Assessment of calibration in multistate models

Rosetta, Lara

Assessment of calibration in multistate models

Lara Rosetta

MSc Applied Mathematics

26 February 2026

Supervisors

Prof. dr. M. Fiocco
Dr. M. Spreafico

Leiden University
Mathematical Institute

Contents

ABSTRACT

Assessment of calibration is a well known topic in statistics; however, established methodologies for multistate models used for risk prediction are still lacking. In this thesis, methods for assessing the calibration of predicted transition probabilities from a multistate model are discussed, with a particular focus on approaches that produce calibration plots.

The existing literature proposes several approaches to this problem, often focusing on specific types of multistate models, such as competing risks models. Previous studies have provided promising insights into methods based on pseudo-values based on the Aalen-Johansen estimator, binary logistic regression with inverse probability of censoring weights (BLR-IPCW) and multinomial logistic regression with inverse probability of censoring weights (MLR-IPCW).

The main contribution of this research is a systematic evaluation of these methodologies, originally developed for large samples, when applied to smaller sample settings commonly observed in medical research.

To evaluate the performance of each method in estimating calibration curves for predicted transition probabilities, both bias and variability of the calibration methods were investigated using a simulation study. Data were generated using a clock-forward approach under increasing levels of association between the outcome and the censoring mechanism, assuming conditional independence given the predictor variables. The performed study provided practical guidance on minimal sample size requirements, the choice of evaluation time, and computational considerations that are critical for obtaining reliable calibration estimates in multistate settings.

As an illustration of real data application, a multistate model for Ewing sarcoma data was estimated, and the calibration of the model in predicting transition probabilities between states was assessed. This further highlighted the challenges associated with sparse transitions in higher-order states and show how model simplification may be necessary to enable

calibration in small datasets.

Chapter 1

Introduction

In medical statistics, multistate models are used in situations involving diseases with multiple stages of development or complex scenarios in which a patient experiences different types of illnesses and clinical events. Being able to incorporate all this information into a single model provides useful insights into how certain events influence the risk of experiencing others. For example, in studies that analyze tumor progression, local recurrence or the appearance of metastasis can impact patient's survival.

The importance of these models lies in their ability to predict the development of similar cases and provide guidance on medical decision-making. For this reason, evaluating the performance of a model is essential to determine how reliable its predictions are when applied to new data not used in the development of the model.

Calibration is a tool commonly used for this purpose, but it is still not fully developed for multistate models. The existing literature suggested methodologies for specific types of multistate, such as competing risks and the illness-death model; however, the work of Pate *et al*[1] was the first to discuss different methods to assess calibration in a more general multistate models, where the only assumption is that all patients enter the study in the same state.

The article proposed four methods to assess either mean calibration or moderate calibration (discussed in Chapter 4). The Aalen-Johansen (AJ) method is specific for mean calibration and the pseudo-values method is used for moderate calibration, while binary logistic recalibration with inverse probability of censoring weights (BLR-IPCW) and multinomial logistic regression with inverse probability of censoring weights (MLR-IPCW) provide results for both calibration types.

Many aspects of calibration in multistate models still require further investigation, even within the models and settings presented by Pate *et al*[1]. One of these is testing the limits of calibration in small samples, as medical studies often include only a few hundreds of individuals, rather than the thousands considered in the literature.

The limitations of calibration assessment in small samples are the focus of this thesis. The problem is examined through a simulation study (Chapter 5) and a real-data example (Chapter 6) based on the Euro Ewing 2012 trial[2]. The performance of AJ, pseudo-values, BLR-IPCW, and MLR-IPCW methods is compared in both chapters. The simulation study considers progressively smaller sample sizes, eventually reaching a size at which the functions of the `calibmsm`[3] package, developed by Pate *et al*, no longer converge.

Chapter 2

Survival analysis

A statistical analysis that aims to study and model the time elapsed between a starting time (when the study starts) and the occurrence of an event is called Survival Analysis. It is largely used in medical studies as it can model events such as the recurrence or development of an illness.

In this chapter the basic quantities used in survival analysis are introduced.

2.1 Fundamental notions

2.1.1 Survival function

Let T be a non-negative random variable that represents the elapsed time between the start of the study and the observation of the event of interest.

Definition 2.1 (Survival Function)**.** *If T is continuous, $f(u)$ can be considered as the approximate probability that the event will occur at time u and the probability of an individual surviving beyond time t is defined as*

$$S(t) = \mathbb{P}(T > t) = \int_t^\infty f(u)du.$$

Assuming that event time and censoring time are independent, a non parametric estimate of the survival function is given by the Kaplan-Meier estimator

$$\widehat{S}_{KM}(t) = \prod_{s \leq t} \left(1 - \frac{\Delta \overline{N}(s)}{Y(s)}\right)$$

where $\Delta \overline{N}(s)$ is the number of events at time s and $Y(s)$ is the number of subjects at risk at s.

If T is a discrete random variable, the survival function is defined as

$$S(t) = \mathbb{P}(T > t) = \sum_{t_j > t} p(t_j)$$

where T can take the value t_j, $j = 1, 2, ...$, $t_1 < t_2 < ...$ and $p(t_j) = \mathbb{P}(T = t_j)$.

2.1.2 Hazard function

Definition 2.2 (Hazard Function)**.** *The hazard function is a non-negative function defined as*

$$h(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T < t + \Delta t | T \geq t)}{\Delta t}.$$

The hazard function represents the instantaneous risk of the event happening right after time t, conditional on survival up to time t.

When T is continuous, the cumulative hazard function is defined as

$$H(t) = \int_0^t h(u)du = -\ln[S(t)].$$

This implies

$$S(t) = e^{-H(T)}.$$

When T is discrete, the cumulative hazard function is defined as

$$H(t) = \sum_{t_j \leq t} h(t_j).$$

An estimate of the cumulative hazard function is given by the Nelson-Aleen estimator:

$$\widehat{H}_{NA}(t) = \sum_{s \leq t} \frac{\Delta \overline{N}(s)}{Y(s)}$$

which can also be written as

$$\widehat{H}_{NA}(t) = \int_0^t \frac{d\Delta \overline{N}(s)}{Y(s)}.$$

If the sample is large, there is very little difference between $\widehat{H}_{NA}(t)$ and $\widehat{H}_{KM}(t) = -ln(\widehat{S}_{KM}(t))$.[4]

2.2 Likelihood

Before introducing the notion of likelihood in survival analysis, we must consider an important phenomenon that frequently appears in medical studies. We call an observation *right-censored* when we have information about the lifetime of the individual only before a certain time. In this case we only know that the subject did not experience the event before that time. If we only know that the individual experienced the event before the start of the study, the observation is *left-censored*, while if we only know that the event happened in a certain interval, the observation is *interval-censored*.

Right-censoring is the most common occurrence. A simple example is a study with a fixed end date. In this case all the individuals are observed until the end of the study and no information is collected after that.

If we assume the censoring mechanism and the time to event to be independent, the likelihood

function that takes into consideration right, left and interval censoring observations is as follows:

$$L \propto \prod_{i \in D} f(x_i) \prod_{i \in R} S(C_r) \prod_{i \in L} (1 - S(C_l)) \prod_{i \in I} [S(L_i) - S(R_i)]$$

where

- X: *lifetimes*;
- C_r: *right censoring*;
- C_l: *left censoring*;
- D: *deaths*;
- R: *right-censored observations*;
- L: *left-censored observations*;
- I: *interval-censored observations*;

Restricting the model to a right-censoring-only scenario, we can write the likelihood as

$$L = \prod_{i=1}^{n} \mathbb{P}(t_i, \delta_i) = \prod_{i=1}^{n} (f(t_i))^{\delta_i} (S(t_i))^{1-\delta_i}$$

where

- $\delta = \mathbb{1}(X \leq C_r)$: binary indicator variable indicating whether the lifetime is observed or not;
- $T = min(X, C_r)$.

It is easy to show that

$$L = \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}.$$

Since the equation $f(t_i) = h(t_i)S(t_i)$ holds, it is possible to write the likelihood as

$$L = \prod_{i=1}^{n} h(t_i)^{\delta_i} e^{-H(t_i)}.$$

2.3 Proportional hazard Cox model

To quantify the effect of covariates on the hazard, the semi-parametric Cox model can be used; it is defined as

$$h(t|Z) = h_0(t)e^{\beta^T Z}$$

where $h_0(t)$ is the non-negative baseline hazard, Z is the vector of covariates known at baseline, and β is the column vector of the regression coefficients. The baseline hazard $h_0(t)$ is the hazard obtained when the covariates assume the reference values.

The label semi-parametric comes from the fact that the baseline hazard is studied non-parametrically, while the parametric part refers to the covariates. It is also called a proportional hazard (PH) model, since the ratio of the hazards of two individuals does not depend on time:

$$\frac{h(t|Z_1)}{h(t|Z_2)} = \frac{h_0(t)e^{\beta^T Z_1}}{h_0(t)e^{\beta^T Z_2}} = e^{\beta^T (Z_1 - Z_2)}.$$

The Cox model allows us to investigate the effect of the covariates on the survival.

If we are interested in studying not only baseline variables, but also covariates $Z(t)$ that are time-dependent, then the model is written as

$$h(t|Z(t)) = h_0(t)e^{\beta^T Z(t)}.$$

2.3.1 Likelihood in a Cox model

The regression coefficients β can be estimated by maximizing the partial likelihood function that depends on the parameters β. There are different methods to obtain the likelihood function and they depend on the assumptions we are willing to make and on the data that we are taking into consideration.

The first way to obtain a partial likelihood is to assume that only one individual at time can experience the event. In this case, considering $\mathcal{R}(t) = \{i|T_i > t\}$ to be the set of individuals at risk at time t, we can write the contribution of an individual to the likelihood as

$$L_j(\beta) = \frac{h_0(t_j)e^{\beta^T Z_j}}{\sum_{l\in\mathcal{R}(t_j)} h_0(t_j)e^{\beta^T Z_l}} = \frac{e^{\beta^T Z_j}}{\sum_{l\in\mathcal{R}(t_j)} e^{\beta^T Z_l}}.$$

Then the partial likelihood equation is

$$L(\beta) = \prod_{i=1}^{n} \left(\frac{e^{\beta^T Z_i}}{\sum_{l\in\mathcal{R}(T_i)} e^{\beta^T Z_l}} \right)^{\delta_i}.$$

Once the estimation of the coefficients $\hat{\beta}$ is obtained, using the Breslow's estimator of the baseline cumulative hazard it is possible to obtain

$$\widehat{H_0(t)} = \sum_{t_j<t} \frac{1}{\sum_{l\in\mathcal{R}(t_j)} e^{\hat{\beta}^T Z_l}}.$$

If we do not want to rely on the previous assumption, we can employ a likelihood that considers the possibility of multiple individuals experiencing the event at the same time.

The Breslow's likelihood satisfies such requirements. Let m be the number of distinct observed event times $t_1, ..., t_M$ and d_i the number of events happening at time t_i. The partial likelihood is defined as

$$L(\beta) = \prod_{i=1}^{m} \frac{e^{\sum_{j=1}^{d_i} Z_j \beta}}{\left(\sum_{l \in \mathcal{R}(t_i)} e^{\beta Z_l}\right)}.$$

It is also possible to use a different approach and calculate separate likelihoods in m subgroups of the sample. In this case we maximize the likelihood of each stratum $k = 1, ..., m$

$$L_k(\beta) = \prod_{i=1}^{n} \left(\frac{e^{\beta^T Z_i}}{\sum_{l \in \mathcal{R}_k(t_i)} e^{\beta^T Z_l}} \right)^{\delta_i}$$

where $R_k(t_i)$ are the individual of the stratum k at risk at time t_i. Thus, the partial likelihood of the sample is

$$L(\beta) = \prod_{k=1}^{m} L_k(\beta)$$

2.3.2 Testing the proportional hazard assumption

The proportional hazards (PH) assumption in a Cox model can be tested using several approaches. The two most common are Time-Dependent Covariates Test and Schoenfeld residual–based tests.

To test whether the PH assumption for a specific time-fixed covariate Z_1 holds, an interaction term between the covariate and a function of time $g(t)$ (e.g., log(time)) can be incorporated in the Cox model

$$h(t|Z_1) = h_0(t) e^{\beta_1 Z_1 + \beta_2 Z_2(t)} = h_0(t) e^{\beta_1 Z_1 + \beta_2 (Z_1 g(t))}$$

where $Z_2 = Z_1 \times g(t)$. This model can be tested for the proportional hazard assumption using a null hypothesis $\beta_2 = 0$. If the coefficient for $Z_2(t)$ is significant, then the PH assumption is violated. If the hazard rates of two individuals at time t are compared, the hazards are

proportional only if β_2 is equal to 0.

$$\frac{h(t|Z_1)}{h(t|Z_1^*)} = e^{\beta_1(Z_1 - Z_1^*) + \beta_2 g(t)(Z_1 - Z_1^*)}$$

Under the PH assumption, scaled Schoenfeld residuals should show no association with time. A formal test (e.g., Grambsch–Therneau test[5]) assesses whether residuals vary systematically over time.

2.3.3 Stratified proportional hazard model

If a covariate violates the PH assumption, a stratified Cox model can be estimated, allowing different baseline hazards across strata as follows:

$$h_j(t|Z(t)) = h_{0j}(t)e^{\beta^T Z(t)}, \qquad j = 1, ..., J$$

where j is the stratum, $h_{0j}(t)$ is the baseline hazard of stratum j, $Z(t)$ is the vector of (time-dependent) covariates, and β is the column vector of regression coefficients.

