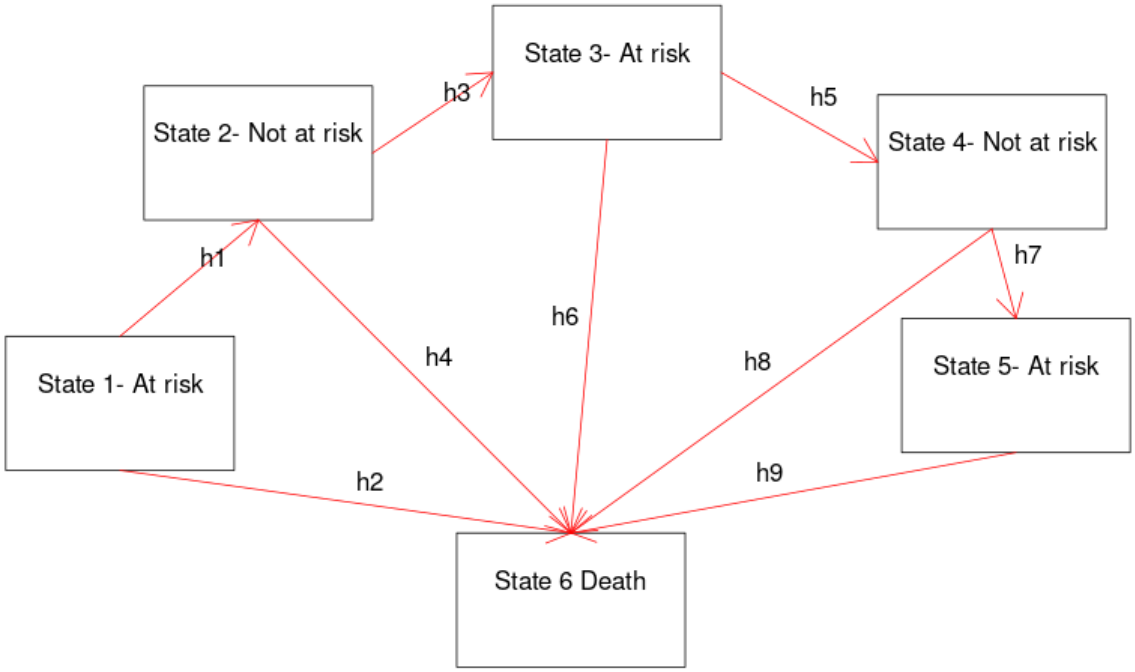
Applying a multi-state model is, therefore, a more general non-frailty approach compared to the aforementioned models. The baseline transition intensity rates can be modelled parametrically (102) (instead of semi-parametrically), with more choices of timescales (see Section 6.1), flexible modelling of the covariate effects and capability of deriving estimates of probability and probability-based measures. Multi-state models can also easily account for competing, terminal events.

4.9.2.2 Allowing for gaps at risk-time

In a setting of recurring events, there can be gaps between periods during which subjects are at risk for new events. For example, if hospitalization is the recurrent event of interest, an individual cannot be at risk for a re-hospitalization if they are still hospitalized. In a recurrent events setting both in the presence or absence of a terminal event, frailty models accommodate for these gaps at risk-time by simply measuring time since the end of the previous recurrent event (end of previous hospitalization). Multi-state models can also allow for such gaps at risk-time as each state can have duration. Figure 4.1 below depicts a recurrent event process with the presence of a terminal event, with time-gap periods (States 2 and 4) that the individual is not at risk. We can think of these periods as a hospitalization, where the individual cannot be at risk for next hospitalization. Only after discharge does the individual become at risk again for hospitalization (States 3 and 5). Multi-state models with recurrent couples of At risk/ Not

at risk periods are used in the recurrent structures of **Study III**, allowing for taking into account of the not at risk periods when estimating the measures of interest.

Figure 4.1. Recurrent event process with the presence of a terminal event, with gaps at risk-time periods under a multi-state structure.



4.9.2.3 Summing measures over recurring states

In some multi-state settings, it may be of interest to obtain an estimate on the total probability of being in a set of specific states or the total length of time, such as (i) summing all terminal event states in a competing risks model to get all-cause probability of death; (ii) summing all non-terminal events to get probability of being alive and restricted mean survival time, or (iii) summing over all recurrent events states to get the total probability of being in a recurrent event state. In **Study III**, apart from studying the probability of being separately in each medication cycle across the follow-up via the use of recurrent multi-state models, one can also study the total probability of being in a medication cycle, when there can be multiple medication cycles and discontinuation periods across the follow-up time by summing up the transition probabilities for each medication cycle. This summation can be done for time since the start of the follow-up but also for time since entering the j^{th} medication cycle onwards. In the appendix of **Study III**, following the same rationale, I also sum the expected length of stay in medication cycles to derive estimates of total expected length of stay in a set of states, in this case, medication cycles.

Let us use as example the recurrent multi-state structure of **Study III** (check Figure 5.3b or 5.5f), for which a semi-Markov model was used. Consider a stochastic process $Y(t)$ with space of states $= 1, \ldots, L$. Let State 1 be the starting state (start of follow-up) and L be the terminal state. The even numbered states, $A = \{2, 4, \ldots, L-1\}$ can be the set of states of interest (for

example medication cycle states) with a_j the j^{th} element of set A, and $j \in J_A = \{1,2,\dots,N_A\}$, N_A being the number of medication cycle states/ elements of set A. Similarly, the uneven states, $B = \{3,5,\dots,L-2\}$ can be the set of non-absorbing states that we are not interested in (for example discontinuation period states) with b_k the k^{th} element of set B. Let t be the time of prediction, s the time of left truncation, and r_j the time of entering the j^{th} medication cycle. We are interested only in estimates either since the start of the follow-up ($s = 0$) or immediately upon entering the j^{th} medication cycle ($s = r_j$).

The total probability of being in any medication cycle (set of states A) up to time t since since the start of follow-up ($s = 0$) in the initial state can be defined as:

$$P(Y(t) \in A \,|\, Y(0) = 1) = \sum_{j \in J_A} P\big(Y(t) = a_j \big| Y(0) = 1\big) \tag{53}$$

The probability of being in the j^{th} medication cycle up to time t given entering it at $s = r_j$, can be defined as:

$$P\big(Y(t) = a_j \big| Y(r_j) = a_j\big) \tag{54}$$

Let us now split the set of states of interest in two subsets based on the j^{th} medication cycle, with $A_{j^-} = \{a_{j^-}, j^- \in J^-\}$ being the subset of all medication cycles before the j^{th} one and $A_{j^+} = \{a_{j^+}, j^+ \in J^+\}$ being the subset of all medication cycles from the j^{th} one and after, with $J^- = \{1,\dots,j-1\}$ and $J^+ = \{j,\dots,N_A\}$.

Then, the total probability of being in the j^{th} medication cycle or any subsequent medication cycle across time t since start of follow-up given entering the j^{th} cycle at time $s = r_j$ is:

$$P\big(Y(t) \in A \big| Y(r_j) = a_j\big) = P\big(Y(t) \in A_{j^+} \big| Y(r_j) = a_j\big) = \sum_{j^+ \in J^+} P\big(Y(t) = a_{j^+} \big| Y(r_j) = a_j\big) \tag{55}$$

The total restricted expected length of stay in all medication cycle states (set A) until time t since the start of the follow-up (State 1), can be defined as the integral from 0 to t of the transition probability of equation 53:

$$\int_0^t \sum_{j \in J_A} P\big(Y(u) = a_j \big| Y(0) = 1\big)\, du \tag{56}$$

The total restricted expected length of stay in the j^{th} medication cycle and all subsequent medication cycles across time t given entering the j^{th} cycle at time r_j, can be defined as the integral from r_j, to t of the transition probability of equation 55:

$$\int_0^t \sum_{j^+ \in J^+} P\big(Y(u) = a_{j^+} \big| Y(r_j) = a_j\big)\, du \tag{57}$$

According to Section 4.8.4, the predictions made under a semi-Markov model given entering a state at time r can be reported either on the time since the start of follow-up or on the time since entering each state.

5 RESULTS

5.1 STUDY I

In **Study I**, I assessed how, in a competing risk setting, the choice of timescale for a competing event can influence the estimates of the cause-specific CIFs, for a range of different scenarios (shape of baseline other cause mortality rate, standard deviation of age at diagnosis, sample size and non-proportional hazards). Specifically, for the competing events of death due to colon cancer and death due to other causes, I wanted to evaluate the performance while estimating the cause-specific CIFs for both events, if the other cause mortality rate is a function of attained age but time since diagnosis is used as a timescale instead, while also modelling the effect age at diagnosis in the model with different levels of complexity. I also presented standardized cause-specific CIFs which can be a useful tool when interest lies in assessing the overall effect of a covariate of interest on the cause-specific CIFs, as they allow comparability of different groups as well as addressing causal questions (103).

The choice of timescale for one competing event is likely to have less of an impact on the CIF of another event compared to the CIF of the event itself. This can be easily understood if we consider the components of the CIF function in a competing risk setting with two competing events, which have three components, the survival functions from the first and the second competing event ($k = 1,2$) and the hazard rate for the event under study ($k = 1$ or 2). In **Study I**, under all modelling approaches, colon cancer mortality (first competing event, $k = 1$) is assumed to be a function of time since diagnosis.

As discussed in Section 4.7 regarding the definition of the CIF for colon cancer mortality (CIF_1), under the assumption that other cause mortality is a function of time since diagnosis (equation 49 for $k = 1$) versus the assumption that it is a function of attained age (equation 50), we can observe that these alternative modelling approaches have only one out of the three functions modelled differently for CIF_1 (colon cancer mortality). The other cause mortality rate is modelled as a function of time since diagnosis $h_2^{time}(t|\boldsymbol{Z_i}, a_{0_i})$ instead of $h_2^{age}(a_{0_i} + t|\boldsymbol{Z_i}, a_{0_i})$, resulting in using the term $S_2^{time}(t|\boldsymbol{Z_i}, a_{0_i})$ for the cause-specific survival function for event $k = 2$ instead of $S_2^{age}(a_{0_i} + t|\boldsymbol{Z_i}, a_{0_i})$.

When modelling other cause mortality rate, function $h_2^{age}(a_{0_i} + t|\boldsymbol{Z_i}, a_{0_i})$ is not equivalent with $h_2^{time}(t|\boldsymbol{Z_i}, a_{0_i})$ so CIF_1 from equation 49 will not be equal to the CIF_1 from equation 50, but we can expect those two modelling approaches to yield similar estimations of the CIFs for death due to colon cancer if the effect of age at diagnosis a_0 as a covariate is flexibly modelled.

Similarly, regarding the definition of the CIF for other cause mortality (CIF_2), under the assumption that other cause mortality is a function of time since diagnosis (equation 49 for $k = 2$) versus the assumption that it is a function of attained age (equation 51), we can observe that these alternative modelling approaches have two out of the three components of the CIF are

modelled differently for CIF_2 (other cause mortality), that is $h_2^{time}(t|\boldsymbol{Z_i}, a_{0_i})$ and $S_2^{time}(t|\boldsymbol{Z_{i_i}}, a_{0_i}) = \exp\left(-\int_0^t h_2^{time}(w|\boldsymbol{Z_i}, a_{0_i})\, dw\right)$ instead of $h_2^{age}\left(a_{0_i} + t\middle|\boldsymbol{Z_i}, a_{0_i}\right)$ and $S_2^{age}\left(a_{0_i} + t\middle|\boldsymbol{Z_i}, a_{0_i}\right) = \exp\left(-\int_0^t h_2^{age}\left(a_{0_i} + w\middle|\boldsymbol{Z_i}, a_{0_i}\right) dw\right)$.

In this case, depending on the choice of the timescale, two out of the three components of the cumulative incidence functions of equations 49 and 51 will differ, so any difference between the terms $h_2^{time}(t|\boldsymbol{Z_i}, a_{0_i})$ and $h_2^{age}\left(a_{0_i} + t\middle|\boldsymbol{Z_i}, a_{0_i}\right)$ will have a bigger impact in the estimation of CIF of other cause mortality.

Factors such as the shape of baseline other cause mortality rate, standard deviation of age at diagnosis, sample size and non-proportional hazards can influence the structure of the risk set and thus differentially influence the mortality rates estimation on the attained age and the time since diagnosis timescale, causing $h_2^{time}(t|\boldsymbol{Z_i}, a_{0_i})$ and $h_2^{age}\left(a_{0_i} + t\middle|\boldsymbol{Z_i}, a_{0_i}\right)$ to diverge. Still, if the other cause mortality rate is modelled with time since diagnosis timescale and the effect of age at diagnosis is modelled flexibly then these terms should be quite similar. In that case, even if the other cause mortality rate is a function of attained age and we model it as a function of time since diagnosis, the impact on the CIFs may be quite small.

Based on scenarios with varying values of the aforementioned factors and modelling approaches using time since diagnosis as the timescale for other cause mortality and including age at diagnosis as a covariate with an effect of increasing complexity, I explored the bias and other performance measures in the estimation of the CIFs for colon cancer and other cause mortality. For standard deviation of age, values of 10 and 15 years were explored. For sample size, values of 500 and 2000 individuals were explored. For hazard proportionality, a proportional hazards assumption and a non-proportional hazards assumption of a covariate of interest (gender) was explored and for baseline other cause mortality rate three different shape mortality rates as a function of age were considered (details of DGM in manuscript of **Study I**). For gender, the non-proportional hazard assumption on the attained age timescale assumed a protective effect of gender against other cause mortality that diminishes as attained age progresses.

There were four modelling approaches (Approach a- Attained age, Approach b- Linear, Approach c- Splines and Approach d- Splines/Int), all of which modelled the colon cancer mortality rate with the same FPSM, that is, using time since diagnosis as the timescale, with 5 df for the baseline hazard, with age at diagnosis included in the model using restricted cubic splines with 5 knots (4 df) and proportional hazards assumed for gender (same as the underlying DGM). The approaches differ in the way they model other cause mortality (Table 5.1).

Table 5.1. Description of the four different modelling approaches used in **Study I** while modelling other cause mortality, with Approach a- Attained age being the approach of reference.