Under the null hypothesis the coefficients do not change in different strata.

Chapter 3

Multistate models

A multistate process is a stochastic process $X(t)$ evolving in continuous time $t \in [0, \tau]$, with $\tau \leq +\infty$, and taking values in a finite state space S.

Multistate models allow us to investigate the probability of transitioning from a state to another; moreover the hazard function of the basic one-event model has corresponding values for each possible transition in the multistate model.

Definition 3.1 (Transition probability)**.** *Given states $i, j \in S$, times $t, s \in \mathbb{R}^+$ with $s \leq t$, and the history at time s denoted as H_{s^-}, the transition probability from i at time s to j at time t is defined as*

$$p_{i,j}(s,t) = \mathbb{P}(X(t) = j | X(s) = i, \mathrm{H}_{s^-}).$$

The matrix whose (i,j)-element is $p_{i,j}(s,t)$ is called the transition probability matrix $\mathbf{P}(s,t)$.

Definition 3.2 (Transition intensity)**.** *Given states $i, j \in S$, the transition intensity at time t is defined as*

$$h_{ij}(t) = \lim_{\Delta t \to 0} \frac{p_{ij}(t, t+\Delta t)}{\Delta t}.$$

The matrix whose (i,j)-element is $h_{ij}(t)$ is the transition intensity matrix $\mathbf{h}(\mathrm{t})$. The diagonal

elements of this matrix are defined as $h_{ii}(t) = -\sum_{i \neq j} h_{ij}$.

Definition 3.3 (Cumulative hazard)**.** *Given states $i, j \in S$, the cumulative hazard for the transition from i to j is defined as*

$$H_{ij}(t) = \int_0^t h_{ij}(u)du.$$

The matrix whose (i, j)-element is $H_{ij}(t)$ is the cumulative transition intensity matrix $\mathbf{H}$(t).

Definition 3.4 (Total hazard)**.** *The total hazard out of state $i \in S$, denoted by $h_{i\bullet}(t)$, is defined as*

$$h_{i\bullet}(t) = \sum_{i \neq j \in S} h_{ij}(t).$$

The *total hazard* represents the probability of leaving a state.

Definition 3.5 (Total cumulative hazard)**.** *The total cumulative hazard out of state $i \in S$, denoted by $H_{i\bullet}(t)$, is defined as*

$$H_{i\bullet}(t) = \int_0^t h_{i\bullet}(u)du.$$

Definition 3.6 (State occupation probability)**.** *The state occupation probability is the probability that the process is in a given state $j \in S$ at time t:*

$$\pi_j(t) = \mathbb{P}(X(t) = j) = \sum_{i \in S} \pi_i(0) p_{ij}(0, t).$$

4

3.1 Markov assumption

Definition 3.7 (Markov assumption)**.** *If* $\forall i, j \in S, \forall t, s \in \mathbb{R}^+$ *with* $s \leq t$

$$p_{ij}(t,s) = \mathbb{P}(X(t) = j | X(s) = i, \mathcal{H}_{s^-}) = \mathbb{P}(X(t) = j | X(s) = i),$$

then the Markov assumption is satisfied. This means that the transition probability of the process depends only on the current state i *and the current time* s *but not on the previous history.*

When a model satisfies the Markov assumption, it is called Markovian model.

3.2 Counting time

There are two approaches to counting time in multistates models: *clock forward* and *clock reset*. If the *clock forward* method is applied, then the time is measured from t_0, when the process is in the starting state. In contrast, under the *clock reset* method, the time is reset to zero as soon as the process enters a new state. The *clock reset* method implies that the conditions of the Markov property are met.

3.3 Inference on multistate models

In a similar fashion as before, maximizing the likelihood provides estimates of the effect of covariates on transitions.

Let I be the set of m individuals and n_i the number of states visited by an individual $i \in I$, s_{ij} the j^{th} state entered by individual i at time t_{ij} time. The contribution to the likelihood of a single transition for individual i is

$$h_{s_{ij}, s_{i,j+1}}(t_{i,j+1}) \exp\{-(H_{s_{ij}}(t_{i,j+1}) - H_{s_{ij}}(t_{ij}))\}.$$

Let $l \in S$ represent the state the individual can possibly enter next. The contribution of state s_{ij} with respect to individual i is

$$\prod_{s_{ij} \to l \in S} h_{s_{ij},l}(t_{i,j+1})^{\mathbb{1}(s_{i,j+1}=l)} \exp\{-(H_{s_{ij}}(t_{i,j+1}) - H_{s_{ij}}(t_{ij}))\}.$$

Considering all individuals and states in the model yields the formula of the general likelihood.

If the final state is an absorbing state, then the general likelihood is defined as

$$\prod_{s_{i0} \to l \in S} h_{s_{i0},l}(t_{i1})^{\mathbb{1}(s_{i1}=l)} \exp\{-(H_{s_{i0},\bullet}(t_{i1}) - H_{s_{i0},\bullet}(t_{i0}))\} \times$$

$$\prod_{s_{i1} \to l \in S} h_{s_{i1},l}(t_{i,2})^{\mathbb{1}(s_{i2}=l)} \exp\{-(H_{s_{i1},\bullet}(t_{i2}) - H_{s_{i1},\bullet}(t_{i1}))\} \times$$

$$\times \cdots \times$$

$$\prod_{s_{i,n_i-1} \to l \in S} h_{s_{i,n_i-1},l}(t_{i,n_i})^{\mathbb{1}(s_{i,n_i}=l)} \exp\{-(H_{s_{in_i-1},\bullet}(t_{i,n_i}) - H_{s_{in_i-1},\bullet}(t_{in_i-1}))\}.$$

If the final state is transient, then the general likelihood includes a term that takes into account the right-censored observations and their censoring time c_i, as follows:

$$\prod_{s_{i0} \to l \in S} h_{s_{i0},l}(t_{i1})^{\mathbb{1}(s_{i1}=l)} \exp\{-(H_{s_{i0},\bullet}(t_{i1}) - H_{s_{i0},\bullet}(t_{i0}))\} \times$$

$$\prod_{s_{i1} \to l \in S} h_{s_{i1},l}(t_{i,2})^{\mathbb{1}(s_{i,2}=l)} \exp\{-(H_{s_{i1},\bullet}(t_{i2}) - H_{s_{i1},\bullet}(t_{i1}))\} \times$$

$$\times \cdots \times$$

$$\prod_{s_{i,n_i-1} \to l \in S} h_{s_{i,n_i-1},l}(t_{i,n_i})^{\mathbb{1}(s_{i,n_i}=l)} \exp\{-(H_{s_{in_i-1},\bullet}(t_{i,n_i}) - H_{s_{in_i-1},\bullet}(t_{in_i-1}))\} \times$$

$$\exp\{-(H_{s_{in_i-1},\bullet}(c_i) - H_{s_{in_i-1},\bullet}(t_{in_i}))\}.$$

If we are interested in estimating transition probabilities, assuming that a model is Markovian, it is possible to obtain the transition probabilities using the Chapman-Kolmogorov equations and the Kolmogorov forward equation.

Definition 3.8 (Chapman-Kolmogorov equations)**.** *Let $i, j, l \in S$ be states and $s, u, t \in \mathbb{R}^+$ denote times with $s \leq u \leq t$. The Chapman-Kolmogorov equation for transition probability from i to j is given by:*

$$p_{ij}(s,t) = \sum_{l \in S} p_{il}(s,u) p_{lj}(u,t).$$

From the Chapman-Kolmogorov equations it follows that the transition probability matrix can be written as:

$$\mathbf{P}(s,t) = \mathbf{P}(s,u)\mathbf{P}(u,t).$$

Then, considering the Kolmogorov forward equation, defined as

Definition 3.9 (Kolmogorov forward equation)**.**

$$\frac{\partial}{\partial t}\boldsymbol{P}(s,t) = \boldsymbol{P}(s,t)\boldsymbol{h}(t),$$

the following formula of the transition probability matrix can be obtained

$$\mathbf{P}(s,t) = \prod_{u\in(s,t]} \{\mathbf{I} + d\mathbf{H}(u)\}.$$

The cumulative hazard using the Nelson-Aalen estimator is given by

$$\widehat{H}_{ij}(t) = \int_0^t \mathbb{1}(Y_i(u) > 0)\frac{dN_{ij}(u)}{Y_i(u)}$$

where $Y_i(t)$ is the number of individuals in state i at time $t-$, $N_{ij}(t)$ is the number of transitions $i \to j$ observed in $[0,t]$, and Y_i is the number of subjects at risk at t_i.

3.4 Well-known models

3.4.1 Competing risks model

A competing risk model in the simpliest form of multistate models which presents an initial state and several absorbing states, as shown in Figure 3.1. Entering one of the absorbing states implies that the individual will never experience the others. For this reason we talk about competing risks.

3.4.2 Illness-death model

A simple and well-known multistate model is the progressive illness-death model shown in Figure 3.2. Let $S = \{1,2,3\}$ be the state space and $1 \to 2$, $1 \to 3$, $2 \to 3$ the possible transitions. State 1 (Health) is the starting state, state 2 (Diseased) is a transient state and state 3 (Death) is an absorbing state. In this model, the only way a patient may not reach the third state is through censoring.

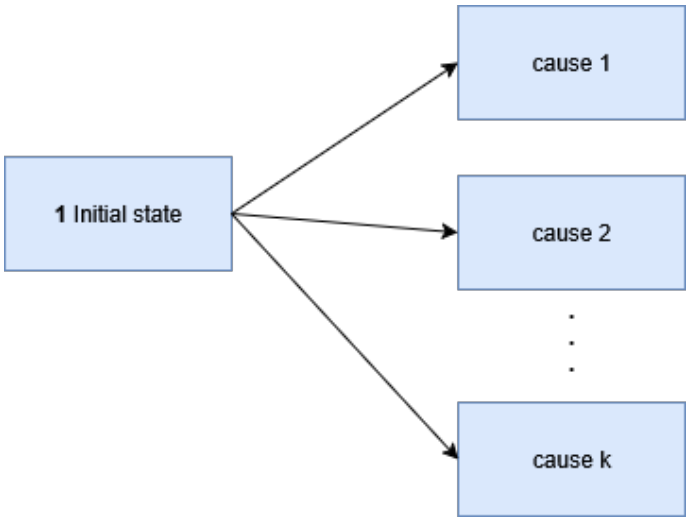


Figure 3.1: Competing risk model with k absorbing states.

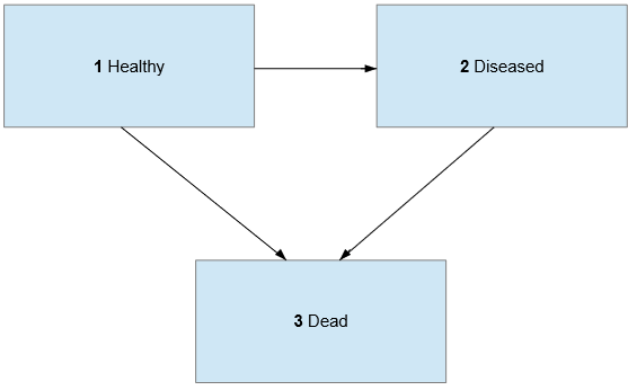


Figure 3.2: Illness-death model with three states: 1) starting (Health), 2) transient (Diseased), and 3) absorbing (Death).

Chapter 4

Assessment of calibration

4.1 Calibration

Calibration of a model refers to the agreement between predicted and observed outcomes. *Internal* calibration represents the case where both predicted and observed outcomes are obtained from the same dataset used to develop the model, whereas *external* calibration employs datasets that are different from the one on which the model was developed.

Lack of internal calibration is associated with issues of poor model fit and misspecification of the fitted model, whereas lack of external calibration is usually related to overfitting.[6] Typically, a calibration plot, which shows observed outcomes against predicted values, is used to evaluate the calibration of a fitted model.[7] Intuitively, when the same dataset is used to estimate the parameters of the model and obtain the calibration plot, the calibration slope is expected to be equal to one. When calibration is assessed using a different dataset, the slope may be different from one[8]. A slope smaller than one indicates overfitting.

Although statistical tests are also widely employed to evaluate calibration, the primary focus of this thesis is on graphical assessment.

4.1.1 Four levels of calibration

There are four forms of calibration corresponding to increasingly strict levels: mean calibration (calibration-in-the-large), weak calibration, moderate calibration and strong calibration.[9,10]

1) Mean calibration

Mean calibration evaluates the agreement between the average predicted risk compared to the observed event rate. It is uninformative in the context of internal validation methods, such as bootstrapping.

2) Weak calibration

Weak calibration considers only the average prediction effect and is assessed using the calibration slope and intercept: a slope value equal to 1 and an intercept value of zero imply that on average the predicted risk is neither systematically overestimated nor underestimated.

3) Moderate calibration

Moderate calibration assesses whether the estimated risks correspond to observed event proportions. It can be investigated using flexible calibration curves, which ideally are close to the diagonal line representing perfect agreement.

4) Strong calibration

Strong calibration requires that "the predicted risk corresponds to the observed proportion for every possible combination of predictor values". This level of calibration is rarely used, as it imposed strict conditions that are hard to satisfy in practice.

4.1.2 Calibration in different types of outcomes

The assessment of calibration is a well-established methodology in the literature for different types of outcomes, i.e., continuous, binary, polytomous, and survival outcomes; however, corresponding methodology for multistate models is still under development.

For binary outcomes, the most common approach is to use smoothing techniques to estimate observed probabilities of the outcome, which are then compared to predicted probabilities. The Locally Estimated Scatterplot Smoothing (LOESS) algorithm is often chosen for this purpose. It is a non-parametric procedure based on local weighted regressions that produces a smooth curve from a scatter plot. This method is considered more flexible, although parametric logistic recalibration has also been shown to provide good performance.[11]

An alternative strategy is based on the Hosmer-Lemeshow (H-L) test, which involves stratifying the sample according to the predicted probabilities of the outcome and, within each stratum, calculating both the mean predicted probability and the empirical estimated probability, which are then compared.

Similarly to the binary case, calibration for polytomous outcomes can be assessed using Multinomial Logistic Regression (MLR), which is particularly useful to analyze calibration slopes and intercepts, as well as through non-parametric methods.[12]

Complications arise when working with survival (or time-to-event) outcomes, as calibration evaluates the agreement between predicted and observed risks at specific points in time. When multiple time points are deemed relevant to the analysis, calibration must be evaluated separately for each time point. Moreover, the observed data consist of the time-to-event, rather than the probability of an outcome. As in the binary setting, the two most common approaches are stratification and employment of smoothed calibration curves.[13]

As discussed above, the methodology for calibration in multistate models is still in development. In 2024 Pate *et al*[1] evaluated several methods for assessing calibration in multistate

models with particular attention to calibration plots. These methods include the Aalen-Johansen estimator, pseudo-values, binary logistic recalibration using inverse probability of censoring weights (BLR-IPCW), and a nominal recalibration framework using multinomial logistic regression with inverse probability of censoring weights (MLR-IPCW). The article reported better performances of the latter three methods, which will be the focus of this thesis.

4.2 Methods

The methods presented in this section focus on cohorts in which all the individuals start in the same state (state $i = 1$ at time $s = 0$) and are limited to assessing calibration for the transition probabilities from that initial state to state k at time t, that is $p_{1k}(0, t)$. The notation is hence simplified to $p_k(t)$.

4.2.1 Aalen-Johansen estimator

The Aalen-Johansen estimator, or empirical transition matrix, is the generalization of the Kaplan-Meier estimator in multistate models.

Taking into account the Nelson-Aalen estimator of the cumulative hazard risk $\widehat{H}_{ij}(t)$, with $i \neq j$, it is possible to obtain the estimator $\widehat{H}_{ii}(t) = -\sum_{i \neq j} \widehat{H}_{ij}(t)$. Let $\widehat{\mathbf{H}}(t)$ be the matrix of dimension $(k+1) \times (k+1)$, where k is the number of states of the model.[14] The transition probability matrix can then be estimated using the Aalen-Johansen estimator, written as a matrix product integral $\widehat{\mathbf{P}}(s, t) = \prod_{(s,t]} \{\mathbf{I} + d\widehat{\mathbf{H}}(u)\}$.[15]

The Aalen-Johansen estimator provides estimates of the observed risk of state k, denoted by obs_k^{AJ}, allowing the evaluation of the *mean* calibration of the transition probabilities in a

cohort of n individuals as follows:

$$\frac{1}{n}\left(\sum_{i=1}^{n} obs_k^{AJ} - \hat{p}_k^i(t)\right)$$

where $\hat{p}_k^i(t)$ is the predicted transition probability to state k for individual i.

Aalen-Johansen estimator relies on the assumption of random censoring. If this condition is not met, the estimator needs to be calculated within subgroups where the independence between censoring mechanism and survival times holds. These subgroups are defined by covariates or by predicted transition probabilities of each state k. In the latter case, the cohort is divided into equal sized subgroups based on the predicted transition probabilities $\hat{p}_k^i(t)$ for each state k. For each state, mean calibration is assessed by subtracting the mean predicted transition probability from the subgroup-specific average of the Aalen–Johansen estimator.

This method works because individuals with similar predicted probabilities are more comparable, making censored subjects more representative of the uncensored ones within each group. It is particularly useful when many covariates influence both the outcome and the censoring mechanism, making the alternative subgrouping approach impractical.

The Aalen-Johansen estimator is a consistent estimator only under the Markov assumption and, in the particular case where every subject starts in the same initial state, it is consistent even in non-Markov models. Datta and Satten[16] and Glidden[17] showed that, in non-Markov models, the Aalen-Johansen estimator is a consistent estimator of the state occupation probability $\pi_k(t) = \mathbb{P}(X(t) = k) = \sum_{j \in S} \pi_j(0) p_{jk}(0, t)$ for each $k \in S$, where S is the set of states in the model.

When all subjects start in the initial state 1 at time 0, $\pi_k(0) = 0 \ \forall k > 1$ and $\pi_1(0) = 1$. Consequently, the state occupational probabilities correspond to the vector of transition probabilities out of state 1. Then the transition probabilities considered in this thesis corre-

spond to the occupation probabilities $\pi_k(t) = p_{1k}(0,t) = p_k(t)$.

If all individuals begin in the same initial state, and interest only is in estimating transition probabilities from that state, the Aalen-Johansen estimator is sufficient for assessing calibration.

4.2.2 Pseudo-values

In survival analysis the presence of censoring introduces the challenge of handling incomplete data.[18] When considering the parameter $\theta = E(f(X))$, where $f(X_i)$ is a function of the survival time, it is evident that it would not be possible to calculate estimators like $\hat{\theta} = \frac{1}{n}\sum_i f(X_i)$, since $f(X_i)$ might not be observed for some i. Instead, a solution is represented by pseudo-values, which can be considered the contribution of a subject to the expectation. A pseudo-value, defined as

$$\hat{\theta}^i = n * \hat{\theta} - (n-1) * \hat{\theta}^{-i}$$

where $\hat{\theta}^{-i}$ is a well-behaved estimator $\hat{\theta}$ calculated in the cohort with individual i removed, can be used as a substitute for the possibly incomplete observation $f(X_i)$. It is important to notice that this approach is valid if $\hat{\theta}^{-i}$ is used for every $i = 1, ..., n$.

In this thesis, the Aalen-Johansen estimator is used to obtain the pseudo-value $\hat{\theta}_k^i(t)$ of individual i at time t for the transition to state k. Once obtained $\hat{\theta}_k^i(t)$, it is then possible to assess calibration. Specifically:

1. *mean* calibration can be calculated as follows:

$$\frac{1}{n}\left(\sum_{i=1}^{n} \hat{\theta}_k^i(t) - \hat{p}_k^i(t)\right);$$

2. the slope $\hat{\beta}$ obtained by fitting the regression model

$$\hat{\theta}_k^i(t) = \alpha + \beta * \hat{p}_k^i(t)$$

is used to assess *weak* calibration;

3. a logistic regression model with restricted cubic splines smoother is employed to assess *moderate* calibration:

$$\text{logit}(\hat{\theta}_k^i(t)) = \text{rcs}(\text{logit}[\hat{p}_k^i(t)]). \tag{4.1}$$

After fitting model (4.1) in the validation cohort, the observed probabilities $\widehat{obs}_k^{PV,i}$, derived from the fitted values, can be used to construct the calibration plot $\left\{\hat{p}_k^i(t), \widehat{obs}_k^{PV,i}(t)\right\}$ for transition to state k.

Since the Aalen-Johanes estimator is derived under the assumption of independent censoring, violations of this assumption requires the identification of subgroups where censoring is independent and the estimator remains valid. Two approaches may be considered: (i) when the censoring mechanism and the outcomes are conditionally independent given the covariates, subgroups may be defined based on the covariates; otherwise, (ii) individuals can be ordered by $\hat{p}_k^i(t)$ and then divided into subgroups.

4.2.3 Binary logistic recalibration using inverse probability of censoring weights (BLR-IPCW)

Binary logistic regression (BLR) is often used to assess calibration of binary outcomes and can be adapted to the multistate setting through an indicator variable $I_k(t)$ which equals 1 if an individual occupies state k at time t and 0 otherwise.[1] Given the estimated occupation probabilities $p_k^i(t)$, BLR can be applied to a validation cohort via a logistic regression model

with restricted cubic splines with four knots:

$$\text{logit}(I_k(t)) = \text{rcs}(\text{logit}[\hat{p}_k^i(t)]).$$

This provides the fitted values $\widehat{obs}_k^{BLR,i}(t)$, analogous to the pseudo-value method, allowing the construction of the calibration plots $\left\{\hat{p}_k^i(t), \widehat{obs}_k^{BLR,i}(t)\right\}$.

For uncensored data, the observed probabilities $\mathbb{P}(X(t) = k \mid \hat{p}_k(t))$ can be estimated directly. In the presence of censoring, calibration may be biased if the censoring is not independent. This can be corrected using inverse probability of censoring weights

$$w_t = \frac{1}{G(\min(t, t_{abs}), Z)},$$

where Z denotes baseline covariates, t_{abs} is the time until entry into an absorbing state, and $G(t, Z)$ is the censoring distribution, estimated from the time to censoring (with censoring not possible after entry into an absorbing state).

Calibration plots are used to assess *moderate* calibration, while *mean* and *weak* calibration are evaluated using the following regression model:

$$\text{logit}[I_k(t)] = \alpha + \beta * \text{logit}[\hat{p}_k(t)]. \tag{4.2}$$

Specifically, *weak* calibration is assessed by estimating $\hat{\beta}$ from model (4.2) without imposing any constraints on the calibration slope, whereas *mean* calibration is evaluated by fixing $\beta = 1$.

When β is fixed to 1, the corresponding fitted values are

$$\widehat{obs}_k^{BLR-mean,i}(t) = \text{logit}^{-1}\left(\hat{\alpha} + \text{logit}[\hat{p}_k^i(t)]\right)$$

and the *mean* calibration is calculated as follows:

$$\frac{1}{n}\left(\sum_{i=1}^{n} \widehat{obs}_k^{BLR-mean,i}(t) - \hat{p}_k^i(t)\right).$$

4.2.4 Nominal recalibration framework using multinomial logistic regression and inverse probability of censoring weights (MLR-IPCW)

Polytomous logistic regression allows the fitting of submodels that compare each outcome category with a chosen reference category. This approach can be extended to the multistate setting by considering the polytomous variable $I_X(t) = k$, $k \in \{1, ..., K\}$, which indicates the state occupied by an individual at time t.[1]

To model the outcome $I_X(t)$, a multinomial logistic regression with vector spline smoother is used:

$$\ln\left[\frac{\mathbb{P}(I_X(t) = k)}{\mathbb{P}(I_X(t) = 1)}\right] = \alpha_k + \sum_{h=2}^{K} \beta_{k,h} * s_k(\widehat{LRP}_h) \tag{4.3}$$

for $k > 1$. For each individual i in the validation cohort, the log-ratio of probabilities (LRP) for each state k relative to state 1 is defined as:

$$\widehat{LRP}_k^i = \ln\left[\frac{\hat{p}_k^i(t)}{\hat{p}_1^i(t)}\right].$$

The observed event probabilities $\widehat{obs}_k^{MLR,i}(t)$ are then obtained as:

$$\widehat{obs}_1^{MLR,i}(t) = \frac{1}{1 + \sum_{l=2}^{K} \exp\left(\hat{\alpha}_l + \sum_{h=2}^{K} \hat{\beta}_{l,h} * s_l\left(\widehat{LRP}_h\right)\right)},$$

$$\widehat{obs}_k^{MLR,i}(t) = \frac{\exp\left(\hat{\alpha}_k + \sum_{h=2}^{K} \hat{\beta}_{k,h} * s_k\left(\widehat{LRP}_h\right)\right)}{1 + \sum_{l=2}^{K} \exp\left(\hat{\alpha}_l + \sum_{h=2}^{K} \hat{\beta}_{l,h} * s_l\left(\widehat{LRP}_h\right)\right)}, \quad k > 1.$$

The calibration plot for state k is then given by the set of points $\left\{\hat{p}_k(t), \widehat{obs}_k^{MLR,i}(t)\right\}$ and can be used to assess *moderate* calibration.

Since dependence between the censoring mechanism and the outcome induces bias, the same inverse probability of censoring weighting (IPCW) strategy used for binary logistic recalibration is applied to the multinomial logistic regression.

As Pale *et al*[1] reported, the MLR-IPCW approach substantially differs from the pseudo value and the BLR-IPCW approach. First, MLR-IPCW is the only method that does not apply a "one versus all" strategy when estimating event probabilities and therefore it guarantees that the estimated probabilities sum to one for each individual. Second, calibration plots are not obtained using flexible models that provide smoothed curves, instead they take the form of scatter plots. This means that individuals sharing the same predicted transition probability for state k may differ in their predicted probabilities for other states and thus in their observed event probabilities. This provides a more informative assessment of calibration than the BLR-IPCW and pseudo-value approaches.

To assess *mean* and *weak* calibration, the following multinomial logistic regression model is estimated in the validation cohort

$$ln\left[\frac{\mathbb{P}(I_X(t) = k)}{\mathbb{P}(I_X(t) = 1)}\right] = \alpha_k + \beta_k * \widehat{LRP}_k, \tag{4.4}$$

where β_k is fixed to 1 when assessing *mean* calibration, whereas it is estimated without constraints when assessing *weak* calibration. The estimated coefficient $\hat{\alpha}_k$ is the calibration intercepts from the calibration framework of Van Hoorde *et al*[11].