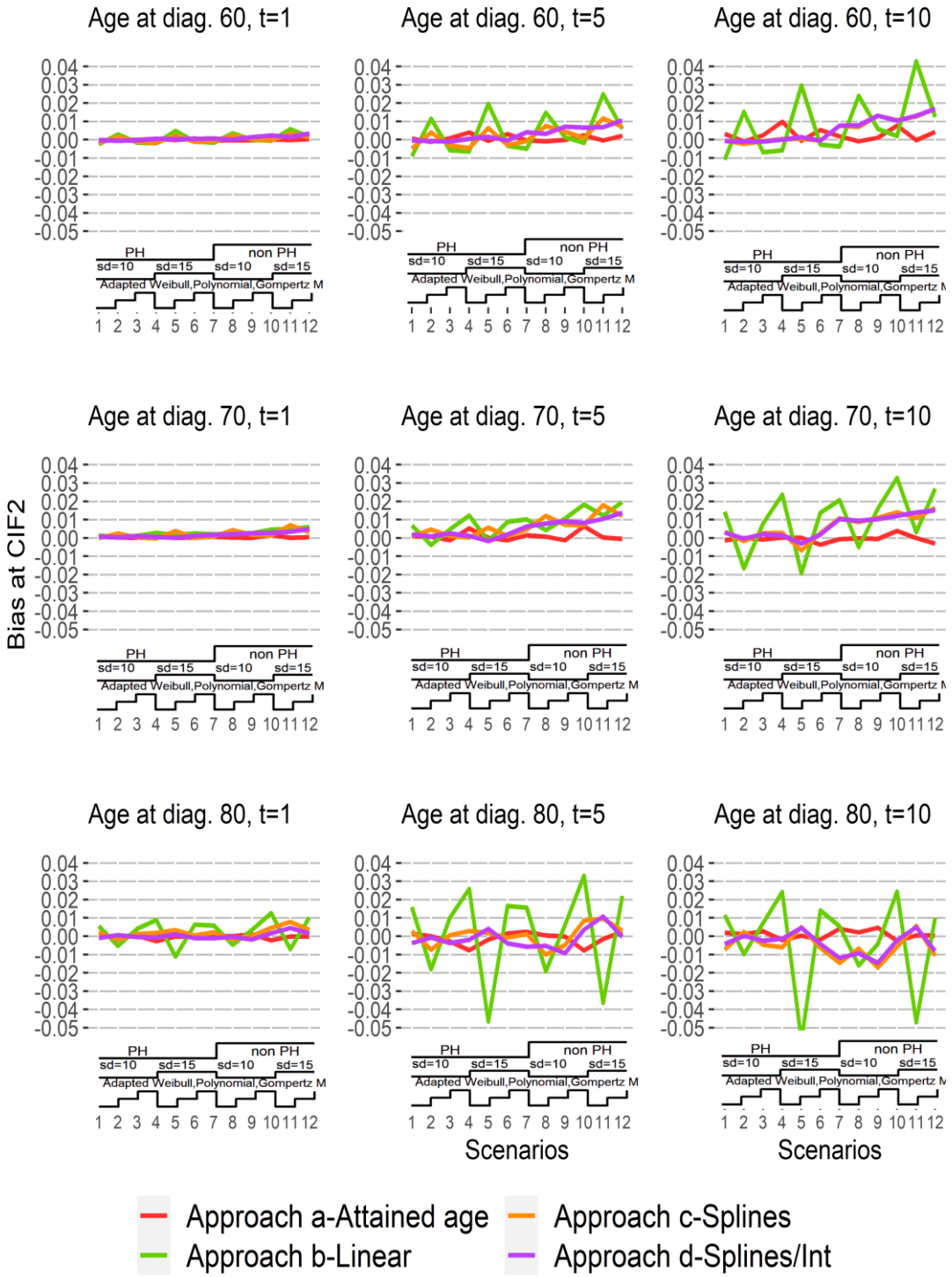
<table>
<tr><th></th><th>Timescale for other cause mortality</th><th>Age at diagnosis</th><th>Gender</th></tr>
<tr><td>Approach a- Attained age</td><td>Attained age</td><td>Not included as a covariate</td><td rowspan="4">Main effect plus restricted cubic splines with 3 df for the time dependent effect on the timescale of use.</td></tr>
<tr><td>Approach b- Linear</td><td rowspan="3">Time since diagnosis</td><td>Linear effect of age</td></tr>
<tr><td>Approach c- Splines</td><td>Age modelled using restricted cubic splines with 4 df</td></tr>
<tr><td>Approach d- Splines/Int</td><td>Age modelled using restricted cubic splines with 4 df + interaction with timescale via restricted cubic spline function of 3 df</td></tr>
</table>

Regarding the cause-specific CIF for colon cancer, the bias of all three approaches that use time since diagnosis as the timescale for both events that we can refer to as common timescale approaches (Approach b- Linear, Approach c- Splines, Approach d- Splines/Int), with different complexity of the effects of age at diagnosis in the other cause mortality model was low (<0.0035). The coverage was close to the nominal 95% and the precision level was very close with the Approach a-Attained age (relative precision close to 0), under all the different scenarios.

Regarding the bias in the cause-specific CIF for other causes (Figure 5.1), Approach b- Linear, which models age at diagnosis in the other cause mortality rate with a simple linear effect is highly sensitive to the shape of the baseline other cause mortality for most ages at diagnosis and times since diagnosis for most scenarios, presenting large bias. This modelling approach presents a large overall degree of bias, with very low coverages under some scenarios (Adapted Weibull, standard deviation of age at diagnosis equal to 15, Non-proportional hazards of gender for age at diagnosis 70) and higher precision compared to Approach a-Attained age (reference approach). Regarding the approaches that use time since diagnosis as the timescale for other cause mortality rate and which include age at diagnosis in the model with sufficient complexity (Approach c-Splines, Approach d- Splines/Int), scenarios under non-proportional hazards of a covariate in the model for other cause mortality (here gender) on the attained age scale, tends to lead to an increase in bias, showing that the time-varying effects of a covariate on the other cause mortality rate that is a function of attained age (as assumed in the DGM) cannot be fully captured by cause-specific hazard models that assume the hazard rate to be a function of time since diagnosis. This can lead to bias greater than 0.01 in the cause-specific CIF, especially for $t = 5, 10$. Variance in age at diagnosis and shape of the baseline other cause mortality rate may

influence the bias for high ages at diagnosis (e.g 80) via influencing the risk sets population, which tend to be small for old ages. Sample size (500 versus 2000) does not seem to influence the degree of bias of any of the approaches.

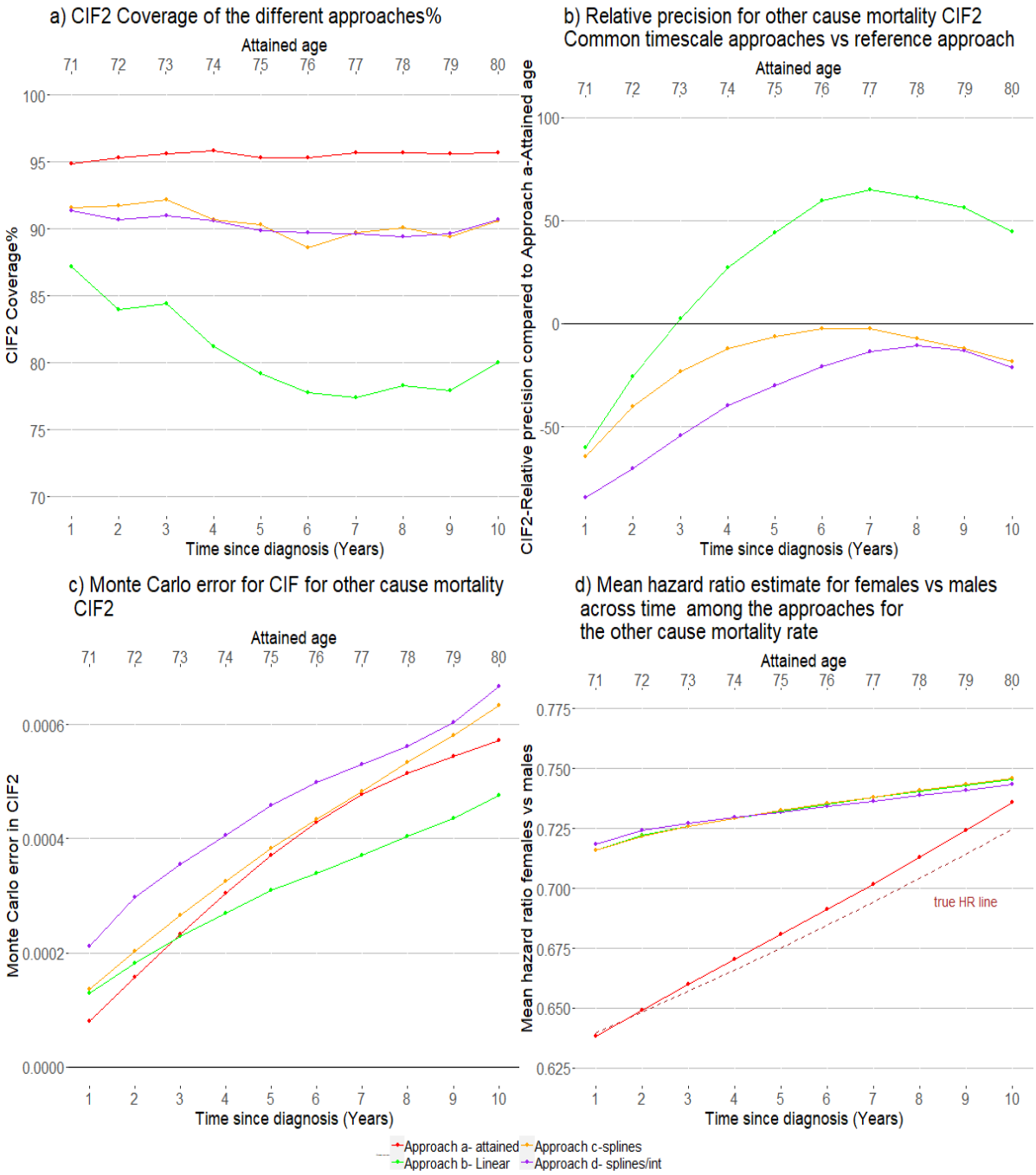
Figure 5.1. Nested loop line plot of bias in $CIF_2(t)$ from each approach over the scenarios. Note: The bias of the different approaches is given for ages at diagnosis (60, 70, 80) and times since diagnosis (1, 5, 10).



Source: Article published based on **Study I** (104)

There is also a trade-off between the modelling complexity of the effect of age at diagnosis in the model and the model precision, with the common timescale approaches showing lower precision compared to the reference Approach a- Attained age, with a maximum relative precision of -27% for Approach c- Splines and -53% for approach d- Splines/Int, at scenario 1 and age at diagnosis 70 years old.

Figure 5.2. a) CIF_2 coverage over the different approaches, b) Relative precision of CIF_2 estimation (Approaches b- Linear, c- Splines and d- Splines/Int versus reference Approach a- Attained age) c) Monte-Carlo standard error during the CIF_2 estimation over the approaches and d) mean estimated hazard ratio over the approaches. All these performance measures and estimations are given over time since diagnosis for a female diagnosed at $a_0 = 70$ years of age



In Figure 5.2, I present the coverage, Monte-Carlo standard error, relative precision and the estimated hazard ratio of females versus males across time since diagnosis under scenario 9 (Standard deviation of age at diagnosis=10, Gompertz-Makeham other cause baseline mortality rate, Non-proportional hazards of the female gender for other cause mortality rate on the attained age timescale, and sample size of 2000) for individuals diagnosed at 70 years of age. In Figure 5.2a and 5.2b we can observe that for Approach b- Linear, relative precision (compared to the reference approach) in estimating CIF_2 is positive but %coverage is quite low. This is expected, as Approach b- Linear is biased under most scenarios, so will lead to low coverage of the true value of CIF_2. In Figure 5.2b we can observe the trade-off of the common timescale approaches that model age at diagnosis flexibly using splines but at the same time this greater complexity leads to negative relative precision in estimating CIF_2, especially in the first years after diagnosis. We can also observe in Figure 5.2d, that, no matter the complexity that the common timescale approaches include age at diagnosis in the other cause mortality hazard model, they produce biased estimates of the hazard ratio of females versus males (projected on the attained age scale), when that variable has a time varying effect on the other cause mortality rate (non-proportional hazard) on the attained age timescale. This bias in the hazard ratio estimates for other cause mortality rate is reflected in the bias of CIF_2 for females in scenarios of non-proportionality (Scenarios 7-12).

In **Study** I, I also used standardized CIFs to evaluate the CIFs over a common covariate distribution and compare groups keeping the rest of the covariate distribution common. No matter the complexity when modelling the effect of age at diagnosis for other cause mortality rate, the estimates derived from the implementation of the technique on the Swedish Cancer Registry colon cancer data, are almost identical between the common timescale approaches. The estimates of the different timescales approach and the common timescale approaches are similar but not identical.

In summary, in **Study I**, I explored how the choice of timescale when modelling the other cause mortality rate (choosing time since diagnosis as the timescale instead of attained age when the underlying other cause mortality is a function of attained age) can influence the estimation of the CIFs, exploring different levels of various factors that may effect the estimations as well as different levels of complexity when modelling the effects of age at diagnosis. Given that the other cause mortality rate is a function of attained age, modelling it as a function of time since diagnosis results in negligible bias in CIF for death due to colon cancer and small bias in CIF for other cause mortality when the effect of age at diagnosis is modelled with sufficient complexity. However, if a covariate has time-varying effects on the attained age scale, those effects are not fully captured when the other cause mortality is modelled as a function of time since diagnosis, no matter the modelling complexity of the effect of age at diagnosis, resulting in small but not negligible bias in the CIF for other cause mortality.

5.2 STUDY II

In survival settings, the disease pathway of interest may consist of more than two states, such as a competing risks setting with multiple competing events or multi-state setting with multiple initial, intermediate and absorbing states. In addition, most of the measures of interest are functions of time, changing over the evolution of the process, such as the transition intensity rates under non time-homogeneous models (models not assuming constant intensity rates over time), the transition probabilities, the restricted expected length of stay in a state or a set of states, the probability of ever visiting a state and more. These factors render the evaluation of the effect of different covariate patterns on the disease process challenging. For example, while the effect of a covariate on each separate transition intensity rate is well defined and can be estimated, the overall effect of that covariate on the whole process is not, due to having to account for competing states at each step of the process. Thus, I argue that graphical displays can lead to better understanding and communication of the overall process evolution over time for different measures of interest. Different types of graphs can also be of great help in getting an intuition of the overall effect of a covariate pattern on the measure of interest (e.g length of stay in an illness-free state), as well as the comparison of different covariate patterns (e.g different in length of stay in an illness-free state between individuals of two different profiles). The aforementioned attributes of a multi-state setting plus the need to be able to communicate multi-state structures and analysis results in an easy and meaningful way to wider research audiences, provided the motivation for the development, in **Study II,** of an interactive application in RShiny, called MSMplus, that is able to read in results from multi-state model structures and analyses and portray them in a plethora of novel interactive plots, across time and covariate patterns.

Measures supported by MSMplus

The measures currently supported by MSMplus are:

- Transition probabilities/ State occupancy probabilities
- Transition intensity rates
- Total restricted expected length of stay in each state
- Probability of ever visiting a state
- Differences and ratios for the aforementioned measures among different covariate patterns

Under homogeneous (or piecewise homogeneous) Markov processes, extra measures are supported:

- Expected single period of occupancy
- Probability that each state is next
- Expected first passage time from a given state
- Expected number of visits to a state

Feeding the results in MSMplus

In order to portray the structure, descriptives and statistical analysis results of the application of a multi-state model, MSMplus requires two files as input. The first file contains information about the multi-state structure (number of states, number of transitions and transition matrix) as well as optional descriptive statistics (frequency of individuals in each state across time). This information is used by the application to build the multi-state structure. Instead of a file, the user can specify the multi-state structure directly on MSMplus platform. The second file contains the estimation results of the multi-state model analysis for the different measures over time and over the different covariate patterns. The reason I developed MSMplus to read in estimation results and not raw data to be analyzed internally, is that, due to potential ethical reasons and restrictions of data usage, the raw research data cannot or should not be uploaded online.