To estimate the *mean* calibration on the probability scale, observed event probabilities for

each individual i in the validation cohort need to be computed by using the following expressions:

$$\widehat{obs}_1^{MLR-mean,i}(t) = \frac{1}{1 + \sum_{l=2}^{K} \exp(\hat{\alpha}_l + \widehat{LRP}_l^i)},$$

$$\widehat{obs}_k^{MLR-mean,i}(t) = \frac{\exp(\hat{\alpha}_k + \widehat{LRP}_k^i)}{1 + \sum_{l=2}^{K} \exp(\hat{\alpha}_l + \widehat{LRP}_l^i)}, \quad k > 1,$$

and *mean* calibration for state k can then be calculated as follows:

$$\frac{1}{n} \left(\sum_{i=1}^{n} \widehat{obs}_k^{MLR-mean,i}(t) - \hat{p}_k^i(t) \right).$$

As usual, *weak* calibration can be assessed by fitting model (4.4) without constraints and evaluating the corresponding calibration slope $\hat{\beta}_k$.

Chapter 5

Simulation

A simulated study on the previously discussed methods of assessing *mean* and *moderate* calibration in multistate models was conducted. Data generating methods and results of the simulation are reported in this chapter. The methods are illustrated by following the ADEMP[19] structure and reporting data generating mechanisms, estimands, methods and performance. The code used for this thesis, available on GitHub[20], was based on the work of Pate *et al*[21].

5.1 Aims

The goal of the simulation study was to evaluate the performance of each calibration assessment method under three scenarios: random censoring (RC), weakly associated censoring (WAC), and strongly associated censoring (SAC).

The work of Pate *et al*[1] only considered large ($n = 200000$) and small samples, where the latter included 3000 or 1500 observations. However, these are considerable amounts of patients and medical studies are often based on much smaller datasets. Therefore, this thesis investigates the performance of calibration assessment methods for smaller sample sizes.

5.2 Data-generating mechanisms

5.2.1 Model

Data were generated following the clock-forward method in a 5-states model (Figure 5.1) based on the multistate model proposed by Putter *et al*[22] to analyze data from the EORTC 10854 breast cancer trial. Every patient starts in state 1 after surgery and may progress towards local recurrence (state 2), distant metastasis (state 3) and death (state 5); state 4 represents the combined occurrence of local recurrence and distant metastasis.

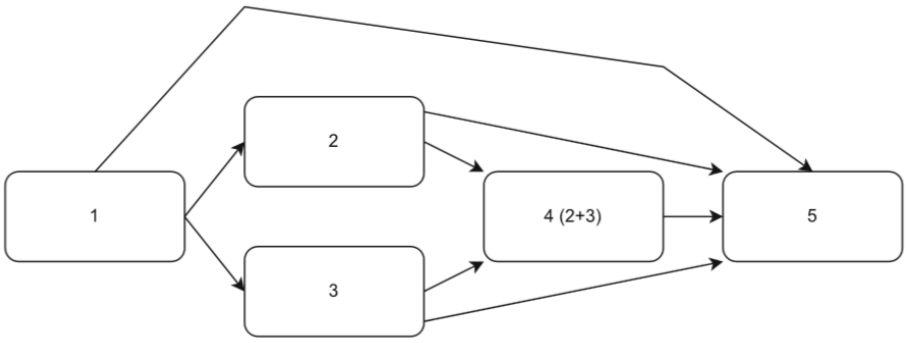


Figure 5.1: Simulated multistate model.

In the article by Putter *et al*[22], the multistate model was estimated on the breast cancer trial data and the effect of risk factors on each transition was studied, as well as the effect of the time at which intermediate events occured. The estimated survival probabilities reported in the article were used by Pate *et al*[1] to choose the evaluation period of 7 years and to generate the baseline hazards in the simulation process, as discussed below.

5.2.2 Simulation process

Due to the high computational cost required by the simulation study, an initial sample of 1000000 subjects was initially generated. For each subsequent analysis or simulation repetition, the required observations were then randomly sampled from this pre-generated dataset, rather than simulating a new dataset each time.

The first step of the simulation consisted in randomly drawing for each subject the values of eight transition-specific baseline covariates $\mathbf{Z} = (Z_{12}, Z_{13}, Z_{15}, Z_{24}, Z_{25}, Z_{34}, Z_{35}, Z_{45})$ from a normal Gaussian distribution.

The baseline hazard and survival functions for each transition $i \rightarrow j$ were simulated by assuming a Weibull distribution:

$$\lambda_{ij}(t) = \alpha_{ij}^{k_{ij}} \cdot k_{ij} \cdot t^{k_{ij}-1}$$

$$S_{ij}(t) = e^{-(\alpha_{ij} \cdot t)^{k_{ij}}}.$$

For each transition $i \rightarrow j$, the shape parameter k_{ij} was set to 1, in which case the Weibull reduces to an exponential distribution and the baseline hazard rates $\lambda_{ij}(t) = \alpha_{ij}$ remain constant over time. The rate parameters α_{ij} were calculated by considering the survival probability at 7 years, as follows:

$$s_{ij} = S_{ij}(t = 7 \text{ years})$$

$$\alpha_{ij} = \frac{-log(s_{ij})}{7 \cdot 365.25}.$$

Once the baseline hazards were obtained, the survival time for transition $i \rightarrow j$ could be calculated using an exponential distribution with hazard $\alpha_{ij} \cdot e^{0.5 \cdot Z_{ij}}$.

Following the approach adopted by Pate *et al*[1], the values $\{s_{ij}\}$ were initially chosen equal to the survival probabilities estimated at 7 years for the breast cancer trial[22] ($s_{12} = 0.90, s_{13} = 0.80, s_{15} = 0.99, s_{24} = 0.55, s_{25} = 0.95, s_{34} = 0.70, s_{35} = 0.15, s_{45} = 0.05$); however, the low number of subjects in some states did not allow the assessment of calibration in samples smaller than 1500 individuals. Therefore, alternative survival probabilities were specified ($s_{12} = 0.40, s_{13} = 0.25, s_{15} = 0.99, s_{24} = 0.55, s_{25} = 0.95, s_{34} = 0.50, s_{35} = 0.15, s_{45} = 0.20$), while maintaining the same model and evaluation time, in order to allow the investigation

of smaller sample sizes.

The true transition probabilities $p_{k,true} = p_k$ into each state k were saved to allow comparison with the estimates obtained later using different methods. The formulas for the true transition probabilities are provided in Appendix A.

5.2.3 Censoring

The next step in the simulation experiment was to build datasets under three censoring scenarios: random (RC), weakly (WAC), and strongly (SAC) associated censoring.

Each dataset was obtained by generating censoring times from an exponential distribution with hazard $\lambda_C \cdot e^{\boldsymbol{\beta}_{cens}\mathbf{Z}}$ and vector parameter $\boldsymbol{\beta}_{cens} = (\beta_{12}, \beta_{13}, \beta_{15}, \beta_{24}, \beta_{25}, \beta_{34}, \beta_{35}, \beta_{45})$. The value of $\lambda_C = 5005$ was chosen to obtain 0.4 probability of censoring under the RC scenario at 7 years with $\boldsymbol{\beta}_{cens} = 0$. The parameters $\beta_{cens} = 0.125$ and $\beta_{cens} = 0.25$ were chosen to generate a WAC and SAC scenarios, respectively.

The censoring was then added to the original dataset, obtaining three separate datasets of 1000000 subjects each.

5.2.4 Predicted transition probabilities

Finally, the last step in the data preparation consisted of deterministically generating the values of three types of predicted transition probabilities $\hat{p}_k$: one set of perfectly predicted transition probabilities, $\hat{p}_{k,true}$, equal to the true transition probabilities already calculated, and two sets of miscalibrated predicted probabilities, $\hat{p}_{k,miscal1}$ and $\hat{p}_{k,miscal2}$, obtained by adding a vector v_j to the the log-odds, then converting back to probabilities and normalizing.

The vectors v_j, with $j = 1, 2$ indicating the corresponding miscaliibrated set, were defined as

$$v_1 = (0.5, 0.25, 0, -0.25, -0.5), \qquad v_2 = (0.5, -0.5, -0.5, -0.5, 0.5)$$

and each entrance $v_{j,k}$ $(k = 1, ..., 5)$ was integrated into the log-odds

$$\hat{p}_{k,miscalj} = \frac{e^{\log\left(\frac{p_{k,true}}{1-p_{k,true}}\right)+v_{j,k}}}{1 + e^{\log\left(\frac{p_{k,true}}{1-p_{k,true}}\right)+v_{j,k}}}$$

and finally normalized as follows

$$\hat{p}_{k,miscalj} = \frac{\hat{p}_{k,miscalj}}{\hat{p}_{1,miscalj} + \hat{p}_{2,miscalj} + \hat{p}_{3,miscalj} + \hat{p}_{4,miscalj} + \hat{p}_{5,miscalj}}.$$

This strategy induced a miscalibration, leading to over predict the transition to state 1 and 2 and under predict the transition to state 4 and 5 for $\hat{p}_{k,miscal1}$. For $\hat{p}_{k,miscal2}$ the over predicted states were 1 and 5, while 2, 3 and 4 were under predicted.

5.2.5 ALICE

Part of the data-generation process and later the evaluation of mean and moderate calibration involved extremely time consuming operations. In particular, calibration was assessed through functions available in the R package `calibmsm`[3], that could not be run on a local computer in reasonable time.

The code used in this project was adapted to make use of the ALICE high-performance computing (HPC) resources provided by Leiden University. HPC enabled the parallelization of critical parts of the code, as explained in the following section.

5.2.6 Calibration in small samples

For each combination of scenario (RC, WAC, and SAC) and type of predicted probabilities ($\hat{p}_{k,true}$, $\hat{p}_{k,miscal1}$, $\hat{p}_{k,miscal2}$), the assessment of calibration was conducted on samples of size $n = 700$, 500, and 400. However, the lack of patients in states 1, 2 and 3 after 7 years allowed only the evaluation of *mean* calibration in samples of 400 subjects under the strongly associ-

ated censoring (SAC) scenario, as the simulations did not converge for *moderate* calibration. Therefore, only *mean* calibration results are reported for $n = 400$.

Mean calibration was calculated by considering 50 batches of simulations, each initialized with a different seed and consisting of 1000 simulations, resulting in a total of $B = 50 \times 1000 = 50000$ repetitions. The values of the true mean calibration, as well as the calibrations obtained using Aalen-Johansen (AJ), Binary Logistic Recalibration (BLR), and Multinomial Logistic Regression (MLR) were saved for each simulation, and the results from all batches were finally combined.

Similarly, for *moderate* calibration, the methods employed were pseudo-values, BLR and MLR. Batches of 200 simulations, instead of 1000, were performed, resulting in a total of $B = 50 \times 200 = 10000$ repetitions.

5.3 Estimands and other targets

The estimand of interest in each simulation was the calibration of the predicted transition probabilities, which compared $\{\widehat{p}^i_{k,true}, \widehat{p}^i_{k,miscal1}, \widehat{p}^i_{k,miscal2}\}$ to p^i_k.

When evaluating *mean* calibration for each state k, the estimand of interest was the average difference between the predicted transition probabilities $\widehat{p}^i_k \in \{\widehat{p}^i_{k,true}, \widehat{p}^i_{k,miscal1}, \widehat{p}^i_{k,miscal2}\}$ and the true transition probabilities

$$\frac{1}{n}\left(\sum_{i=1}^{n} p^i_k - \widehat{p}^i_k\right)$$

Moderate calibration was assessed by plotting the estimated transition probabilities against the observed event proportions, with the set of coordinates $\{\widehat{p}^i_k, p^i_k\}$ being the estimand of interest.

If the estimates are obtained using the MLR-IPCW method, they come in the form of a scatter plot $\left\{\widehat{p}^i_k, \widehat{obs}_k^{MLR,i}\right\}$, that has already the same form as the estimand. Mean-

while, pseudo-values and BLR-IPCW methods return smoothed curves. To be able to compare these with the desired estimate $\{\widehat{p}_k^i, \widehat{p}_{k,true}^i\}$, the latter needs to be transformed into a smoothed curve. To achieve this result, the true transition probabilities were regressed on the predicted transition probabilities, using restricted cubic splines with four knots $\text{logit}(\widehat{p}_{k,truesmooth}^i) = \text{rcs}(\text{logit}[\widehat{p}_k^i])$.

5.4 Methods

The assessment of calibration was conducted using the methods discussed in Chapter 4. In particular, for *mean* calibration the methods employed were Aalen-Johansen estimator, BLR-IPCW and MLR-IPCW, while in *moderate* calibration pseudo-values, BLR-IPCW and MLR-IPCW were used. Since the pseudo-values method relies on the Aalen-Johansen estimates, there was no reason to consider the pseudo values too when assessing *mean* calibration. On the other hand, the Aalen-Johansen method cannot provide calibration plots, while the pseudo-value can. For this reason, the latter was chosen to assess *moderate* calibration.

Both the Aalen-Johansen and the pseudo-values methods required the use of subgroups to satisfy the assumption of independence between the censoring mechanism and the survival times, as explained in Sections 4.2.1 and 4.2.2. For both cases 10 subgroups were chosen based on the predicted transition probabilities.

For the BLR-IPCW and MLR-IPCW methods, problems related to censoring are managed by using weights that equal the inverse probability of censoring (see Section 4.2.3). Therefore, in the WAC (weakly associated censoring) and SAC (strongly associated censoring) scenarios the probability of censoring was obtained by fitting a proportional hazard Cox model with baseline covariates Z, while in random censoring (RC) scenario no covariates were used.