The input files can be created manually or automatically. In Stata and R the input files can be created automatically via Stata command options and R packages developed for this purpose. In Stata, if `msboxes` command is used and option `interactive` is specified, the first input file for MSMplus will be created. Then, if command `predictms` is used and, once again, the option `interactive` is specified, the second input file will be created. In R, I have created the MSMplus package, which contains a function called `msboxes_R` that creates the first input file for MSMplus and three wrapper functions, `msmjson`, `mstatejson` and `flexsurvjson`, that call internally the `msm`, `mstate` and `flexsurv` packages, perform the analyses, restructure the analyses results and create the second input file. In case that the MSMplus user performs a multi-state analysis in another programming language (e.g. SAS, Python) or in R but not using the `msm`, `mstate` or `flexsurv` libraries, a manually created csv file with the estimation results can be provided to the application, under certain naming and structure rules, specified both on the platform of MSMplus and the Appendix of the relevant publication in BMC Research Methodology.

How to access MSMplus

MSMplus was originally built to be an online tool so it is directly accessible at https://nskbiostatistics.shinyapps.io/MSMplus. However, I also created a version that can be locally launched via the MSMplus package I developed in R. Below is the code needed to locally launch MSMplus:

```
library("devtools")
remotes::install_github("nskourlis/MSMplus", build_vignettes = TRUE, dependencies= TRUE, force = TRUE)
library(MSMplus)
MSMplus::runMSMplus()
```

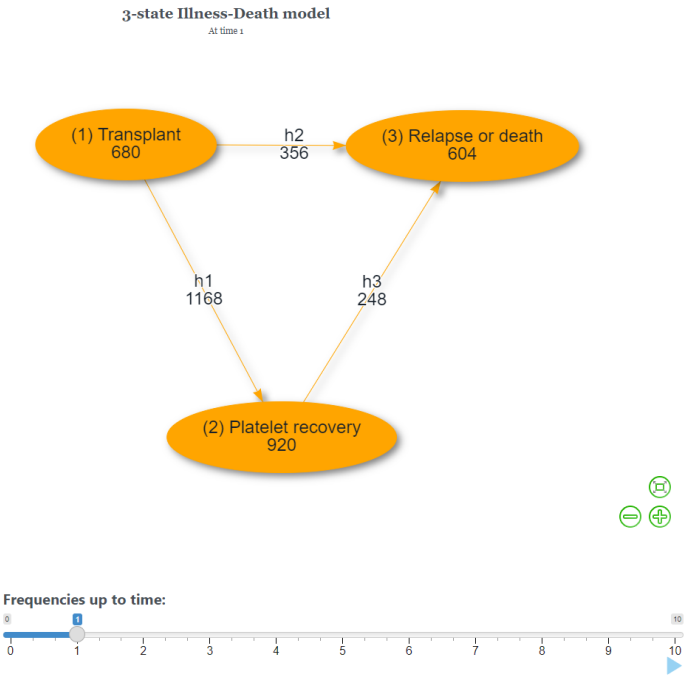
Creating multi-state graphs

By specifying the number of states and the transition matrix, either directly on the platform of MSMplus or via creating an input file, the user can create graphs of multi-state structures. These structures can vary from simple ones Figure 5.3a (3 state Illness-death model) to complex ones such as Figure 5.3b which portrays a multi-state structure with recurrent couples of medication

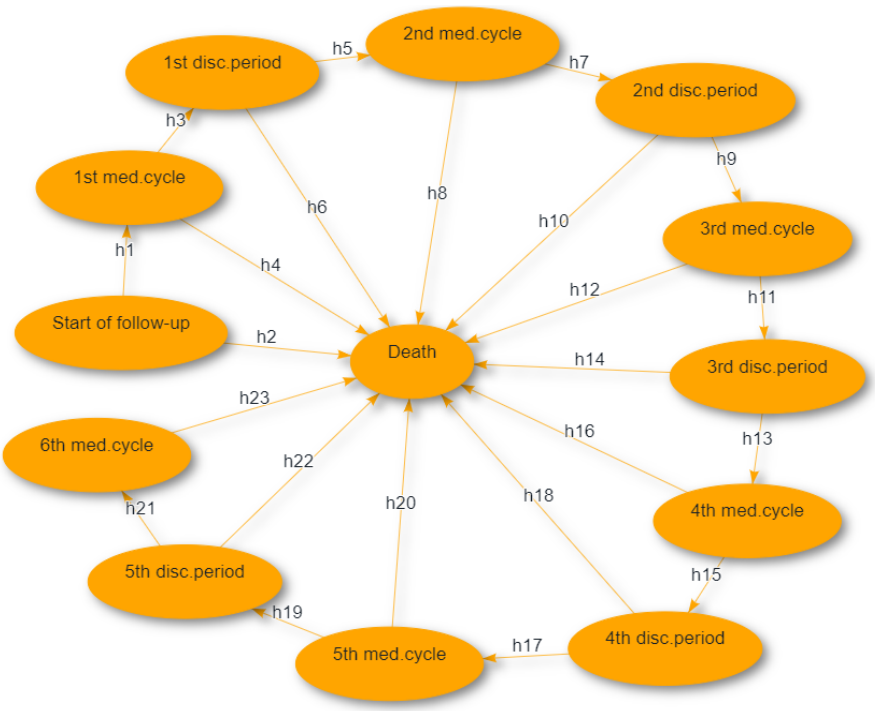
cycle- discontinuation period states with death as absorbing state (this structure is discussed in more detail in **Study III**). Via a slidebar, the number of individuals found in each state, and the number of individuals that have experienced each transition can get depicted across time.

Figure 5.3. a) Multi-state structure for the 3-state Illness-Death model based on the EBMT toy dataset with frequencies of people being in each state and experienced each transition by the first year since the start of follow-up. b) Multi-state structure with recurrent couples of medication cycle/discontinuation states used in **Study III**.

a)



b)



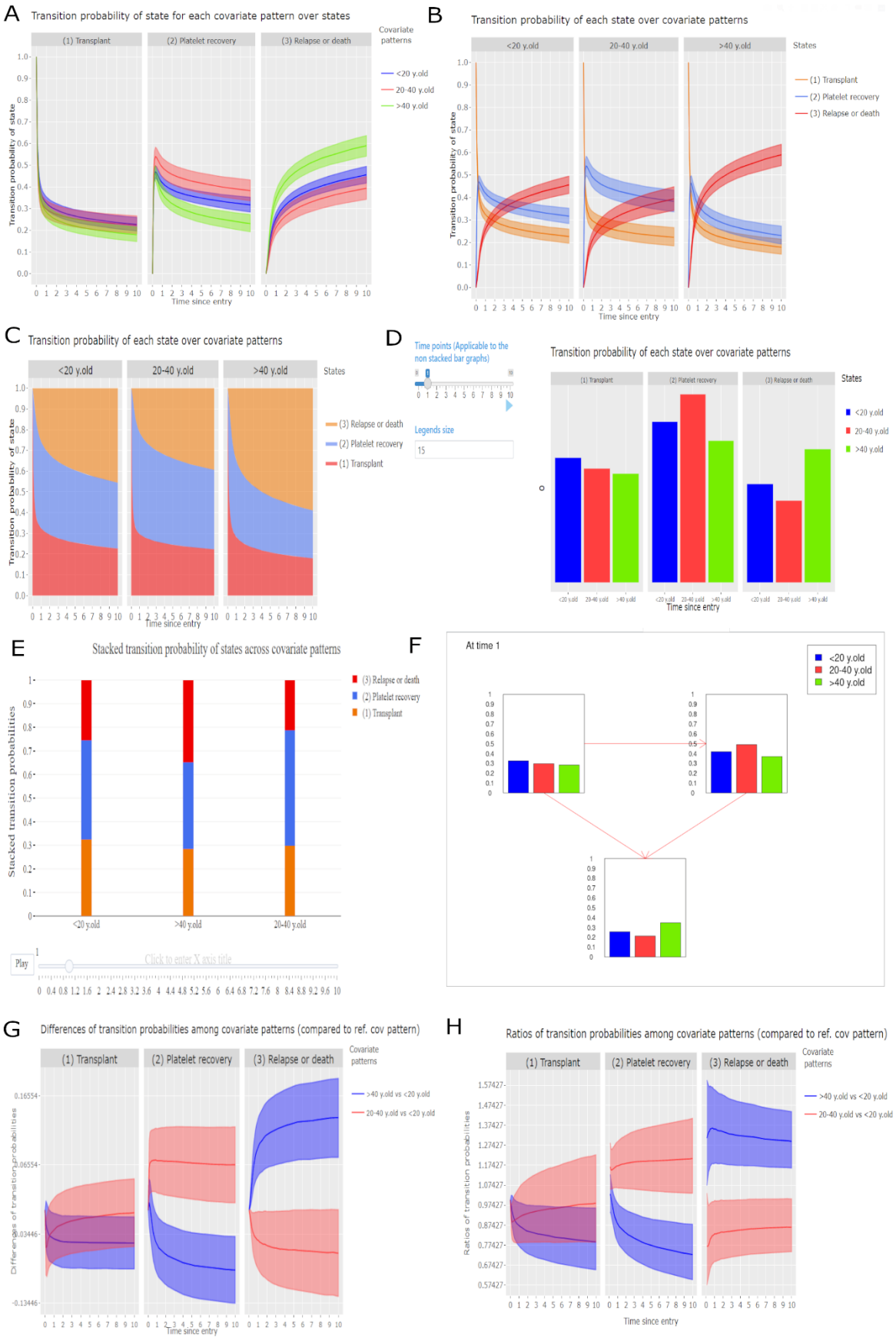
Communicating results

Via a plethora of interactive graphs, the estimation results can be communicated in alternative ways that can serve a high-quality communication of how the multi-state process evolves over time for the different estimated measures across the covariate patterns of interest. Moreover, it is unlikely that one measure can provide an overall summary of the process. Greater understanding can be obtained through visualization of different measures with the interactivity element of the application leading to a better understanding of how multiple measures change across covariate patterns and over time. I show here few of the graphs generated when providing MSMplus estimation results based on analyzing the EBMT data (toy dataset by the application), using an Illness-Death multi-state model (Figure 5.4) regarding Transition/State occupancy probabilities and other measures.

Figures 5.4A to 5.4F depict the same information, that is the estimated transition probabilities across states for different covariate patterns over time, with alternative ways such as line plots, stacked line plots, bar plots and stacked bar plots with slide bars for exploring the process evolution over dimension of time. The last two subfigures (5.4G and 5.4H), depict the difference and the ratio of the state occupancy probabilities between each covariate pattern for which predictions were derived and a reference covariate pattern that is set as reference. In this case, the three covariate patterns are "<20 years old", "20-40 years old", ">40 years old" with "<20 years old" selected as the covariate pattern of reference for measures comparison.

In summary, MSMplus is an interactive tool built to communicate results of multi-state model analyses in a flexible way, enhancing the understanding of the evolution of the multi-state process. It includes graphs and plots that change over time across different covariate patterns, for different measures. The creation of the input files is also flexible as it allows alternative ways of their creation, both automatic (for certain statistical software and commands/ packages) and manual. The primary aim of the application is to facilitate the communication of relevant research findings to both scientific and general audiences. I argue that more focus should be given by the research community to develop such applications in other fields of statistics, contributing to a more efficient way of communicating results of statistical analyses and evolution of composite processes.

Figure 5.4. A-F) Different displays of transition probabilities for each covariate pattern over states across time, G) Difference and H) Ratios of transition probabilities between covariates over states across time.



5.3 STUDY III

In **Study III**, I developed a series of multi-state models of increasing complexity in order to explore and address a series of research questions regarding the probability of antidepressant medication use among women diagnosed with breast cancer and age-matched cancer-free women from the Swedish population based on Breast Cancer Data Base Sweden 2.0 (BCBaSe 2.0), a register-based research resource (51). I started from simpler research questions such as "What is the probability of medication use initiation?". For such simple research questions, a single event survival analysis or a competing risks analysis suffices. However, more composite research questions such as "What is the total probability of being in a medication cycle since the start of follow-up or upon entering a given medication cycle?" or "What is probability of being in the current medication given entering the 1^{st} , 2^{nd} , 3^{rd} medication cycle or the 1^{st} , 2^{nd}, 3^{rd} discontinuation period?", require more complex multi-state structures to be properly addressed such as bidirectional and recurrent multi-state models. Using appropriate multi-state structures of sufficient complexity allows us to use the full richness of the prescription data of the Swedish Prescribed Drug Register in order to address more realistic and composite research questions of clinical interest.