5.5 Performance measures

For each state k, *mean* calibration was assessed by evaluating the bias of the estimated mean calibration under each method $j \in \{\text{AJ, BLR-IPCW, MLR-IPCW}\}$, defined as follows:

$$\text{Bias}_k^b = \frac{1}{n}\left(\sum_{i=1}^{n} \widehat{obs}_{k,j}^{i,b} - \widehat{p}_k^{i,b}\right) - \frac{1}{n}\left(\sum_{i=1}^{n} p_{k,true}^{i} - \widehat{p}_k^{i,b}\right)$$

where $\widehat{p}_k^{i,b} \in \{\widehat{p}_{k,true}^{i}, \widehat{p}_{k,miscal1}^{i}, \widehat{p}_{k,miscal2}^{i}\}$ is the predicted transition probabilities, $i = 1, \ldots, n$ is the subject index, $b = 1, \ldots, B$ is the repetition. For each sample size, graphs were produced reporting the median bias together with the 2.5 to 97.5 percentile range among repetitions under each combination of type of predicted probabilities, method, and censoring scenario.

Moderate calibration was assessed graphically, using scatter plots or smoothed curves, as explained in Section 5.3. For each sample size, censoring scenario, and type of predicted probabilities, the calibration plots $\left\{\widehat{p}_k^{i}, \widehat{obs}_{k,j}^{i,b}\right\}$ of state k, method j and repetition b were drawn on the same graph to show the variability of the estimated moderate calibration.

5.6 Results

5.6.1 Mean calibration

Figures 5.2 to 5.4 display the median bias, along with percentile ranges, obtained when estimating *mean* calibration under sample sizes $n = 400, 500, 700$, respectively. In each figure, each column refers to a different state k, and each row corresponds to a different type of predicted transition probabilities. Within each panel, percentile ranges are grouped by censoring scenario and colored according to the estimation method (blue: AJ; green: BLR-IPCW; red: MLR-IPCW).

Under the RC and WAC scenarios all methods provide good performances, with no noticeable

variation between samples of different sizes. Under the SAC scenario the bias is more evident, in particular in states 2 and 5 and for the BLR-IPCW and MLR-IPCW methods.

Thus, from these graphs, it can be concluded that the AJ method provides the best performance in estimating *mean* calibration, even under the SAC scenario. Moreover, the sample size does not appear to affect the bias of the *mean* calibration.

5.6.2 Moderate calibration

Moderate calibration can be evaluated using calibration plots as in Figures 5.5–5.22. Each figure refers to a different configuration of type of predicted transition probabilities, censoring scenario, and sample size. Within each figure, each column refers to a different state k, and each row corresponds to a different estimation method (row 1: pseudo-value; row 2: BLR-IPCW; row 3: MLR-IPCW). As explained in Section 5.3, smoothed plots are shown for pseudo-value and BLR-IPCW methods, while scatter plots are used for MLR-IPCW. In both cases, the estimates from each repetition are displayed in red, while the true calibration curve is shown in blue.

The comparison between Figures 5.5 and Figure 5.14, related to perfectly predicted calibration probabilities under RC scenario with $n = 500$ and 700 respectively, shows a small increase in the variability of the plots as the sample size decreases, in particular in state 3. These considerations hold for every pair of graphs of sizes 500 and 700 that share the same cansoring scenario and type of predicted risks.

Across all the graphs, it is evident that the variability is higher in the regions where the density of predicted risks is small. This happens under all three scenarios and the pseudo-values and BLR-IPCW methods, which both produce smoothed curves, present almost the same results.

When predicted risks are miscalibrated, the true calibration is not a straight line, and $\widehat{p}^{i}_{k,miscal1}$ leads to an over-estimation of states 1 and 2 and an under-estimation of states

4 and 5 in most of the plots. Instead, for $\widehat{p}^{i}_{k,miscal1}$, the over-estimation occurs in states 1 and 5, while states 2,3 and 4 are under-estimated. These results are consistent with the way the miscalibrated predicted risks were constructed.

In all cases, no substantial difference is observed between the resulting smoothed curves estimated using the pseudo-value and BLR-IPCW methods methods.

5.6.3 Discussion

All the methods considered for the assessment of *mean* calibration produce very small bias, even in scenarios where the outcome and the censoring mechanism are strongly associated.

In contrast, the comparison of performance for *moderate* calibration shows greater variability.

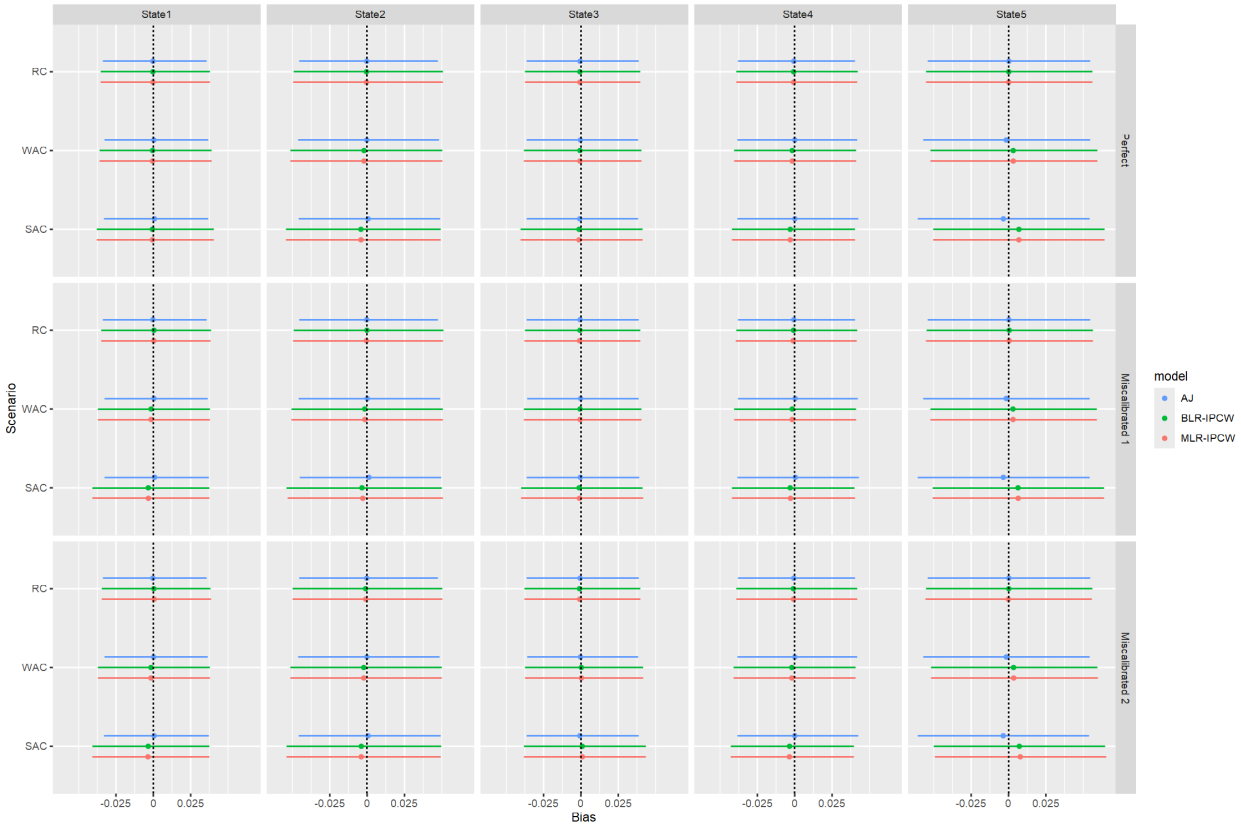


Figure 5.2: Median bias, with 2.5 to 97.5 percentile range, of *mean* calibration estimation under sample size $n = 400$. Each panel refers to a different state and type of predicted risk. Within each panel, results are grouped by censoring scenario and colored according to the estimation method (blue: AJ; green: BLR-IPCW; red: MLR-IPCW).

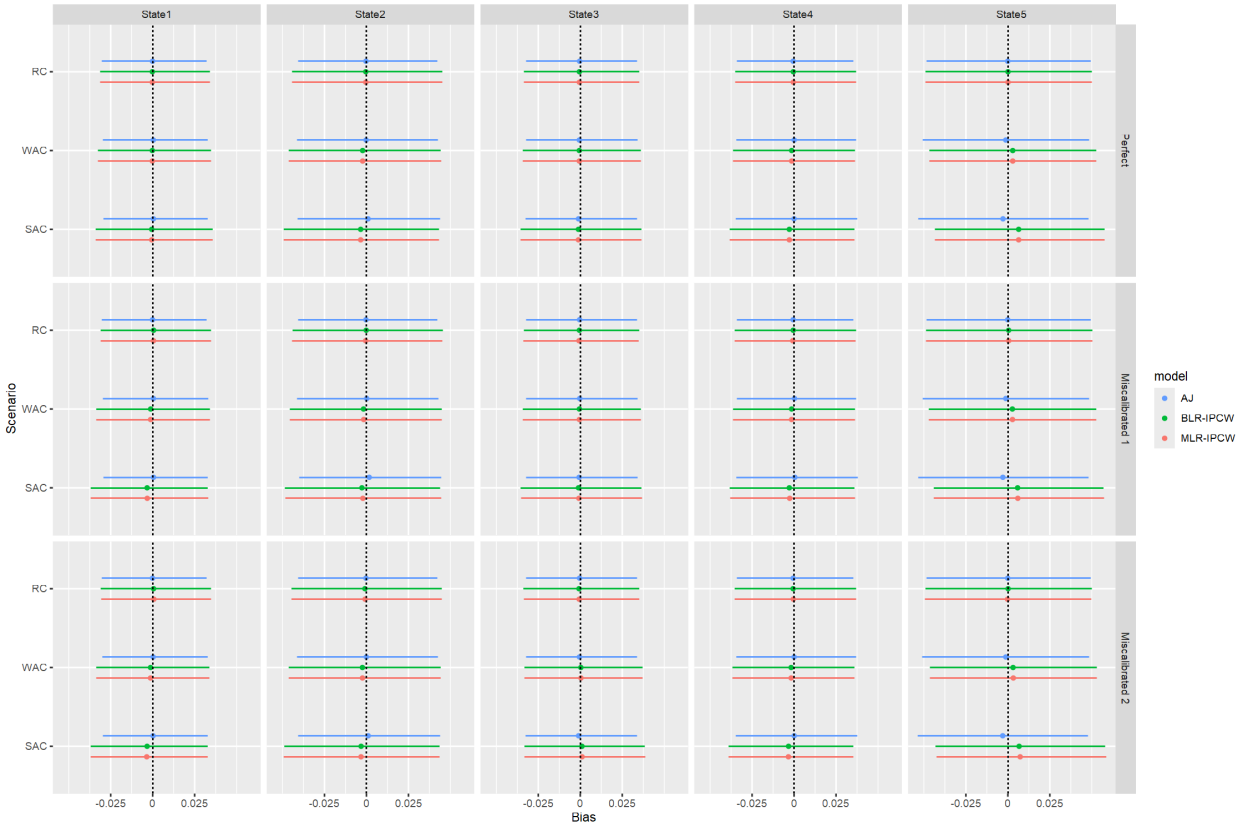


Figure 5.3: Median bias, with 2.5 to 97.5 percentile range, of *mean* calibration estimation under sample size $n = 500$. Each panel refers to a different state and type of predicted risk. Within each panel, results are grouped by censoring scenario and colored according to the estimation method (blue: AJ; green: BLR-IPCW; red: MLR-IPCW).

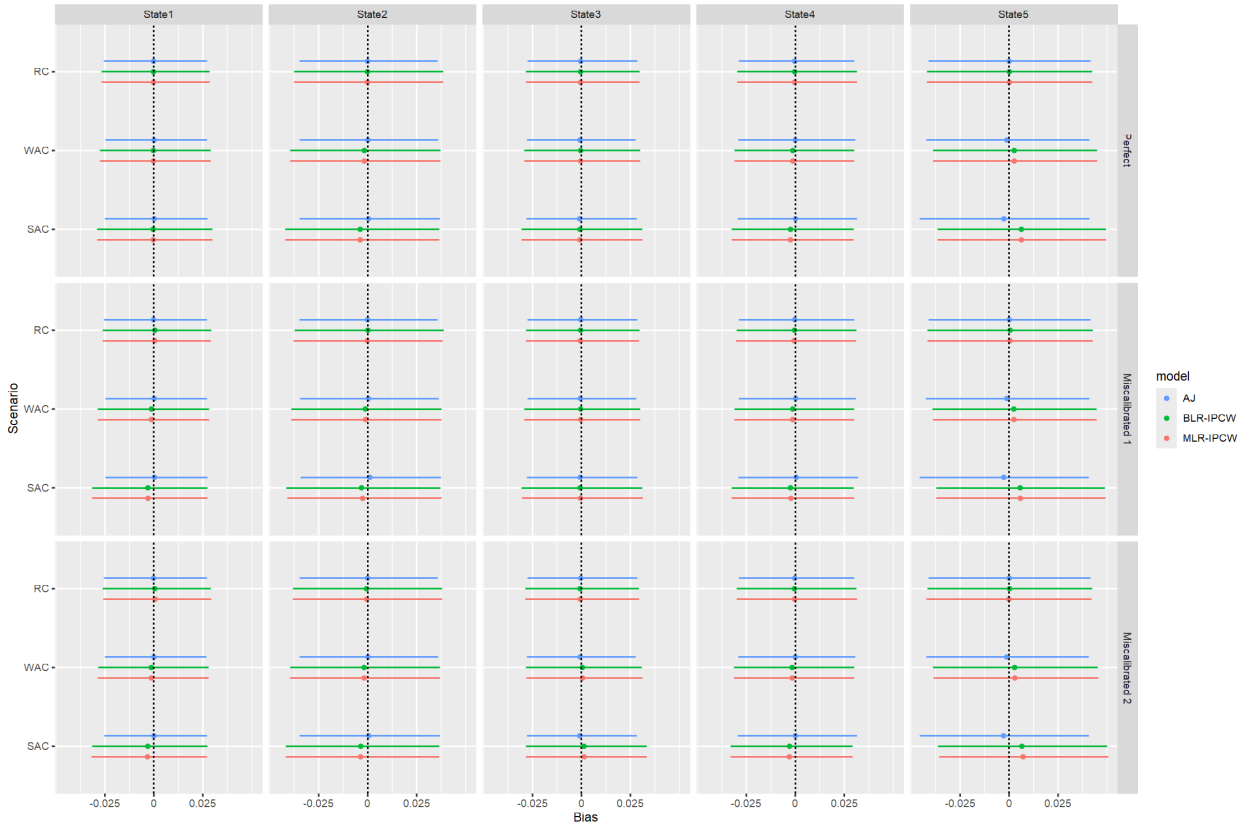


Figure 5.4: Median bias, with 2.5 to 97.5 percentile range, of *mean* calibration estimation under sample size $n = 700$. Each panel refers to a different state and type of predicted risk. Within each panel, results are grouped by censoring scenario and colored according to the estimation method (blue: AJ; green: BLR-IPCW; red: MLR-IPCW).