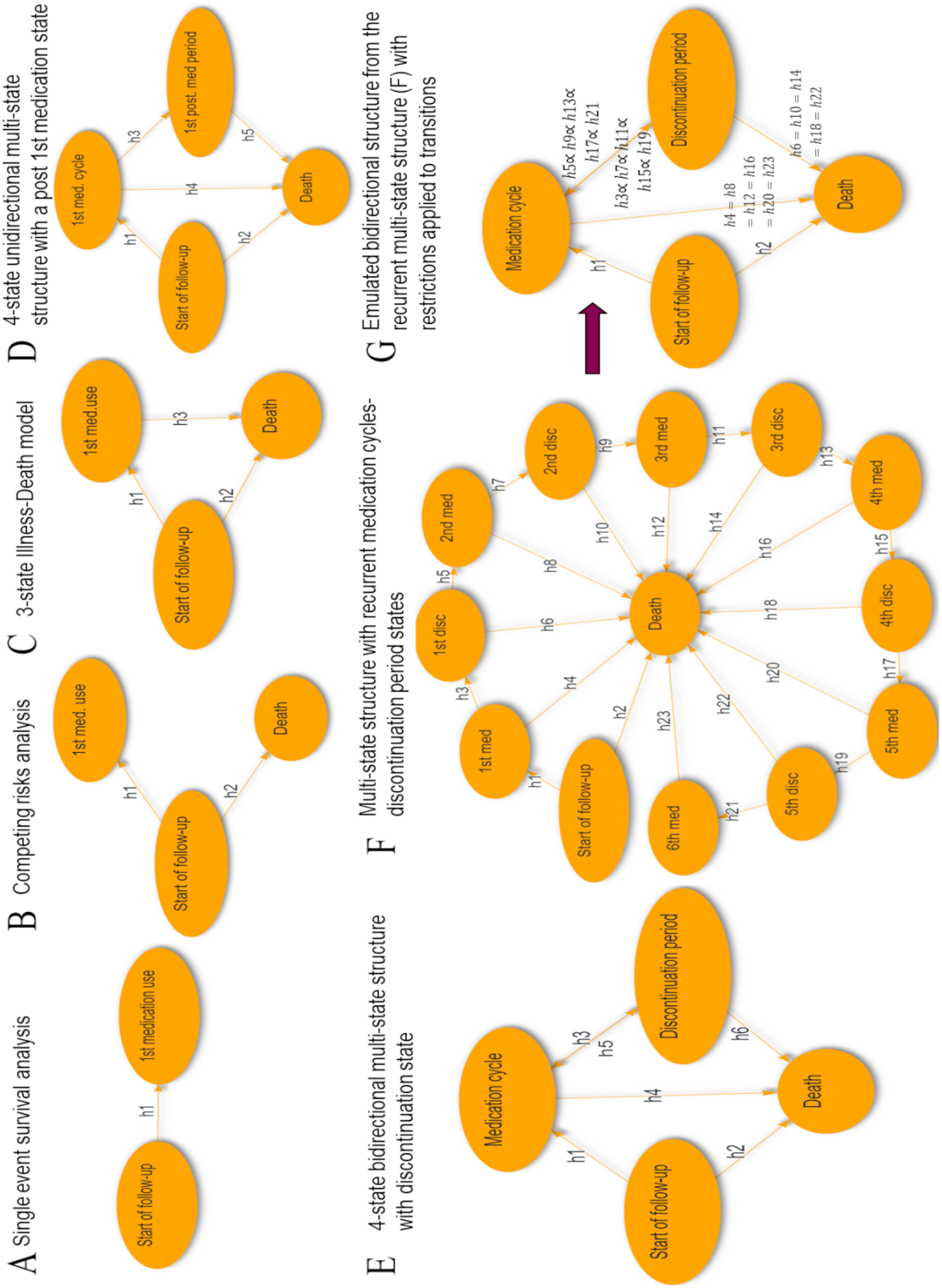
Figure 5.5 below shows the different multi-state structures used in this study, from the simplest to the most complex one. Table 5.2 shows the correspondence between the research questions, the multi-state structures, the model assumptions and the amount of information found in the data that is used in each model.

Table 5.2. Correspondence between each multi-state structure used in Study III, the research question addressed in terms of probabilities of antidepressants use corresponding to the structure and the information used.

Multi-state structure	Research questions answered in terms of probabilities	Information used
Single-event survival analysis of time to antidepressant medication initiation (Fig.5.5A)	What is the probability of ever been prescribed medication in the hypothetical situation that the individual cannot die due to any causes?	Information until first prescription date with censoring due to death, emigration or end of follow-up period
Competing risks for time to medication initiation with death as a competing event (Fig. 5.5B)	What is the probability of ever been prescribed medication up to time t after the start of the follow-up, accounting for the fact that individuals may die?	Information until first prescription date or death, censoring due to emigration or end of follow-up period

3-state Illness-Death model adding a transition from medication initiation to death (Fig. 5.5C)	What is the probability of ever been prescribed medication and still be alive up to time t after the start of the follow-up?	Information until first prescription date and death. Information on subsequent prescription dates not used. Censoring due to emigration or end of follow-up period.
4-state unidirectional multi-state model with a medication discontinuation state (Fig. 5.5D)	What is the probability of being in the **1st** medication cycle since start of follow up/ since entering the **1st** medication cycle?	Information on medication use via prescription dates and defined daily dose (DDD) until the end of the first medication cycle and then only information about death status. Information on prescription dates about subsequent medication cycles not used.
4-state Bidirectional multi-state structure with medication discontinuation state (Fig. 5.5E)	What is the probability of being in **a** medication cycle (or in **a** medication discontinuation period) since the start of follow-up or given entering one?	Use of the entirety of information on medication use (prescription dates and DDD) by an individual until death or censoring due to migration, end of follow-up.
Recurrent events multi-state structure (with or w/o restrictions) (Fig. 5.5F and 5.5G)	• What is the total probability of being in **a** medication cycle since the start of follow-up or given entering the 1st, 2nd, 3rd one? • What is probability of being in **the current** medication given entering the 1st, 2nd, 3rd medication cycle or the 1st, 2nd, 3rd discontinuation period?	Information on medication use (prescription dates and DDD) until the start of the 6th medication cycle and then only information about death status. Information on prescription dates about subsequent medication cycles not used.

Figure 5.5. Multi-state structures used in **Study III** overview.



Interpretations and limitations of the different multi-state structures used in the study

The first three multi-state structures used address research questions regarding the 1st medication use or medication initiation among the individuals of the study sample. Their interpretation has to do with the probability of ever receiving antidepressant medication after the start of the follow-up. This is the kind of research question that can be addressed under these simple multi-state structures, posing a limit to how complex research questions can be asked.

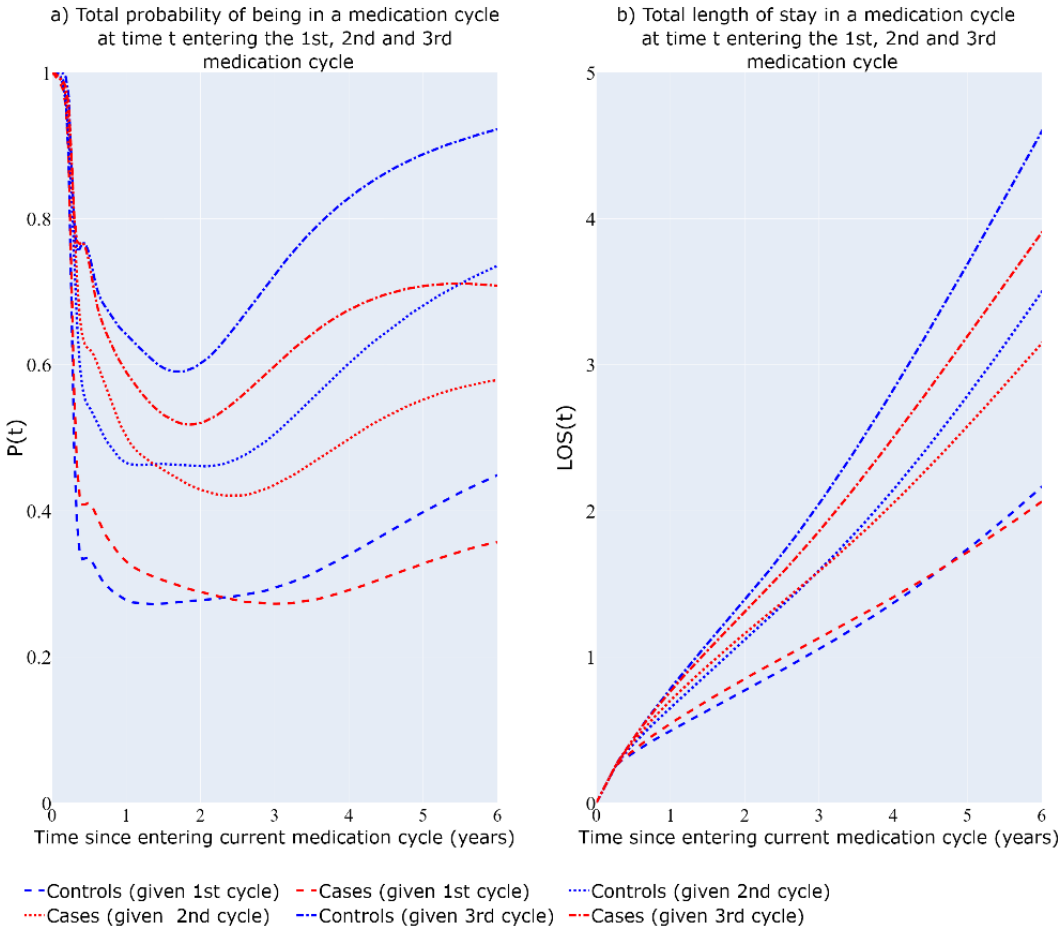
By building up the multi-state structure complexity by adding a post-medication period state, the previous state transforms from 1st medication use to 1st medication cycle, having a duration in time (beginning and end), it is not an instantaneous event, thus allowing the study of the probability of being under the first medication cycle (or other measures). The 4-state bidirectional structure, uses the entirety of the information from each individual about going back and forth between a medication cycle and discontinuation period, allowing for an infinite amount of such transitions. However, when deriving transition probabilities or other probability-based measures such as restricted expected length of stay, time-varying covariates cannot be incorporated to the transition intensity rate models, thus imposing same transition intensity rates from a discontinuation period to a medication cycle (and vice-versa), no matter the number of past medication cycles.

I tackled this issue by using a multi-state structure with recurrent pairs of medication cycles-discontinuation period states, allowing the transition intensity rates to be estimated separately for each new medication cycle and discontinuation period by fitting separate model to each transition. However, this structure can allow only for a finite number of such transitions due to issues of sparse data in high order transitions, while individuals have to be pooled under a semi-absorbing state (6th medication cycle) from which onwards they are considered chronic antidepressant users and can only move towards the state of death. I tried to tackle the data sparsity issue in high order transitions, which can also cause lower precision and convergence issues, by imposing certain restrictions among the transition intensity rates of the structure (Figure 5.5G). However, even in the case of this restricted model, due to extensive memory usage, there are limitations as to how flexibly the baseline transition intensity rates or the covariate effects for each transition can be modelled. It should be noted that the profound limitation of all the structures with medication cycle states are the assumptions used when defining what consists a medication cycle.

An important advantage when using the more complex multi-state structures is that one can derive estimates about the total probability (or total length of stay) of being in a medication cycle across the follow-up or upon entering a medication cycle. In the case of the recurrent multi-state structure, we can do it by summing up the probabilities across the different medication states given a common conditional starting state (Section 4.9.2.3). Figure 5.6a depicts the total probability of an individual, being in a medication cycle (both the current one and all the subsequent ones) as a function of time since entering the 1st , 2nd and 3rd medication cycle, with blue lines for the population comparison group of women and red lines for the

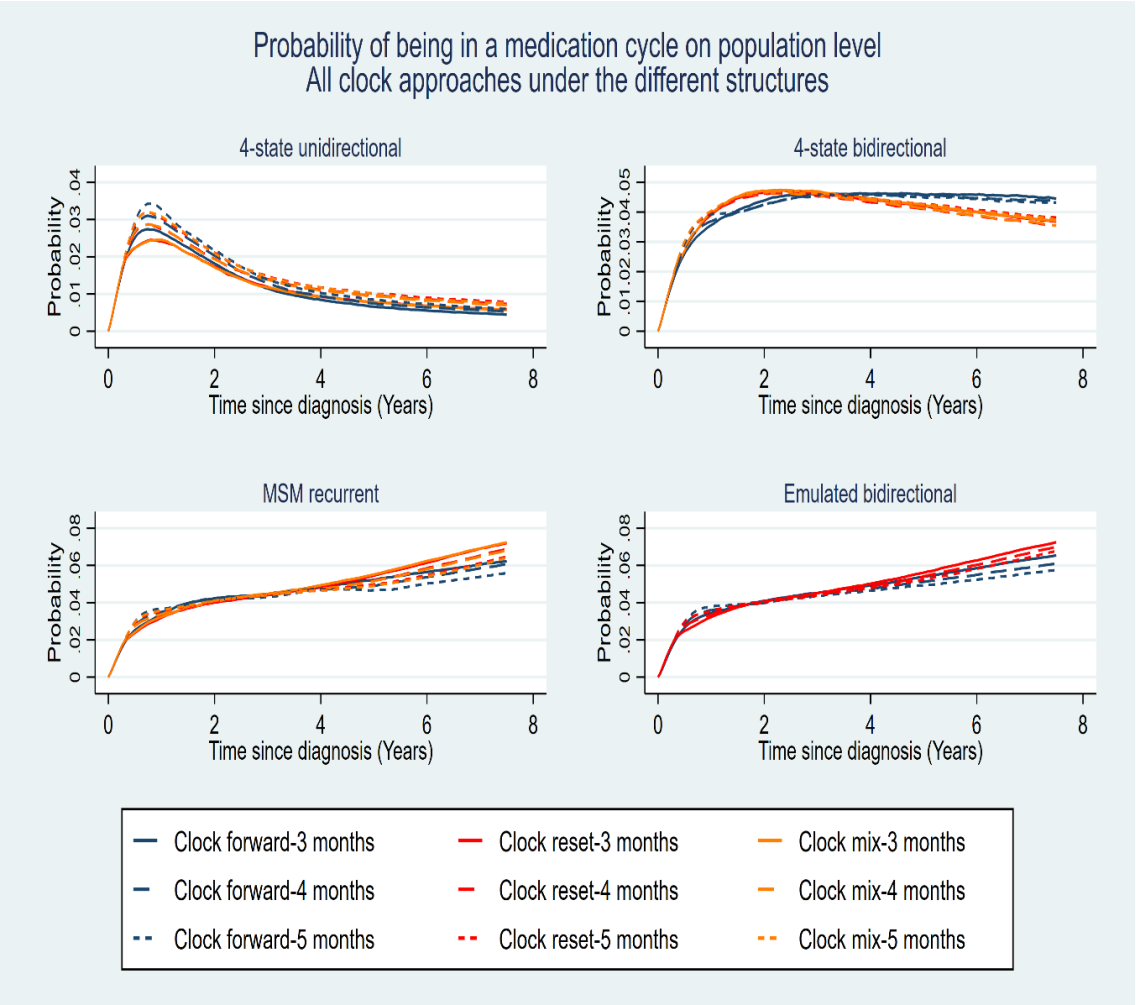
women diagnosed with breast cancer. Upon entering each new medication cycle, the total probability of being in a medication cycle for the rest of the individuals follow-up period tends to increase. Similarly, in Figure 5.6b, the total expected length of stay in medication cycles tends to increase upon entering each new medication cycle both for the comparison group and the BC-diagnosed women.

Figure 5.6. Estimates for the total probability of being in a medication cycle and the total length of stay under medication cycles across time since entering each cycle.



I performed sensitivity analyses to explore how different choices when defining the medication cycles (3,4 and 5 months rule) and when choosing a timescale for the transition intensity rates (Markov assumption, semi-Markov assumption or a mix of the two) can influence the estimations of the different multi-state models applied (Figure 5.7). The so-called 3 months rule in Sweden is the fact that in routine psychiatric practice, oral medications are not likely to be dispensed for more than 3 months at a time (105,106). Based on that rule, I defined whether a dispended medication should be considered as part of the same medication cycle or the beginning of a new cycle, based on the chronological distance with the previous date of dispensed medication of antidepressants. As this decision rule is not necessarily an accurate depiction of what happens in reality, I also used a 4-months and 5-months decision rule in a sensitivity analysis (Figure 5.7).

Figure 5.7. Comparison of the estimate of populational total probability of being in a medication cycle for different definitions of the medication cycles (3 months versus 4 months versus 5 months) under the different clock approaches for the multi-state structures D, E, F and G.