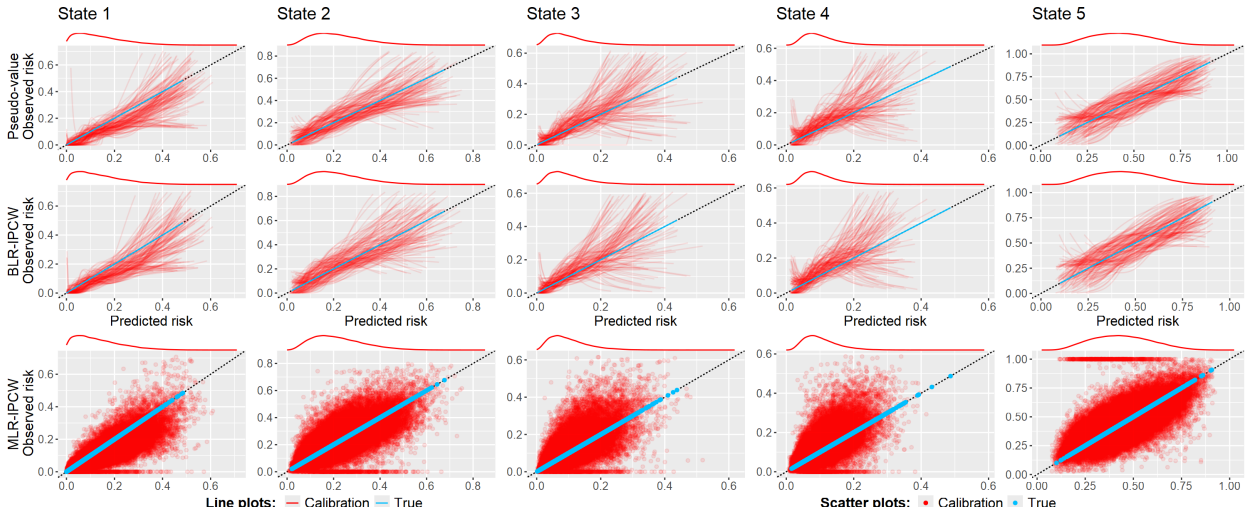


Figure 5.5: *Moderate* calibration plots for perfectly predicted calibration probabilities $\hat{p}^i_{k,true}$, under RC scenario and sample size $n = 500$.

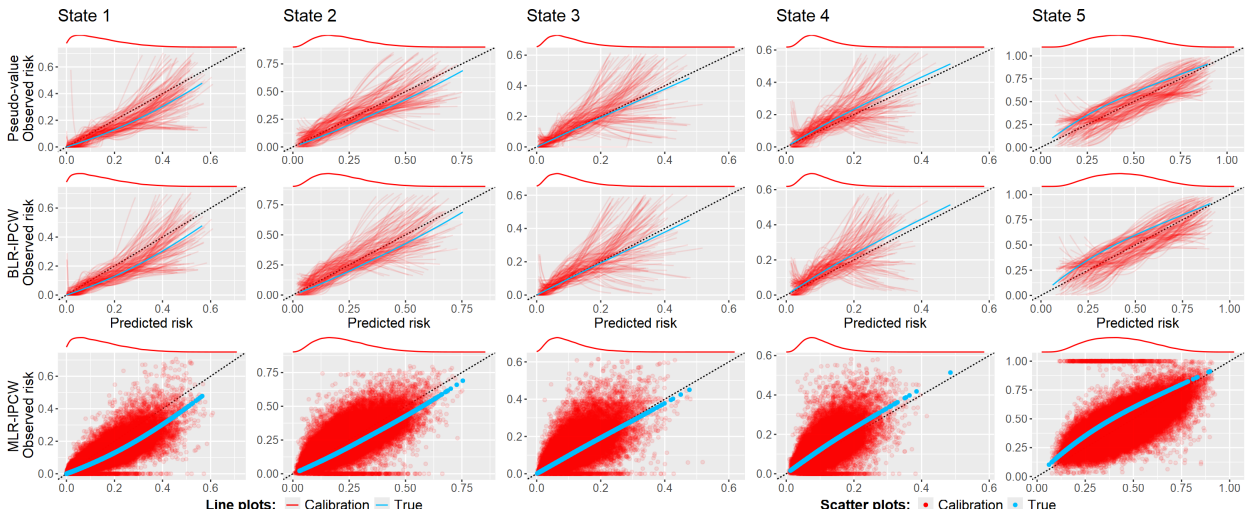


Figure 5.6: *Moderate* calibration plots for miscalibrated predicted calibration probabilities $\hat{p}^i_{k,miscal1}$, under RC scenario and sample size $n = 500$.

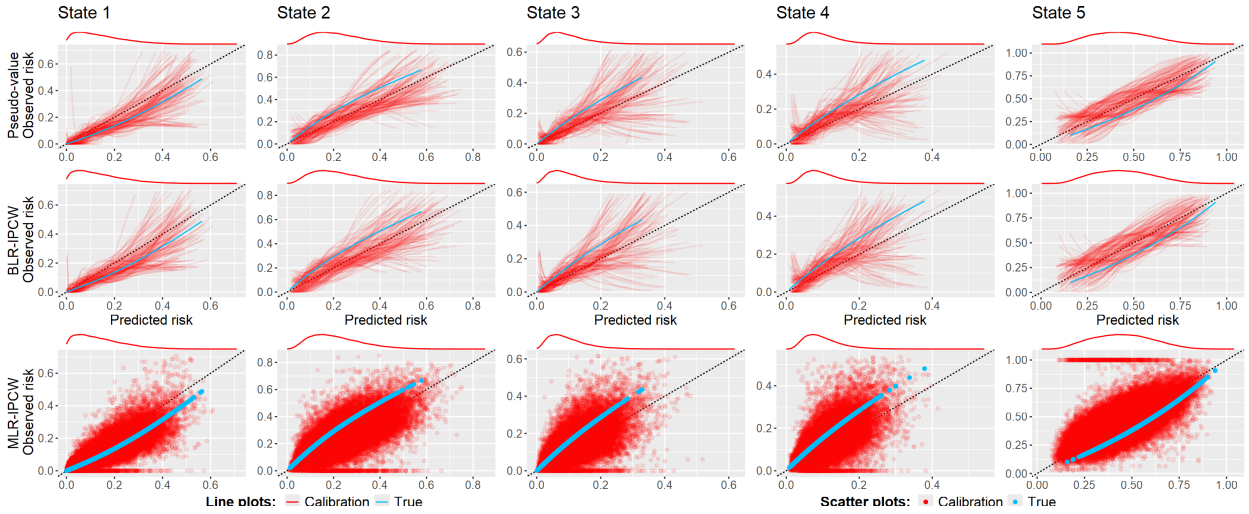


Figure 5.7: *Moderate* calibration plots for miscalibrated predicted calibration probabilities $\hat{p}^i_{k,miscal2}$, under RC scenario and sample size $n = 500$.

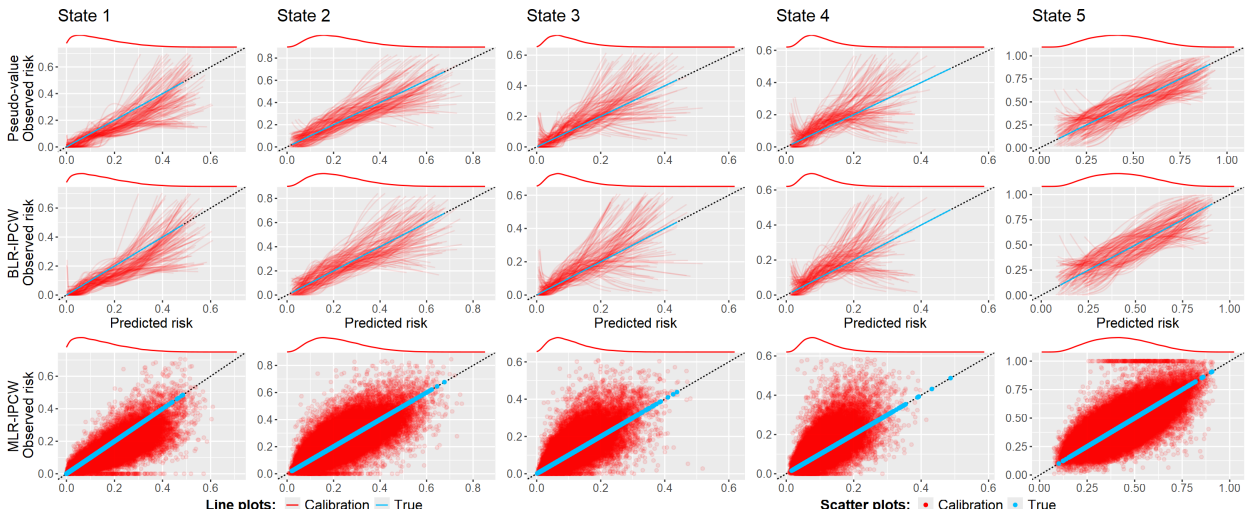


Figure 5.8: *Moderate* calibration plots for perfectly predicted calibration probabilities $\hat{p}^i_{k,true}$, under WAC scenario and sample size $n = 500$.

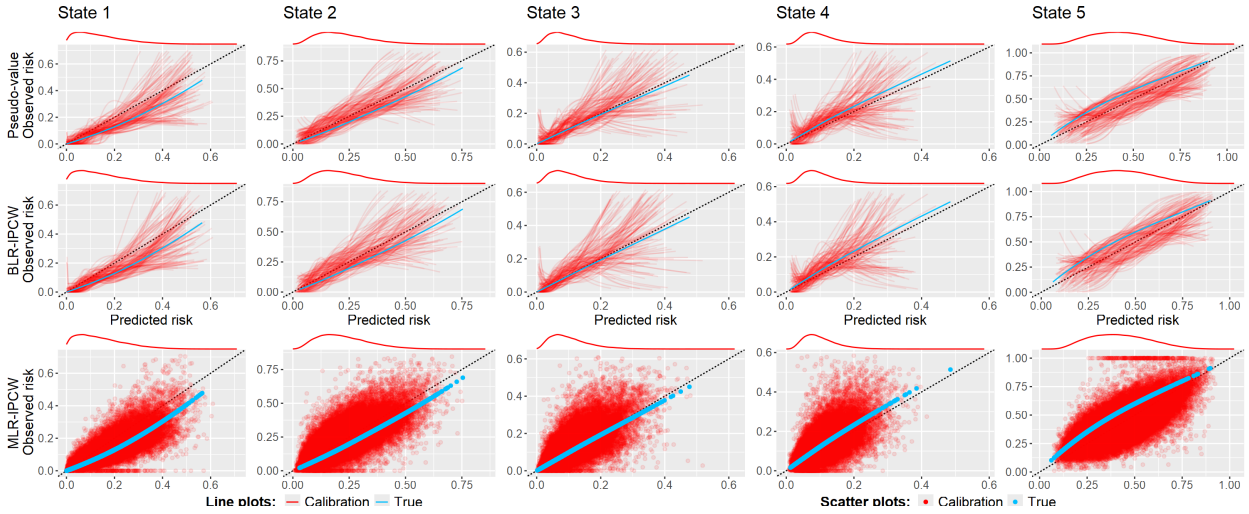


Figure 5.9: *Moderate* calibration plots for miscalibrated predicted calibration probabilities $\hat{p}^i_{k,miscal1}$, under WAC scenario and sample size $n = 500$.

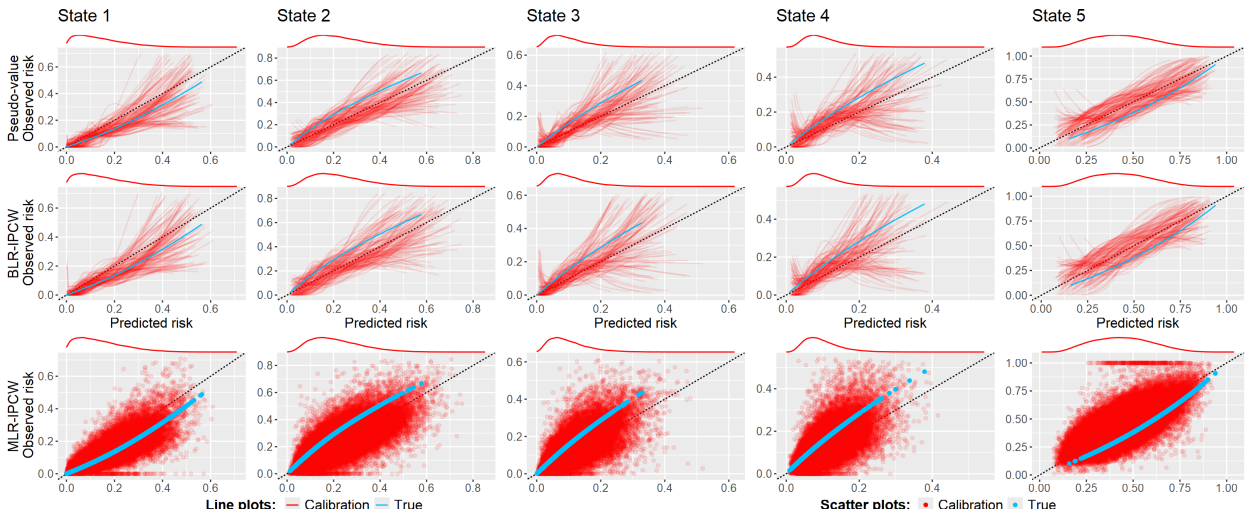


Figure 5.10: *Moderate* calibration plots for miscalibrated predicted calibration probabilities $\hat{p}^i_{k,miscal2}$, under WAC scenario and sample size $n = 500$.

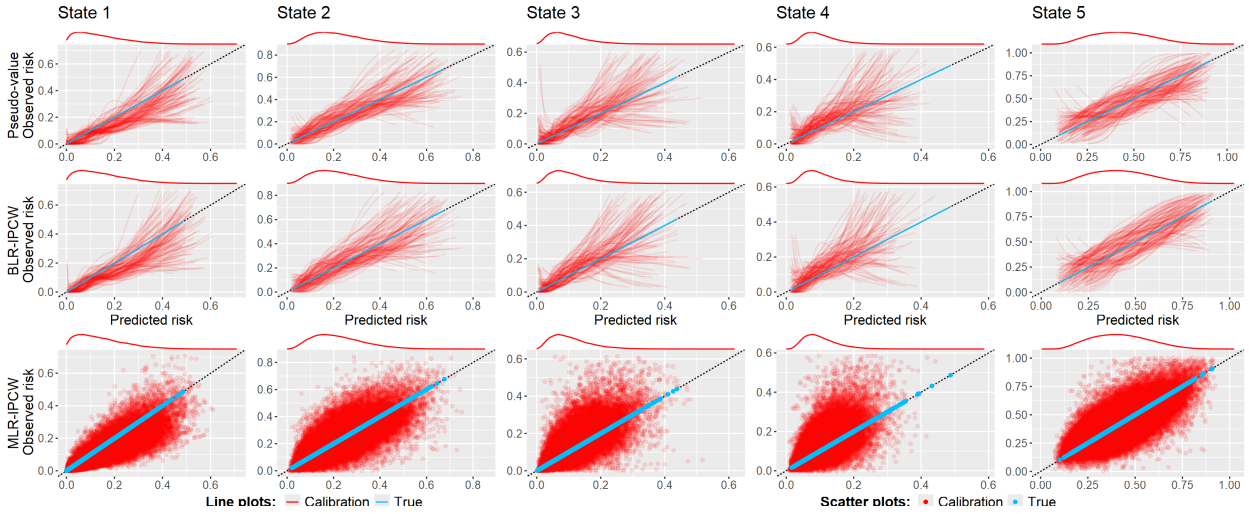


Figure 5.11: *Moderate* calibration plots for perfectly predicted calibration probabilities $\hat{p}^i_{k,true}$, under SAC scenario and sample size $n = 500$.

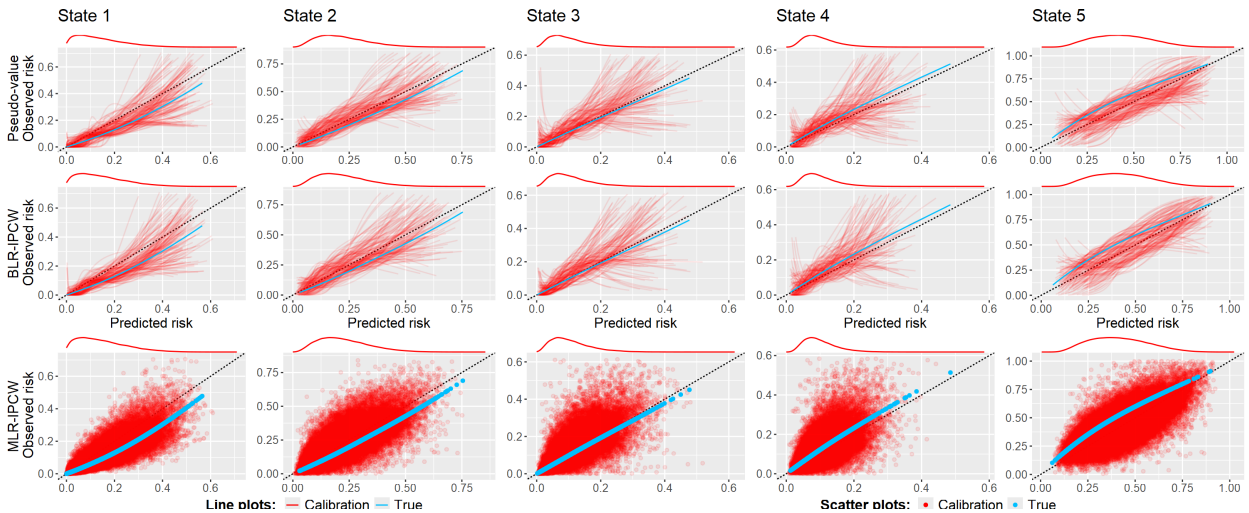


Figure 5.12: *Moderate* calibration plots for miscalibrated predicted calibration probabilities $\hat{p}^i_{k,miscal1}$, under SAC scenario and sample size $n = 500$.

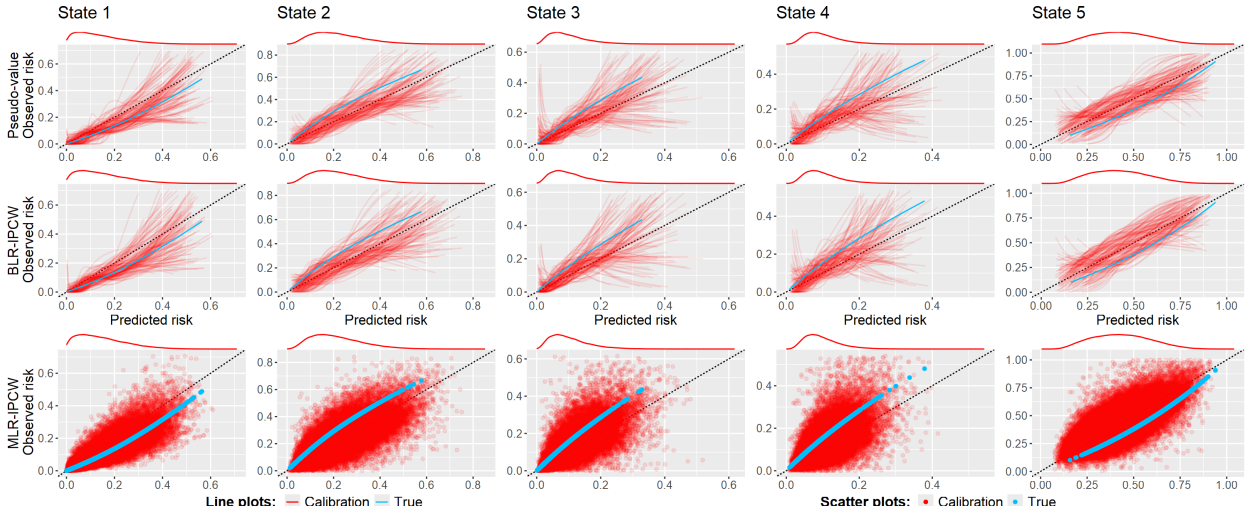


Figure 5.13: *Moderate* calibration plots for miscalibrated predicted calibration probabilities $\hat{p}^i_{k,miscal2}$, under SAC scenario and sample size $n = 500$.

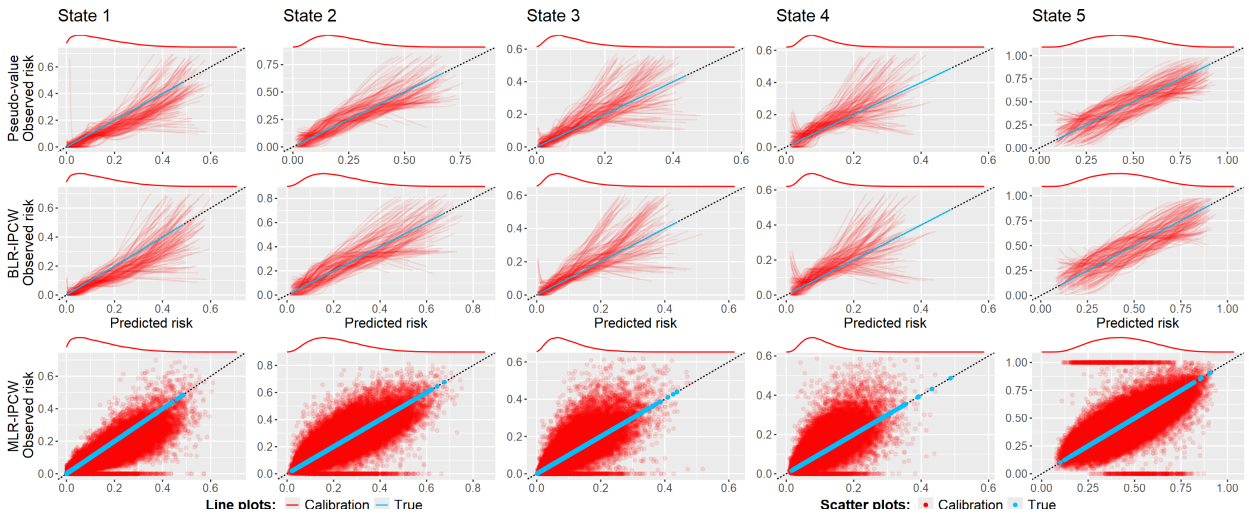


Figure 5.14: *Moderate* calibration plots with perfectly predicted calibration probabilities $\hat{p}^i_{k,true}$, under RC scenario and sample size $n = 700$.

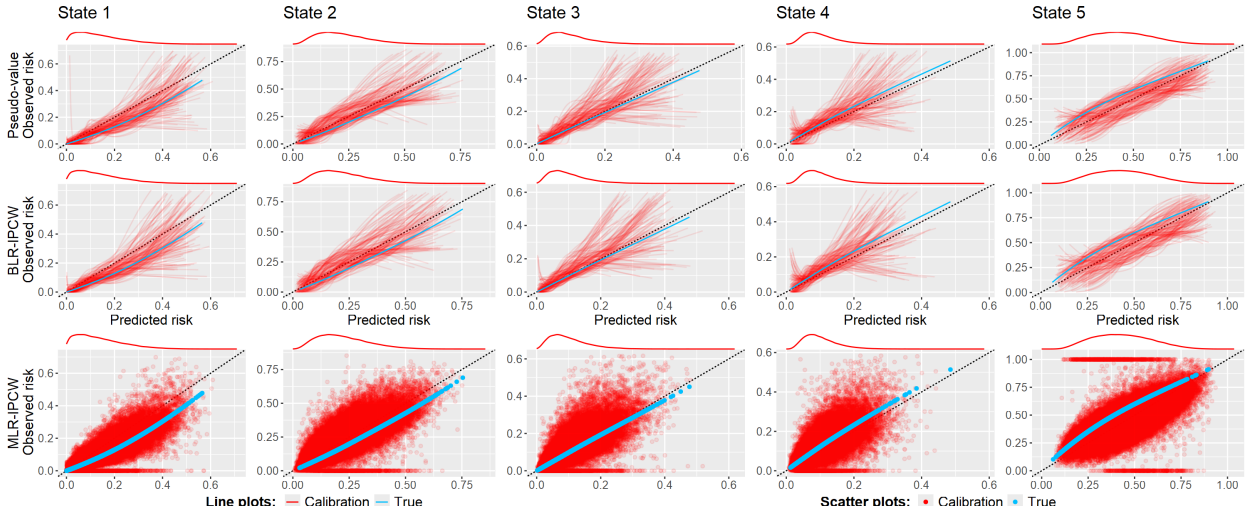


Figure 5.15: *Moderate* calibration plots for miscalibrated predicted calibration probabilities $\hat{p}^i_{k,miscal1}$, under RC scenario andsample size $n = 700$.