In summary, each multi-state structure used in **Study III**, properly addressed specific research questions, with simpler structures such as single-event survival analysis or competing risks addressing simpler research questions and bidirectional and recurrent multi-state structures addressing multiple, composite research questions about the use of antidepressant medication, taking into account the intermittent nature of prescription register data, fully utilizing the available information. For the complex multiple structures, the different definitions of the medication cycles and the different modeling choices (timescales of transitions and sharing information across restrictions) did not have a great in influence in the predicted probabilities of being in a medication cycle as it can be observed in Figure 5.7. However, I argue, that in the presence of several modelling choices, it is always advisable to explore and evaluate different options.

5.4 STUDY IV

In **Study IV**, I evaluated the use of multi-state models in a setting of recurrent events in the presence of a terminal event, when different individuals have different frailties, both for the recurrent and the terminal event process, and there may be an association between the two processes. Under such settings, approaches such as joint frailty models that directly model the variance in the frailty distribution and the association between the two processes is the more common choice. However, when interest lies in studying the marginal probabilities of the terminal and the recurrent event, multi-state models with recurrent event states and an absorbing terminal event state can also be used, indirectly accounting for the frailties via the risk set structure. I performed a simulation using Liu's joint frailty model (Section 4.9.1) as the model for the data generating mechanism (DGM), under different scenarios of variance θ in the gamma frailty distribution, association α between the recurrent and the terminal event process and sample size n during the data generation, with no covariates in the model, allowing for maximum time of observation equal to five years and a maximum number of observed recurrent events equal to ten. My aim was to assess the bias in the predicted probabilities for the terminal and recurrent events when using a recurrent multi-state structure for time $t = 3, 4, 5$ given $0, 1, 2, 3$ past recurrences up to time $t = 1$ year since the start of the follow-up. Two recurrent multistate modelling approaches were used, both of them using FPSMs for the baseline transition intensity rates. The first one, named MSM1, had separately estimated transition intensity rates. The second multi-state approach, named MSM2, had restrictions applied in the estimation of the transition intensity rates, so that all transition intensity rates towards the terminal state are proportional among themselves and all transition intensity rates towards the recurrent states are proportional among themselves.

Regarding the probability of the terminal event, under all scenarios of association between the recurrent and the terminal event processes, frailty variance and sample size, both the restricted and the unrestricted scenarios presented small bias of less than 0.01 event given 0, 1 and 2 past recurrences across the time points of prediction. Under scenarios of positive association between the two processes the bias in the predicted probability of the terminal event given 3 past recurrences was slightly higher than 0.01 for the restricted multi-state model (MSM2). Under adequate sample size (n=2000) and no association between the two processes the bias of the multi-state approaches was negligible, while under positive association and smaller sample sizes (n=500) the bias was still small but not negligible across time after start of follow-up and across number of past recurrences. Sharing information across transitions leads to better overall precision of the estimated probabilities of death but restrictions may lead to bias if they are unrealistic and should therefore be used in moderation, especially for countering data sparsity issues.

Regarding the probability of a new recurrent event, under all scenarios of association between the recurrent and the terminal event processes, frailty variance and sample size, both the restricted and the unrestricted MSM scenarios presented small bias of less than 0.01 event given 0, 1, 2 and 3 past recurrences across the time points of prediction. These results are not shown

in the main manuscript of **Study IV** because the predicted probabilities for the recurrent event from the joint frailty modelling approach had to be derived with a simulation-based approach and not with the analytical approach used for the probability of the terminal event based on Mauguen et al (107).

Figure 5.8. Dot plot of bias in the predicted probabilities for death up to years 2, 3, and 4, given 0, 1, 2, or 3 past recurrences over the three different modeling approaches under the different scenarios of n, α and θ.

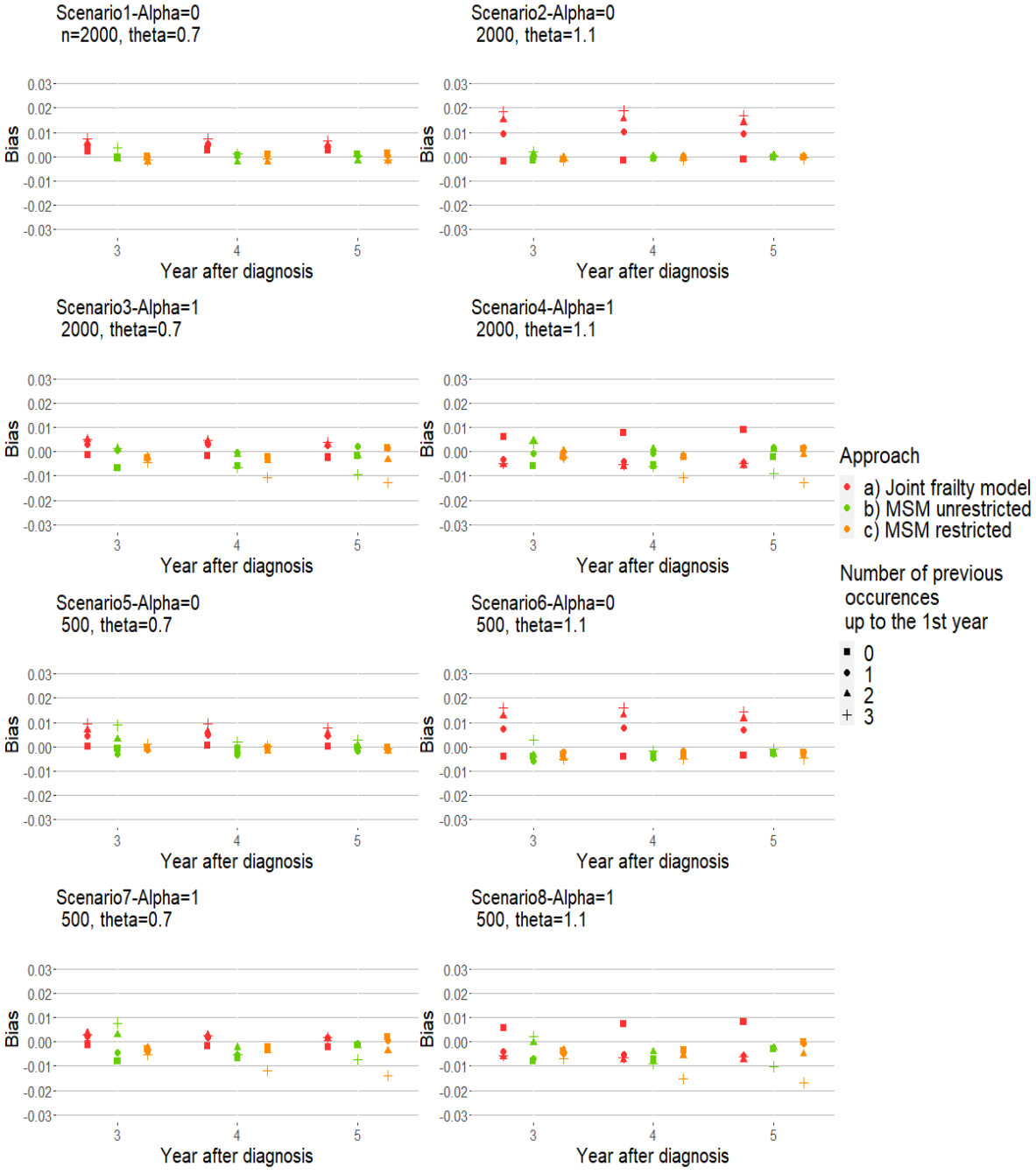
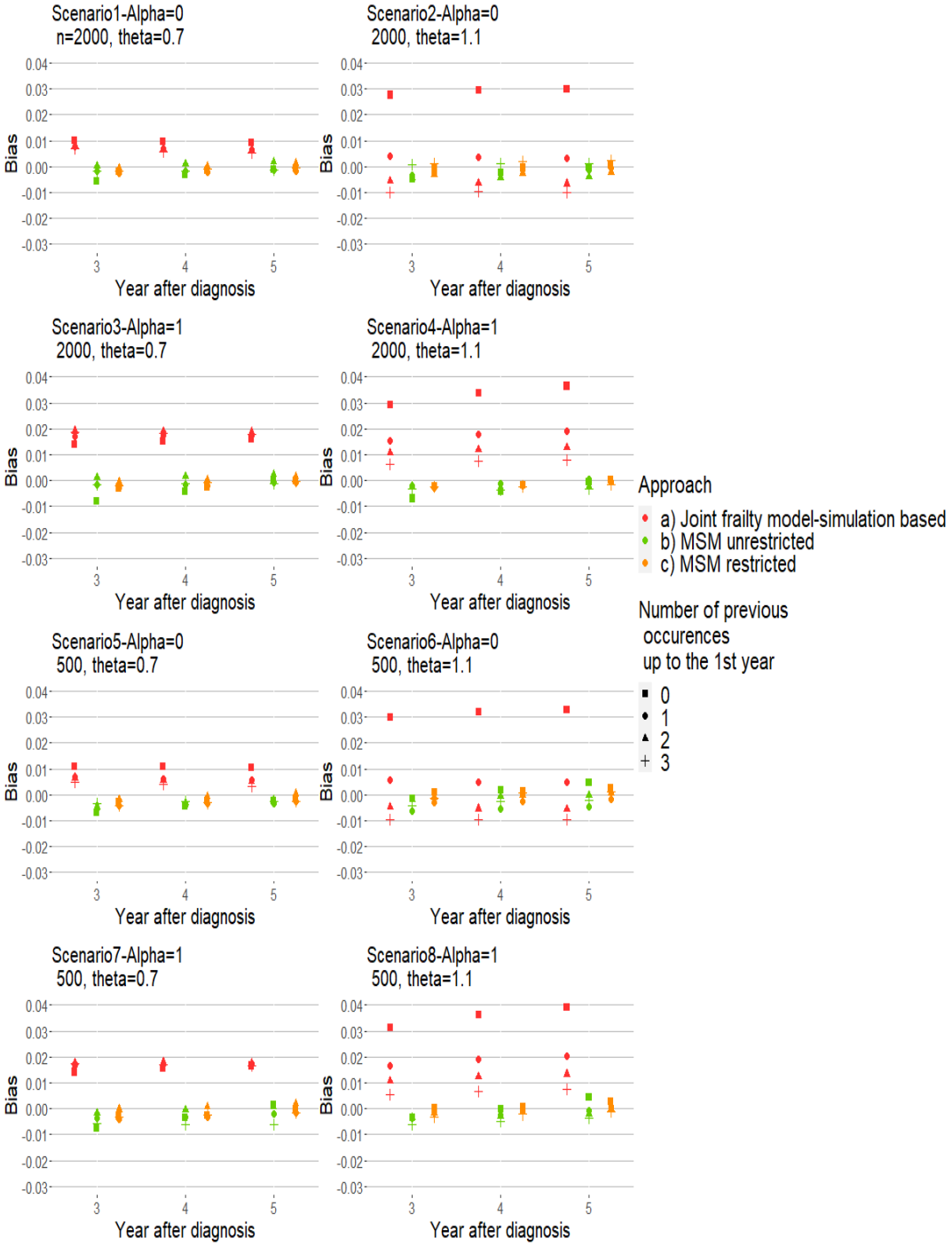


Figure 5.9. Dot plot of bias in the predicted probabilities for a new recurrence to years 2, 3, and 4, given 0, 1, 2, or 3 past recurrences over the three different modeling approaches under the different scenarios of n, α and θ.



We should note that the joint frailty model, as implemented by package `frailtypack` in R, has difficulty in estimating low or high values of variance in the frailty distribution as well as presenting converging issues. This bias in the estimation of the θ parameter in conjuction with small to moderate bias in the estimation of the association parameter α and the baseline hazard rate of recurrence, was reflected as substantial bias in the predicted probabilities of death under no association scenarios with high frailty variance, as the number of past recurrences

increases. The same was observed for the bias in the probability of new recurrence for all scenarios under high variance in frailties, especially given no prior recurrences.

The restricted multi-state modelling approach (MSM2) seemed to have a higher relative precision than the unrestricted multi-state model when both of them have their precisions compared with that of the joint frailty approach, with both multi-state approaches being less precise compared to the joint frailty approach. The unrestricted approach (MSM1) presented a lower precision than the joint frailty model that varies from -48% to almost -75% while the restricted approach MSM2 from -27% to -54% in the estimation of the probabilities for death up to year $t = 3$, given 1 and 2 past recurrences. This is not surprising, as the joint estimation of parameters leads to a smaller number of parameters to be estimated, which in turn leads to lower variance in the estimated parameters and thus predicted probabilities.

The multi-state approaches presented low bias for the prediction of probabilities of new recurrences and death given 0 or a number of previous recurrences, across time since the start of follow-up, when the data generating mechanism is based on an underlying joint frailty model. However, there are a series of issues to be considered when using multi-state models in this setting. The first issue, is a data sparsity issue in higher order states. As state order progresses fewer individuals tend to experience each new transition, resulting in sparsely populated transitions, leading to potential convergence and/or low precision issues. This data sparsity issue can be partly addressed via sharing information across transitions by imposing restrictions, as done in approach MSM2, tackling convergence issues and leading to better precision. However, if the assumptions that are reflected by the imposed restrictions are not realistic, for example, assuming baseline transition intensity rates that are proportional for the different recurrent events while this is not the case, this may result in introducing bias in the predicted probabilities. Another issue is that, when applying a recurrent multi-state structure in a setting of recurrent events in the presence of a terminal event, all the individuals with more than a certain number of recurrences will be pooled together in the last, semi-absorbing, recurrent state from which they can transition only towards the teminal state. These individuals are assumed by the multi-state model to have a common transition intensity rate towards the terminal state, no matter how many previous events each individual has. If the recurrent and the terminal processes are not associated, or the maximum number of observed recurrent events do not greatly surpass the the recurrent states of the multi-state structures, then the common transition intensity rate can be considered as a realistic assumption. However, if the two processes are associated then the predicted probabilities of the terminal events may be prone to bias.

In summary, multi-state models can be used in settings of recurrent events in the presence of a terminal event and the existence of individual frailties, for the prediction of probabilities of a new recurrent event or the terminal event, given no or previous recurrent events, as they indirectly account for these frailties via the risk set structure for each transition. However, careful consideration should be given during their application due to the aforemetioned issues.

6 DISCUSSION

When applying multi-state models, there is a series of modelling and structural choices to consider. In this section, I reflect on these choices and their implications. Non-parametric, semi-parametric or parametric multi-state models can be used in order to study transition probabilities/ state occupation probabilities as well as other measures of interest such as restricted expected length of stay (43,108). However, in the current study I focus on fully parametric multi-state models, with full estimation of the baseline transition intensity rates.

6.1 CONSIDERATION OF DIFFERENT MODELLING CHOICES

6.1.1 Baseline transition intensity rates

6.1.1.1 Transition intensity rate shapes

Baseline transition intensity rates of a multi-state model can be modeled either separately or jointly. When the parameters of the baseline transition intensity rate for a specific transition are estimated solely based on the observed times and events for that transition, then we have transition specific-estimation. In the case of separately estimated transition intensity rates, different transition intensity rates functions can be assumed for different transitions. A model selection process based on the AIC and BIC criteria can be followed separately for each transition in order to choose a baseline transition intensity rate function, ranging from simpler shapes such as exponential or Weibull up to gamma or spline functions. Model selection based on prior knowledge and the amount of information in data is also observed. For example, for transitions for which little a priori knowledge of the shape of the transition intensity rate exists or richer information is available in the form of high number of individuals at risk and number of events, flexible parametric models using spline functions can be used, which are generally not sensitive to knot number and location as long as there are sufficient knots (109–111).

6.1.1.2 Sharing information across transitions/ Joint estimation

Information can be shared across different transitions, by assuming a common shape of the transition rate function or a common transition rate function altogether. This can be achieved by imposing specific restrictions in the parameter estimation of the different baseline transition rate. A way to do that is to use a stacked multi-state model where the transition rate of the first transition is allowed to have a specific hazard shape, for example a spline function of the logarithm of time with four degrees of freedom. Transition indicator variables can then be included in the stacked model with main effects, restricting baseline transition rates for the rest of the transitions to be proportional to the first baseline transition rate. Then, restrictions on these main effects can be imposed, forcing some of them to be equal, resulting in identical baseline transition rates among the selected transitions. On the other hand, under this stacked model we can allow baseline transition rates to differ both in shape and scale from the baseline transition rate of reference (in our case the first one) by including spline interactions between time and the effects of each transition indicator variable of the model (non-proportional

hazards). Such a stacked multi-state model was fit in **Study III** for the "Emulated Bidirectional" structure.

6.1.1.3 Choice of timescales

Another choice when modelling the baseline transition intensity rates of a multi-state model is the timescale used. As mentioned in Section 4.8, there are different approaches to select from, a) the Markov assumption, using time since start of the process as the timescale for all transitions, b) the semi-Markov assumption, using time since entering the current state as the timescale for each transition, c) a mix of the two, where the timescale is selected based on subject –based knowledge, assuming that it is more natural for certain transitions to be functions of the total time since the start of the process and for other transitions to be functions of time since entering the current state, d) multiple timescales, where each baseline transition intensity rate is assumed to be a function of multiple timescales. In the Appendix of **Study III**, I present results from a sensitivity analysis comparing the estimation results regarding transition probabilities based on a Markov assumption, a semi-Markov assumption and a mix of the two (See Section 4.8). **In Study I**, I compared the use of two different timescales (time since diagnosis versus attained age) when modelling the other cause baseline mortality rate in a competing risk setting.

6.1.2 **Covariates**

6.1.2.1 Modelling the covariate effects

Including covariates in a multi-state model is accompanied with a series of modelling choices. In transition intensity-based multi-state models, each transition intensity rate can be thought of as a separate hazard model. Different sets of covariates may be used for different transition rates, and different assumptions can be made about those effects (proportional versus non-proportional hazards). The first step is to decide the criteria upon which a variable will be included as a covariate in each intensity rate model. In case we are interested in getting probability-based measures, we need to include the same variables as covariates in all transition rate models. A model selection process evaluating both the baseline transition intensity rate function selection (Section 6.1.1) and the existence of non-linear covariate effects and non-proportional hazards across the timescale for each covariate in the transition-specific hazard model can be made. I should note that, via the application of FPSM, one can flexibly model the interactions of covariate effects across the timescale of the transition. There are certain choices as to how one can model the covariate effects across transitions.

6.1.2.2 Joint estimation of covariate effects

One choice is to evaluate the covariate effects by modelling each effect separately for each transition. Another choice is to share information about the covariate effects across transitions by imposing restrictions, in a way similar to Section 6.1.1.2. For example, when fitting a stacked multi-state model, if we include a binary variable as a covariate in the model with a main effect (and optionally an interaction term with time to allow for non-proportional

hazards), then a common covariate effect across all transitions is assumed. If we allow an interaction between the covariate effect and the transition indicator variables, then a transition-specific covariate effect is incorporated. Through this process, we can choose to have specific transitions that share the same covariate effects, which may be desirable, in case we want to share information among transitions with more information (more events) and transitions with less information (sparsely populated) for which we are confident to assume that the covariate effect should be the same. In case of abundance of information (high number of events across all transitions), the most liberal choice would be to have unrestricted, transition-specific, time-varying effects (non- proportional hazards) of a covariate for all transition rates which can be induced by allowing triple interactions between transition indicator variable, the covariate and the timescale. However, we should consider that, even with large datasets, such as in **Study III** (more than 110.000 individuals), given multi-state structures with a high number of states, an issue of sparely populated transitions is likely to arise, leading to convergence issues.

6.1.2.3 Time dependent covariates

The multi-state process which we aim to study is a time-dependent process where the state value changes over time. It may be of interest to study how other time-varying factors relate with its evolution over time. A way to study that is to incorporate the time-varying factor as part of the multi-state structure. For example, we can consider the 3-state Illness-Death model (Section Figure 5.3a), where the intermediate state between the initial state of transplantation and the absorbing state of Relapse/Death is the Platelet Recovery state. Instead of having this multi-state structure, we can also choose to use a two state-model with only the initial and the absorbing state (also known as a typical single event survival model), and treat the platelet recovery as a time varying covariate, splitting the time of each individual before and after the platelet recovery and estimate the probability of dying with and without platelet recovery. The benefit of the multi-state structure is that within the multi-state framework, we can also study the probability of the intermediate state of platelet recovery itself, as well as deriving extra measures of interest such as length of stay in a post-transplantation state without a platelet recovery. For time-varying factors that are more complex functions of time, such as continuous or categorical biomarkers that regularly change over time, or even recurrent event processes, joint modelling of longitudinal and multi-state processes can be applied (112).

Given all these modelling choices for the baseline transition intensity rates, the covariate effects and when sharing information across different transitions, it is important to consider which covariates are to be included in the model or perform a sensitivity analysis to evaluate how sensitive the predicted measures of interest are, for different modelling choices.

6.2 STRUCTURAL CHOICES

6.2.1 Correspondence between structure and research question

The choice of multi-state structure usually ensues the data collection, thus provoking the question “Which is the optimal multi-state structure to use given the available data?”. However, each multi-state structure can be used to study multiple endpoints simultaneously and different

multi-state structures may target different underlying quantities, resulting in different interpretation of the estimated measures. In **Study III**, the interpretation of a subset of the estimated transition probabilities depends on the multi-state structure used. For example, under the 4-state Unidirectional model, I estimate the probability of being in the first medication cycle over the follow-up time while under the 4-state Bidirectional model I estimate the probability of being in a medication cycle over the follow-up time. Therefore, the correct question would be to ask "Which is the optimal multi-state structure to use given the research question of interest?"

Based on the disease pathway and the research questions of interest, specific events may be put as intermediate states, competing events may be considered as intermediate or absorbing states, backward transitions to previous states can be considered, recurrent events may be included in the form of recurrent states, gaps between at-risk periods can be incorporated as states, thus shaping the multi-state structure to be used. In **Study III**, I showed that, if someone is interested in the probability of medication initiation and still being alive up to time t after the start of the follow-up, a 3-state Illness-death model suffices. However, for a more composite underlying quantity such as the total probability of being in a medication cycle over the rest of the follow-up upon entering the first medication cycle, the use of a multi-state structure of recurrent couples of medication cycle/ discontinuation period states was needed.

Limitations to the range of potential multi-state structures can be posed by the type and amount of information within the data. If, as aforementioned, the multi-state model analysis is designed after the data collection, then information regarding an event that could serve as an intermediate state or a competing absorbing state may be missing. In that case, only a subset of all the potential structures can be used. This is an important factor to take into consideration during the selection of a multi-state structure.

6.2.2 Limitations

When applying a multi-state model, it is important to consider several factors that may limit either the structural or the modelling choices or both.

Complex multi-state structures typically indicate a high number of states and transitions. Depending on the structure, the addition of even one extra state may lead to the addition of several extra transitions towards that state. It follows that the number of transitions can become unmanageable even with a moderate number of states, leading to an issue of sparsely populated transitions. This data sparsity issue can pose a natural limit to the potential transitions that can be modelled, as it can lead to convergence issues of the transition-specific intensity rate models and low precision in the estimated parameters of the intensity rates and, by extension, low precision in the estimation of the transition probability and probability-based measures. A way to tackle this issue is to share information between transitions using restrictions when estimating the baseline transition rates as described in Section 6.1.1.2, assuming for example proportional baseline transition intensity rates among a cluster of transitions. Applying simpler parametric survival models such as exponential or Weibull can also lead to higher degree of

convergence and higher precision, as fewer parameters are estimated, leading to smaller variance in the estimations. However, these parametric assumptions are quite strong and may not reflect the real underlying hazard rates, potentially inserting bias in the predictions of the multi-state model. Another way to tackle the data-sparsity issue, especially in the case of multi-state structures with recurrent states is to have a semi-absorbing, non-terminal state where all individuals that have already experienced the previous states of the path stay into, until their censoring, the end of the multi-state process or until they transition to the terminal state (See recurrent multi-state structures of **Study III** and **Study IV**). The same issues can arise when modelling the covariate effects. Ideally, we would like to allow for transition-specific, time-varying effects of a covariate of interest for all transitions. However, there may be very few, or even zero individuals of a specific covariate pattern that experience certain transitions. This issue may lead to convergence and precision issues during the estimation of the covariate effects. A way to tackle this issue can be the sharing of information across certain transitions in respect of the covariate effects by imposing restrictions (same effects), as described in 6.1.2.2.

Other restrictions may have to do with the software implementation of multi-state models. For example, stacked multi-state models which can be used for imposing restrictions between baseline transition intensity rates and between covariate effects, can be computationally demanding under a composite transition matrix with high number of transitions, surpassing the maximum memory usage that a system can offer or presenting long execution times (Execution time of restricted recurrent multi-state model of **Study III:** $\backsim$ 12 hours). These issues may lead to a necessary compromise as to how flexibly the transition intensity rates and covariate effects can be modelled (**Study III**, Restricted recurrent/Emulated bidirectional multi-state structure).

Therefore, based on the discussion of Sections 6.1 and 6.2, the realistic question a researcher can afford to ask during multi-state model selection can more accurately be phrased as "Which is the optimal multi-state structure to use given the research question of interest, the information available in the data, and the traits and limitations of the alternative structures?"

6.3 ETHICAL CONSIDERATIONS WHEN APPLYING MULTI-STATE MODELS

In Sections 6.1 and 6.2 we discussed about the different multi-state structures and modelling choices, the different interpretations of measures depending on the structure, the issues and limitations during the selection and application of a multi-state model. As mentioned in Atici et al (113), "it is very important to use biostatistics principles and methods properly in all steps in order to impartially present information obtained through research". Therefore, it is of ethical importance to try to make the best possible structural and modelling choices when applying multi-state models. The research questions of interest should be carefully formulated and appropriate multi-state structures that can address them should be defined. Then, considering limitations in the data information, such as the population of transitions or information availability for the intermediate events, and the potential sharing of information across transitions as described in Section 6.1, a specific multi-state structure should be chosen. The flexibility when modelling the baseline transition intensity rates and the covariate effects

should be explored, via model selection procedures while also considering the precision of the estimates and the convergence of the models. Different restrictions/ assumptions about the relation between baseline intensity rates and covariates among transitions can be explored as well as different timescale approaches via sensitivity analyses. Sensitivity analysis is quite important as it can help in assessing the impact different modelling choices have on estimated measures of interest such as transition probabilities and expected length of stay in a stay (**Study III**). MSMplus, the interactive application I developed in **StudyII,** allows for the visual comparison of results from different multi-state model analyses so it can be a useful tool for sensitivity analyses.

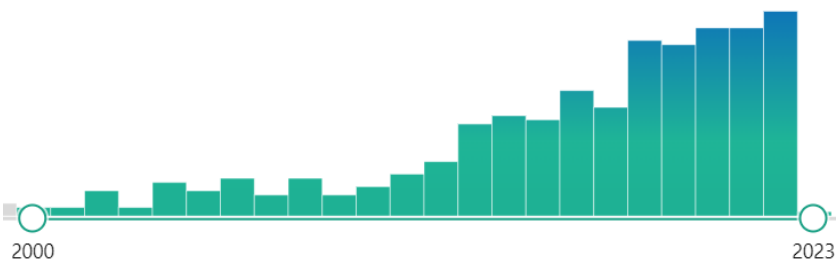
7 CONCLUSIONS

Throughout this work, I aimed to present, evaluate and discuss various related topics regarding competing risks and multi-state models. I focused on the definition and estimation of different measures of interest that can be derived under these models, as well as structural (different multi-state structures) and modelling choices (choice of timescales, sharing information across transitions via restrictions in the estimation process) when applying multi-state models. The notion of structural and modelling choices is also dealt within the manuscripts. The choice of timescale for the other cause mortality rate in **Study I** is essentially a modelling choice when applying competing risks. In **Study III** the use of different multi-state structures and different modelling approaches in regards with the timescales of the transitions are also different structural and modelling choices while trying to explore different research questions when using registry-based repeated prescriptions of antidepressants from the Swedish prescription registry. In **Study IV**, I explored the use of a recurrent multi-state modelling approach (a choice of modelling approach) in a setting of recurrent events under the presence of a terminal event, given a joint frailty model data generating mechanism.

Throughout the different sections, I referred to the plethora of measures that can be estimated via multi-state models across time and among covariate patterns and I stressed the importance of structure and model selection. The effective communication of the structure and estimation results of a multi-state model is therefore of paramount importance in order to deeply understand the multi-state process and conduct sensitivity analyses to assess the impact of the modelling choices (e.g timescales, sharing information across transitions).The RShiny application MSMplus developed in **Study II** was built with those principles at its core. While the first choice is to present the structure and estimation results of one multi-state model, the second choice it provides under its "Aims" label is to actively compare the results from two multi-state models. That is ideal for a quick sensitivity analysis of two multi-state models of the same structure but of different modelling choices.

The use of multi-state models in the epidemiological literature is still limited. However, this is gradually changing, with the number of MSM applications rising both in epidemiological studies and clinical trials focusing on cancer (Figure 7.1). Therefore, the responsible application and communication of multi-state models to the wider community of biomedical and epidemiological research is relevant now more than ever before.

Figure 7.1. Proportion of publications out of a total of 366 in Pubmed when setting as key words "multi-state" and "cancer", filtering from year 2000 up to year 2022.



8 POINTS OF PERSPECTIVE

This work addressed, among others topics, the issue of using different timescales in a competing risk and a multi-state setting. The choice of timescale for each cause-specific hazard model (competing risks setting) or transition-specific intensity rate model (multi-state models) may differ depending on the nature of the event/transition and subject specific knowledge and can influence the estimated cumulative incidence functions and transition probabilities. The magnitude of this influence depends on many aforementioned factors such as indirectly modelling the effect of other timescales as main effects in the model in a linear or non-linear way and with or without interactions with the main timescale the complexity. It may also depend on the number of competing events, the multi-state structure, the risk-set sizes and more. It is therefore of importance for future research to focus on the simultaneous flexible parametric modelling of multiple timescales in a competing risk and a multi-state setting, so that the baseline hazard rate for a competing event or a transition intensity rates of a multi-state model are functions of multiple timescales, also allowing for flexible effects interaction between the different timescales. This way, the estimated measures, such as the cause-specific CIFs, the transition probabilities and other multi-state related measures will be able to be derived as non-linear functions of multiple time-scales and presented not only across one but multiple timescales, serving the better understanding of the disease pathway.

There is a rising interest in the use of multi-state models in cancer clinical trials of phase II and phase III (25,114–116), where the interest lies in evaluating multiple endpoints instead of one, for example evaluating both overall survival and progression-free survival as well as the association between progression and overall survival. Therefore, interest should be given in future research of optimal designs in clinical trials for the use of multi-state structures, securing the desirable type I error and power, and allowing for interim analyses and other clinical trial design traits.

The big datasets available in the last decade, especially in Sweden via data linkage from multiple registers, allow for the use of more complex, high-parametrized models that can address composite research questions and cope with data of complex nature. Joint longitudinal and survival models with one or more biomarkers/longitudinal outcomes can be fit in the form of competing risks and multi-state structures, allowing the study of the relation between these outcomes and the multi-state process.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to the following:

Paul C. Lambert, my main supervisor, for believing in me, giving me the time and the space to grow as a researcher. Every time I was leaping two steps ahead, he was there to remind me to be thorough, to focus on getting the simple stuff working first before moving on to more complicated analyses. I definitely feel I am a much more experienced researcher now, being more methodological and careful when approaching a new problem, with much better solving skills than before. He also put a lot of trust in me which in turn helped me become more responsible at my work and feel more confident. Thank you Paul for your trust and for giving me the opportunity to work at MEB these last four years and meet so many special people.

Therese M-L. Andersson, my co-supervisor, for her advice and feedback which was always precise and accurate, enhancing the quality of my manuscripts, letting me develop my thoughts and ideas even if they were mistaken sometimes. I would also like to thank her for her help when dealing with paperwork in Swedish and always having her office door open so that I could easily drop by and ask for a hint of how to tackle issues that I needed to resolve. Thanks for your kindness and support Therese!

Michael J. Crowther, my co-supervisor, for his vital feedback on coding issues but also during the conceptualization of the aim of the manuscripts. Michael, you have the charisma of conveying difficult concepts in a few sentences in such a cohesive way, making them sound simple for me. Your skills and hard work have been an example for me and would like to wish you good luck in your future endeavors whether you stay outside or return to the academia.

Donghao Lu, my co-author in Study III, whose feedback was detrimental in the development of the manuscript. Donghao, thanks a lot for your prompt feedback, always on spot and in time even though I know your time was very limited.

Mats Lambe, my co-author, thanks to whom I was able to work with the Prescribed Drug Register data of BCBaSe 2.0 database and apply interesting multi-state structures. Mats, thanks a lot for your feedback during the development of the manuscript.

Keith Humphreys, chair of my defense. I deeply enjoyed our conversations about life. You are one of the few people that remain forever young, never making me feel a senior to junior distance during our discussions, always felt more like two friends chatting.

Paul Dickman, head of the survival analysis research group. Paul, thanks a lot for having me as a member in your group. I enjoyed the trips of our group in London, Treviso and Oslo and having nice chats especially during breakfast. Also, thanks a lot for supporting the biostatistics group all these years and for keep pushing for the realization of a master program of biostatistics in KI, a program which is so much needed and will contribute to the growth of the community of biostatisticians!

Marie Jansson, secretary of the biostatistics corridor for her help whenever I needed it and her warm energy even during the Swedish winter. Marie, as with Keith, I really enjoyed that you never made me feel any distance, that we had the chance across the years to chat about politics, religion, our life experiences. You were very tolerant and patient no matter how many times I kept asking for the same project numbers again and again. Above all, a person to say kalimera to every morning. Thanks for everything!

Alessandra Nanni, study coordinator. Alessandra, thanks a lot for your support, not only for the paperwork needed throughout the studies, but also for your advice whenever I needed it.

My corridor colleagues, both seniors and PhD students: Balram Rai, Letizia Orsini, Birzhan Akynkozhayev, Enoch Yi-Tung Chen, Elisavet Syriopoulou, Frida Lundberg, Nurgul Bertykova, Rickard Strandberg, Alessandro Gasparini, Yuliya Leontyeva, Lu Pan, Adam Brand, Iuliana Ciocanea-Teodorescu,Wenjiang Deng, Zheng Ning, Valentin Vancak, Xiaoyang Du, Shuang Hao, Ana Johansson, Rino Bellocco, Cecilia Lundholm, Sven Sandin, Marie Reilly, Yudi Pawitan, Mark Clements, Erin Gabriel, Alex Ploner, and Arvid Sjölander for creating a friendly work atmosphere. This is rare to find nowadays and I really appreciate it. Thanks for having me (and tolerating me)!

My office mates past and present: Elizabeth, Frida, Pablo, Yuliya, Shuang, Rickard, each one with his/her unique style and vibe. Thanks guys for making me comfortable, I hope I made you feel the same!

My friends from MEB Ale, Abi, Pablo, Marco, Ailema, Erwei, Enoch, Mao, Jet, Nita, Laura, Marta, Arvid, Zhung, Betty, Maya, Philippe.

9 REFERENCES

1. Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. Stat Med. 2007 Dec 20;26(11):2389–430.

2. Newland RC, Chan C, Chapuis PH, Keshava A, Rickard MJFX, Young CJ, et al. Competing risks analysis of the effect of local residual tumour on recurrence and cancer-specific death after resection of colorectal cancer: implications for staging. Pathology. 2018 Oct;50(6):600–6.

3. Onozuka D, Nakamura Y, Tsuji G, Furue M. Cancer- and noncancer-specific cumulative incidence of death after exposure to polychlorinated biphenyls and dioxins: A competing risk analysis among Yusho patients. Environ Int. 2021 Feb;147:106320.

4. Pathak M, S Deo SNV, Dwivedi SN, Vishnubhatla S, Thakur B. Comparison of hazard models with and without consideration of competing risks to assess the effect of neoadjuvant chemotherapy on locoregional recurrence among breast cancer patients. J Cancer Res Ther. 2021;17(4):982–7.

5. Rosenberg MA. Competing risks to breast cancer mortality. J Natl Cancer Inst Monogr. 2006;(36):15–9.

6. Shim SH, Lim MC, Lee D, Won YJ, Ha HI, Chang HK, et al. Cause-specific mortality rate of ovarian cancer in the presence of competing risks of death: a nationwide population-based cohort study. J Gynecol Oncol. 2022 Jan;33(1):e5.

7. Tan KS, Eguchi T, Adusumilli PS. Competing risks and cancer-specific mortality: why it matters. Oncotarget. 2018 Jan 26;9(7):7272–3.

8. van Kruijsdijk RCM, Eijkemans MJC, Visseren FLJ. [Competing risks in clinical research]. Ned Tijdschr Geneeskd. 2012;156(46):A5176.

9. Vilaprinyo E, Gispert R, Martínez-Alonso M, Carles M, Pla R, Espinàs JA, et al. Competing risks to breast cancer mortality in Catalonia. BMC Cancer. 2008 Nov 12;8:331.

10. Xu YB, Liu H, Cao QH, Ji JL, Dong RR, Xu D. Evaluating overall survival and competing risks of survival in patients with early-stage breast cancer using a comprehensive nomogram. Cancer Med. 2020 Jun;9(12):4095–106.

11. Zhang S, Ivy JS, Wilson JR, Diehl KM, Yankaskas BC. Competing risks analysis in mortality estimation for breast cancer patients from independent risk groups. Health Care Manag Sci. 2014 Sep;17(3):259–69.

12. de Bock GH, Putter H, Bonnema J, van der Hage JA, Bartelink H, van de Velde CJ. The impact of loco-regional recurrences on metastatic progression in early-stage breast cancer: a multistate model. Breast Cancer Res Treat. 2009 Sep;117(2):401–8.

13. de Boer AZ, Bastiaannet E, Schetelig J, de Glas NA, Manevksi D, Putter H, et al. Breast cancer mortality of older patients with and without recurrence analysed by novel multi-state models. Eur J Cancer. 2022 Oct;174:212–20.

14. Plym A, Johansson ALV, Bower H, Voss M, Holmberg L, Fredriksson I, et al. Causes of sick leave, disability pension, and death following a breast cancer diagnosis in women of working age. Breast. 2019 Jun;45:48–55.

15. Putter H, van der Hage J, de Bock GH, Elgalta R, van de Velde CJH. Estimation and prediction in a multi-state model for breast cancer. Biom J. 2006 Jun;48(3):366–80.

16. Rosner B, Glynn RJ, Eliassen AH, Hankinson SE, Tamimi RM, Chen WY, et al. A Multi-State Survival Model for Time to Breast Cancer Mortality among a Cohort of Initially Disease-Free Women. Cancer Epidemiol Biomarkers Prev. 2022 Aug 2;31(8):1582–92.

17. Vasheghani Farahani M, Ataee Dizaji P, Rashidi H, Mokarian F, Biglarian A. Application of Multi-State Model in Analyzing of Breast Cancer Data. J Res Health Sci. 2020 Jan 5;19(4):e00465.

18. Xu C, Ravva P, Dang JS, Laurent J, Adessi C, McIntyre C, et al. A continuous-time multistate Markov model to describe the occurrence and severity of diarrhea events in metastatic breast cancer patients treated with lumretuzumab in combination with pertuzumab and paclitaxel. Cancer Chemother Pharmacol. 2018 Sep;82(3):395–406.

19. Rotolo F, Dunant A, Chevalier TL, Pignon JP, Arriagada R. Adjuvant cisplatin-based chemotherapy in nonsmall-cell lung cancer: new insights into the effect on failure type via a multistate approach. Annals of Oncology. 2014 Nov 1;25(11):2162–6.

20. Jeong WG, Choi H, Chae KJ, Kim J. Prognosis and recurrence patterns in patients with early stage lung cancer: a multi-state model approach. Transl Lung Cancer Res. 2022 Jul;11(7):1279–91.

21. Conlon ASC, Taylor JMG, Sargent DJ. Multi-state models for colon cancer recurrence and death with a cured fraction. Stat Med. 2014 May 10;33(10):1750–66.

22. Álvaro-Meca A, Akerkar R, Alvarez-Bartolome M, Gil-Prieto R, Rue H, de Miguel ÁG. Factors involved in health-related transitions after curative resection for pancreatic cancer. 10-years experience: a multi state model. Cancer Epidemiol. 2013 Feb;37(1):91–6.

23. Plym A, Clements M, Voss M, Holmberg L, Stattin P, Lambe M. Duration of sick leave after active surveillance, surgery or radiotherapy for localised prostate cancer: a nationwide cohort study. BMJ Open. 2020 Mar 9;10(3):e032914.

24. Conlon ASC, Taylor JMG, Sargent DJ. Multi-state models for colon cancer recurrence and death with a cured fraction. Stat Med. 2014 May 10;33(10):1750–66.

25. Danzer MF, Terzer T, Berthold F, Faldum A, Schmidt R. Confirmatory adaptive group sequential designs for single-arm phase II studies with multiple time-to-event endpoints. Biom J. 2022 Feb;64(2):312–42.

26. Le-Rademacher JG, Peterson RA, Therneau TM, Sanford BL, Stone RM, Mandrekar SJ. Application of multi-state models in cancer clinical trials. Clin Trials. 2018 Oct;15(5):489–98.

27. Xia F, George SL, Wang X. A Multi-state Model for Designing Clinical Trials for Testing Overall Survival Allowing for Crossover after Progression. Stat Biopharm Res. 2016;8(1):12–21.

28. Wu WYY, Nyström L, Jonsson H. Estimation of overdiagnosis in breast cancer screening using a non-homogeneous multi-state model: A simulation study. J Med Screen. 2018 Dec;25(4):183–90.

29. Uhry Z, Hédelin G, Colonna M, Asselain B, Arveux P, Rogel A, et al. Multi-state Markov models in cancer screening evaluation: a brief review and case study. Stat Methods Med Res. 2010 Oct;19(5):463–86.

30. Sutradhar R, Gu S, Paszat LF. Multistate transitional models for measuring adherence to breast cancer screening: A population-based longitudinal cohort study with over two million women. J Med Screen. 2017 Jun;24(2):75–82.

31. Kumar V, Cohen JT, van Klaveren D, Soeteman DI, Wong JB, Neumann PJ, et al. Risk-targeted lung cancer screening: A cost effectiveness analysis. Ann Intern Med. 2018 Feb 6;168(3):161–9.

32. Noordzij M, Leffondré K, van Stralen KJ, Zoccali C, Dekker FW, Jager KJ. When do we need competing risks methods for survival analysis in nephrology? Nephrol Dial Transplant. 2013 Nov;28(11):2670–7.

33. Satagopan JM, Ben-Porat L, Berwick M, Robson M, Kutler D, Auerbach AD. A note on competing risks in survival data analysis. Br J Cancer. 2004 Oct 4;91(7):1229–35.

34. Johansen R. An Empirical Transition Matrix for Non-homogeneous Markov Chains Based on Censored Observations. In 1978 [cited 2023 Jan 2]. Available from: https://www.semanticscholar.org/paper/An-Empirical-Transition-Matrix-for-Non-homogeneous-Johansen/86db6591763c2a7d7285068eb2186d945645e670

35. The statistical analysis of failure time data. By J.D. Kalbfleisch and R.L. Prentice. John Wiley & Sons, Inc., New York, 1980. xi + 321 pp. U.S. $31.50, C $40.35. ISBN 0-471-05519-0. Canadian Journal of Statistics. 1982;10(1):64–6.

36. Fine JP, Gray RJ. A Proportional Hazards Model for the Subdistribution of a Competing Risk. Journal of the American Statistical Association. 1999;94(446):496–509.

37. Lambert PC, Wilkes SR, Crowther MJ. Flexible parametric modelling of the cause-specific cumulative incidence function. Stat Med. 2017 Apr 30;36(9):1429–46.

38. Royston P, Parmar MKB. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Stat Med. 2002 Aug 15;21(15):2175–97.

39. Lambert PC, Royston P. Further Development of Flexible Parametric Models for Survival Analysis. The Stata Journal. 2009 Aug 1;9(2):265–90.

40. Hinchliffe SR, Lambert PC. Flexible parametric modelling of cause-specific hazards to estimate cumulative incidence functions. BMC Med Res Methodol. 2013 Feb 6;13:13.

41. van Houwelingen J (Hans), Putter H. Dynamic predicting by landmarking as an alternative for multi-state modeling: An application to acute lymphoid leukemia data. Lifetime data analysis. 2008 Nov 1;14:447–63.

42. Ieva F, Jackson CH, Sharples LD. Multi-State modelling of repeated hospitalisation and death in patients with Heart Failure: the use of large administrative databases in clinical epidemiology. Stat Methods Med Res. 2017 Jun;26(3):1350–72.

43. Crowther MJ, Lambert PC. Parametric multistate survival models: Flexible modelling allowing transition-specific distributions with application to estimating clinically useful measures of effect differences. Statistics in Medicine. 2017;36(29):4719–42.

44. Jackson C. Multi-State Models for Panel Data: The msm Package for R. Journal of Statistical Software. 2011 Jan 4;38:1–28.

45. Cook RJ, Lawless JF. Statistical Issues in Modeling Chronic Disease in Cohort Studies. Stat Biosci. 2014 May;6(1):127–61.

46. Iacobelli S, Carstensen B. Multiple time scales in multi-state models. Statist Med. 2013 Dec 30;32(30):5315–27.

47. Colzani E, Johansson ALV, Liljegren A, Foukakis T, Clements M, Adolfsson J, et al. Time-dependent risk of developing distant metastasis in breast cancer patients according to treatment, age and tumour characteristics. Br J Cancer. 2014 Mar 4;110(5):1378–84.

48. Putter H, van der Hage J, de Bock GH, Elgalta R, van de Velde CJH. Estimation and Prediction in a Multi-State Model for Breast Cancer. Biometrical Journal. 2006;48(3):366–80.

49. Latouche A, Allignol A, Beyersmann J, Labopin M, Fine JP. A competing risks analysis should report results on all cause-specific hazards and cumulative incidence functions. J Clin Epidemiol. 2013 Jun;66(6):648–53.

50. National Cancer Register [Internet]. Socialstyrelsen. [cited 2023 Jan 3]. Available from: https://www.socialstyrelsen.se/en/statistics-and-data/registers/national-cancer-register/

51. Wadsten C, Wennstig AK, Garmo H, Lambe M, Blomqvist C, Holmberg L, et al. Data Resource Profile: Breast Cancer Data Base Sweden 2.0 (BCBaSe 2.0). Int J Epidemiol. 2022 Jan 6;50(6):1770–1771f.

52. Wennman-Larsen A, Nilsson MI, Saboonchi F, Olsson M, Alexanderson K, Fornander T, et al. Can breast cancer register data on recommended adjuvant treatment be used as a proxy for actually given treatment? Eur J Oncol Nurs. 2016 Jun;22:1–7.

53. Brooke HL, Talbäck M, Hörnblad J, Johansson LA, Ludvigsson JF, Druid H, et al. The Swedish cause of death register. Eur J Epidemiol. 2017;32(9):765–73.

54. Patientregistret [Internet]. Socialstyrelsen. [cited 2023 Jan 3]. Available from: https://www.socialstyrelsen.se/statistik-och-data/register/patientregistret/

55. Wadsten C, Heyman H, Holmqvist M, Ahlgren J, Lambe M, Sund M, et al. A validation of DCIS registration in a population-based breast cancer quality register and a study of treatment and prognosis for DCIS during 20 years. Acta Oncol. 2016 Nov;55(11):1338–43.

56. Löfgren L, Eloranta S, Krawiec K, Asterkvist A, Lönnqvist C, Sandelin K, et al. Validation of data quality in the Swedish National Register for Breast Cancer. BMC Public Health. 2019 May 2;19(1):495.

57. Mattsson B, Wallgren A. Completeness of the Swedish Cancer Register. Non-notified cancer cases recorded on death certificates in 1978. Acta Radiol Oncol. 1984;23(5):305–13.

58. Barlow L, Westergren K, Holmberg L, Talbäck M. The completeness of the Swedish Cancer Register: a sample survey for year 1998. Acta Oncol. 2009;48(1):27–33.

59. The EBMT Patient Registry [Internet]. EBMT. [cited 2023 Jan 3]. Available from: https://www.ebmt.org/ebmt-patient-registry

60. Rondeau V, Gonzalez JR. frailtypack: A computer program for the analysis of correlated failure time data using penalized likelihood estimation. Computer Methods and Programs in Biomedicine. 2005 Nov 1;80(2):154–64.

61. Gonzalez JR, Fernandez E, Moreno V, Ribes J, Peris M, Navarro M, et al. Sex differences in hospital readmission among colorectal cancer patients. J Epidemiol Community Health. 2005 Jun;59(6):506–11.

62. WMA - The World Medical Association-WMA Declaration of Helsinki – Ethical Principles for Medical Research Involving Human Subjects [Internet]. [cited 2023 Jan 3]. Available from: https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/

63. General Data Protection Regulation (GDPR) – Official Legal Text [Internet]. General Data Protection Regulation (GDPR). [cited 2023 Jan 3]. Available from: https://gdpr-info.eu/

64. Omar RZ, Stallard N, Whitehead J. A parametric multistate model for the analysis of carcinogenicity experiments. Lifetime Data Anal. 1995;1(4):327–46.

65. Titman AC. Flexible Nonhomogeneous Markov Models for Panel Observed Data. Biometrics. 2011;67(3):780–7.

66. Król A, Saint-Pierre P. SemiMarkov: An R Package for Parametric Estimation in Multi-State Semi-Markov Models. Journal of Statistical Software. 2015 Aug 27;66:1–16.

67. Blaser N, Vizcaya LS, Estill J, Zahnd C, Kalesan B, Egger M, et al. gems: An R Package for Simulating from Disease Progression Models. J Stat Softw. 2015 Mar;64(10):1–22.

68. Jackson C, Unit MB. Flexible parametric multi-state modelling with flexsurv.

69. Clements M, Liu XR, Christoffersen B, Lambert P, Jakobsen LH, Gasparini A, et al. rstpm2: Smooth Survival Models, Including Generalized Survival Models [Internet]. 2023 [cited 2023 Feb 6]. Available from: https://CRAN.R-project.org/package=rstpm2

70. Mozumder SI, Rutherford MJ, Lambert PC. Direct likelihood inference on the cause-specific cumulative incidence function: a flexible parametric regression modelling approach. Stat Med. 2018 Jan 15;37(1):82–97.

71. Eloranta S, Lambert PC, Andersson TML, Björkholm M, Dickman PW. The application of cure models in the presence of competing risks: a tool for improved risk communication in population-based cancer patient survival. Epidemiology. 2014 Sep;25(5):742–8.

72. Edgren G, Hjalgrim H, Rostgaard K, Lambert P, Wikman A, Norda R, et al. Transmission of Neurodegenerative Disorders Through Blood Transfusion: A Cohort Study. Ann Intern Med. 2016 Sep 6;165(5):316–24.

73. Hinchliffe SR, Seaton SE, Lambert PC, Draper ES, Field DJ, Manktelow BN. Modelling time to death or discharge in neonatal care: an application of competing risks. Paediatr Perinat Epidemiol. 2013 Jul;27(4):426–33.

74. Hultcrantz M, Wilkes SR, Kristinsson SY, Andersson TML, Derolf ÅR, Eloranta S, et al. Risk and Cause of Death in Patients Diagnosed With Myeloproliferative Neoplasms in Sweden Between 1973 and 2005: A Population-Based Study. J Clin Oncol. 2015 Jul 10;33(20):2288–95.

75. The Theory of Stochastic Processes | D.R. Cox | Taylor & Francis eBook [Internet]. [cited 2023 Jan 17]. Available from: https://www.taylorfrancis.com/books/mono/10.1201/9780203719152/theory-stochastic-processes-cox

76. Crowther MJ, Lambert PC. Simulating biologically plausible complex survival data. Statistics in Medicine. 2013;32(23):4118–34.

77. Grand MK, Putter H. Regression models for expected length of stay. Stat Med. 2016 Mar 30;35(7):1178–92.

78. David HA, Moeschberger ML. The theory of competing risks. London: Griffin; 1978. 103 p. (Griffin's statistical monographs & courses).

79. Beyersmann J, Latouche A, Buchholz A, Schumacher M. Simulating competing risks data in survival analysis. Statistics in Medicine. 2009;28(6):956–71.

80. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Statistics in Medicine. 2019;38(11):2074–102.

81. Korn EL, Graubard BI, Midthune D. Time-to-event analysis of longitudinal follow-up of a survey: choice of the time-scale. Am J Epidemiol. 1997 Jan 1;145(1):72–80.

82. Canchola A, Stewart S, Center NCC, Bernstein L. Cox Regression Using Different Time Scales.

83. Efron B. Logistic Regression, Survival Analysis, and the Kaplan-Meier Curve. Journal of the American Statistical Association. 1988 Jun 1;83(402):414–25.

84. Cox DR. Regression Models and Life-Tables. Journal of the Royal Statistical Society: Series B (Methodological). 1972;34(2):187–202.

85. Batyrbekova N, Bower H, Dickman PW, Ravn Landtblom A, Hultcrantz M, Szulkin R, et al. Modelling multiple time-scales with flexible parametric survival models. BMC Medical Research Methodology. 2022 Nov 9;22(1):290.

86. Jackson C. Multi-state modelling with R: the msm package.

87. Wreede LC de, Fiocco M, Putter H. mstate: An R Package for the Analysis of Competing Risks and Multi-State Models. Journal of Statistical Software. 2011 Jan 4;38:1–30.

88. Therneau, T.M. and Grambsch, P.M. (2000) Modeling Survival Data Extending the Cox Model. Springer, Berlin. - References - Scientific Research Publishing [Internet]. [cited 2023 Jan 3]. Available from: https://www.scirp.org/%28S%28351jmbntvnsjt1aadkposzje%29%29/reference/referencespapers.aspx?referenceid=1984629

89. Kelly PJ, Lim LL. Survival analysis for recurrent event data: an application to childhood infectious diseases. Stat Med. 2000 Jan 15;19(1):13–33.

90. Kessing LV, Andersen PK, Mortensen PB, Bolwig TG. Recurrence in affective disorder. I. Case register study. Br J Psychiatry. 1998 Jan;172:23–8.

91. Hougaard P. Frailty models for survival data. Lifetime Data Anal. 1995 Sep 1;1(3):255–73.

92. Mazroui Y, Mathoulin-Pelissier S, Soubeyran P, Rondeau V. General joint frailty model for recurrent event data with a dependent terminal event: Application to follicular lymphoma data. Stat Med. 2012 Dec 20;31(11–12):1162–76.

93. Liu L, Wolfe RA, Huang X. Shared Frailty Models for Recurrent Events and a Terminal Event. Biometrics. 2004;60(3):747–56.

94. Andersen PK, Angst J, Ravn H. Modeling marginal features in studies of recurrent events in the presence of a terminal event. Lifetime Data Anal. 2019 Oct;25(4):681–95.

95. Furberg JK, Andersen PK, Korn S, Overgaard M, Ravn H. Bivariate pseudo-observations for recurrent event analysis with terminal events. Lifetime Data Anal [Internet]. 2021 Nov 5 [cited 2023 Jan 3]; Available from: https://link.springer.com/10.1007/s10985-021-09533-5

96. Andersen PK, Gill RD. Cox's Regression Model for Counting Processes: A Large Sample Study. The Annals of Statistics. 1982;10(4):1100–20.

97. Prentice RL, Williams BJ, Peterson AV. On the Regression Analysis of Multivariate Failure Time Data. Biometrika. 1981;68(2):373–9.

98. Wei LJ, Lin DY, Weissfeld L. Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions. Journal of the American Statistical Association. 1989 Dec 1;84(408):1065–73.

99. Andersen PK, Keiding N. Multi-state models for event history analysis. Stat Methods Med Res. 2002 Apr;11(2):91–115.

100. Amorim LD, Cai J. Modelling recurrent events: a tutorial for analysis in epidemiology. Int J Epidemiol. 2015 Feb;44(1):324–33.

101. Villegas R, Julià O, Ocaña J. Empirical study of correlated survival times for recurrent events with proportional hazards margins and the effect of correlation and censoring. BMC Med Res Methodol. 2013 Jul 24;13:95.

102. Stata Bookstore: Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model [Internet]. [cited 2023 Jan 25]. Available from: https://www.stata.com/bookstore/flexible-parametric-survival-analysis-stata/

103. Syriopoulou E, Mozumder SI, Rutherford MJ, Lambert PC. Estimating causal effects in the presence of competing events using regression standardisation with the Stata command standsurv. BMC Medical Research Methodology. 2022 Aug 13;22(1):226.

104. Skourlis N, Crowther MJ, Andersson TML, Lambert PC. On the choice of timescale for other cause mortality in a competing risk setting using flexible parametric survival models. Biometrical Journal. 2022;64(7):1161–77.

105. Chang Z, Lichtenstein P, Långström N, Larsson H, Fazel S. Association Between Prescription of Major Psychotropic Medications and Violent Reoffending After Prison Release. JAMA. 2016 Nov 1;316(17):1798–807.

106. Fazel S, Zetterqvist J, Larsson H, Långström N, Lichtenstein P. Antipsychotics, mood stabilisers, and risk of violent crime. The Lancet. 2014 Sep;384(9949):1206–14.

107. Mauguen A, Rachet B, Mathoulin-Pélissier S, MacGrogan G, Laurent A, Rondeau V. Dynamic prediction of risk of death using history of cancer recurrences in joint frailty models. Stat Med. 2013 Dec 30;32(30):5366–80.

108. Hill M, Lambert PC, Crowther M. Non-parametric estimation in multi-state survival models: An update to msaj. London Stata Conference 2020 [Internet]. 2020 Sep 11 [cited 2023 Jan 3]; Available from: https://ideas.repec.org//p/boc/usug20/02.html

109. Syriopoulou E, Mozumder SI, Rutherford MJ, Lambert PC. Robustness of individual and marginal model-based estimates: A sensitivity analysis of flexible parametric models. Cancer Epidemiology. 2019 Feb 1;58:17–24.

110. Rutherford MJ, Crowther MJ, Lambert PC. 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. 2015 Mar 4;85(4):777–93.

111. Bower H, Crowther MJ, Rutherford MJ, Andersson TML, Clements M, Liu XR, et al. Capturing simple and complex time-dependent effects using flexible parametric survival models: A simulation study. Communications in Statistics - Simulation and Computation. 2021 Nov 2;50(11):3777–93.

112. Ferrer L, Rondeau V, Dignam JJ, Pickles T, Jacqmin-Gadda H, Proust-Lima C. Joint modelling of longitudinal and multi-state processes: application to clinical progressions in prostate cancer. Stat Med. 2016 Sep 30;35(22):3933–48.

113. Atici E, Erdemir AD. Ethics in a scientific approach: the importance of the biostatistician in research ethics committees. J Med Ethics. 2008 Apr;34(4):297–300.

114. Beyer U, Dejardin D, Meller M, Rufibach K, Burger HU. A multistate model for early decision-making in oncology. Biom J. 2020 Dec;62(3):550–67.

115. Tancredi A. Approximate Bayesian inference for discretely observed continuous-time multi-state models. Biometrics. 2019;75(3):966–77.

116. Meller M, Beyersmann J, Rufibach K. Joint modeling of progression-free and overall survival and computation of correlation measures. Statistics in Medicine. 2019;38(22):4270–89.