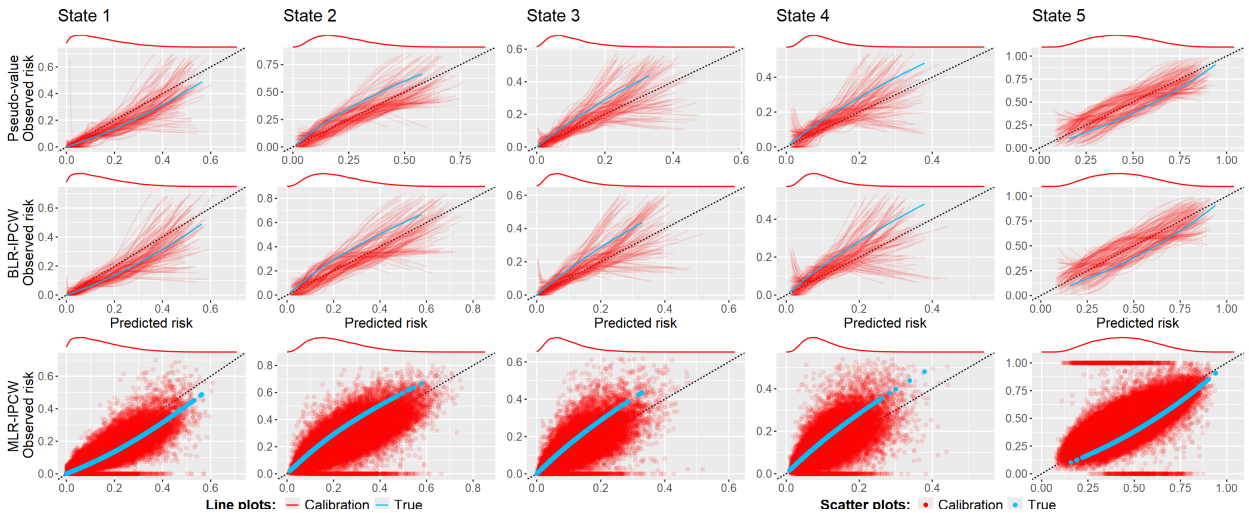


Figure 5.16: *Moderate* calibration plots for miscalibrated predicted calibration probabilities $\hat{p}^i_{k,miscal2}$, under RC scenario and sample size $n = 700$.

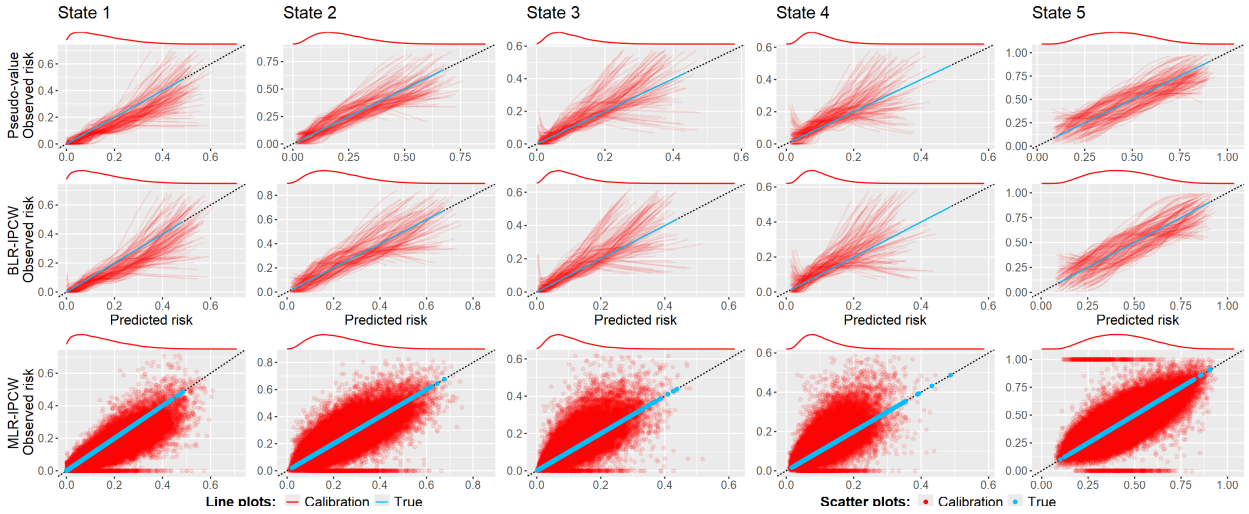


Figure 5.17: *Moderate* calibration plots for perfectly predicted calibration probabilities $\hat{p}^i_{k,true}$, under WAC scenario and sample size $n = 700$.

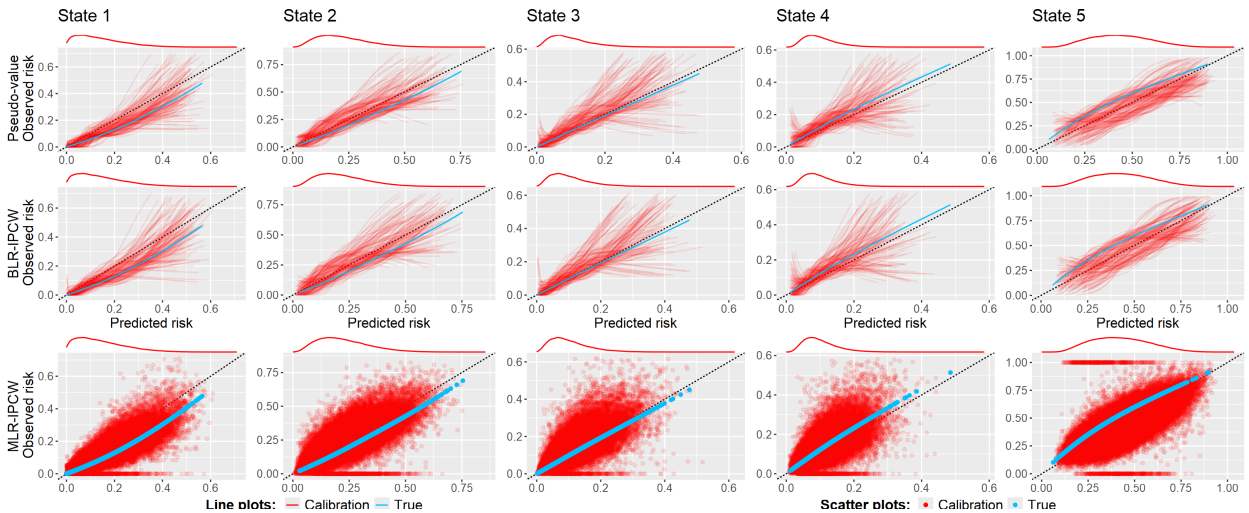


Figure 5.18: *Moderate* calibration plots for miscalibrated predicted calibration probabilities $\hat{p}^i_{k,miscal1}$, under WAC scenario and sample size $n = 700$.

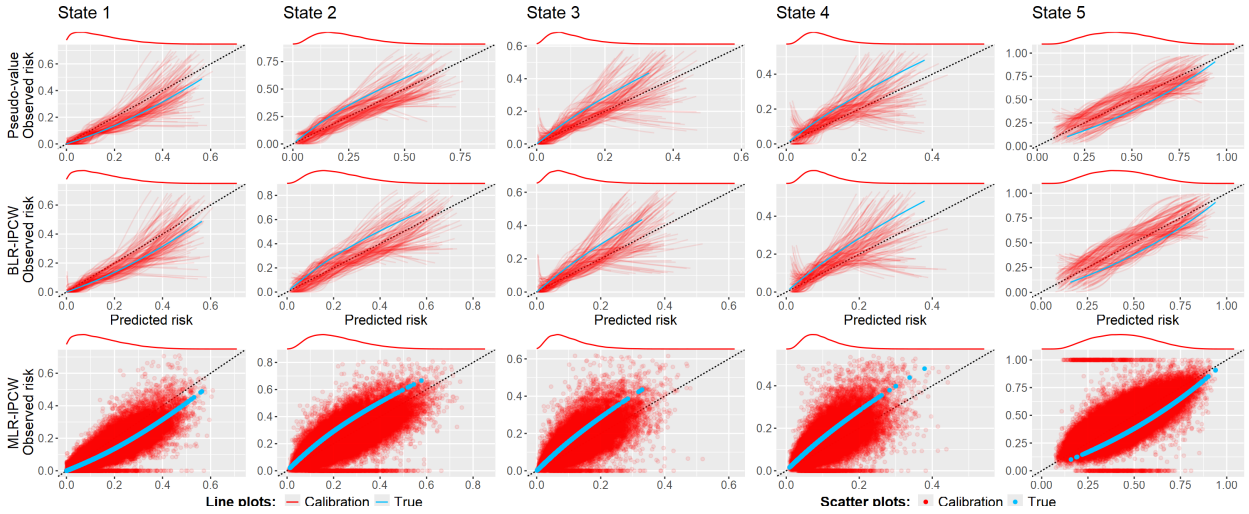


Figure 5.19: *Moderate* calibration plots for miscalibrated predicted calibration probabilities $\hat{p}^i_{k,miscal2}$, under WAC scenario and sample size $n = 700$.

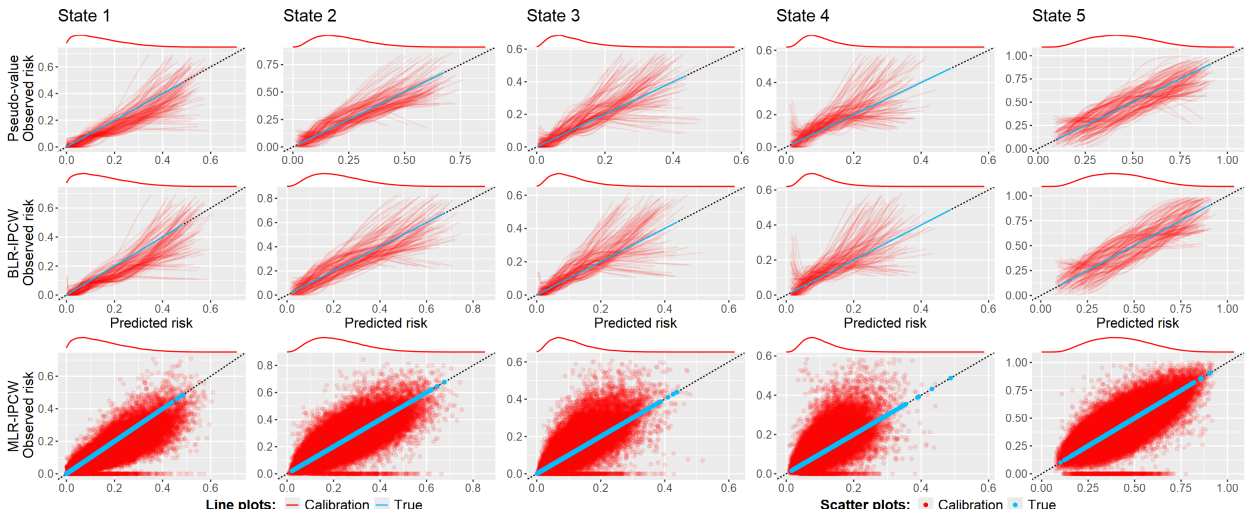


Figure 5.20: *Moderate* calibration plots for perfectly predicted calibration probabilities $\hat{p}^i_{k,true}$, under SAC scenario and sample size $n = 700$.

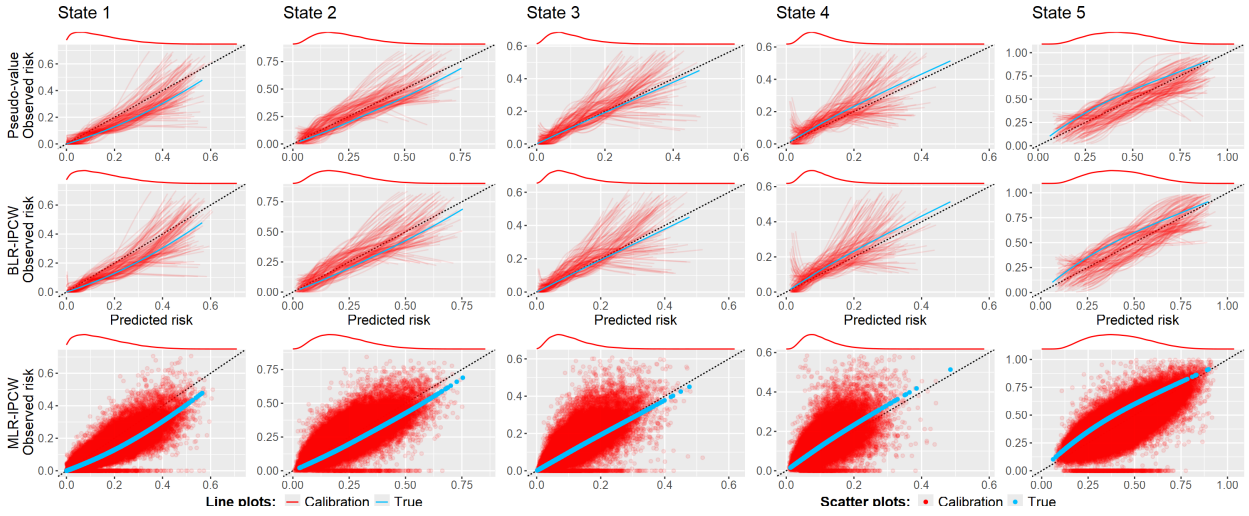


Figure 5.21: *Moderate* calibration plots for miscalibrated predicted calibration probabilities $\hat{p}^i_{k,miscal1}$, under SAC scenario and sample size $n = 700$.

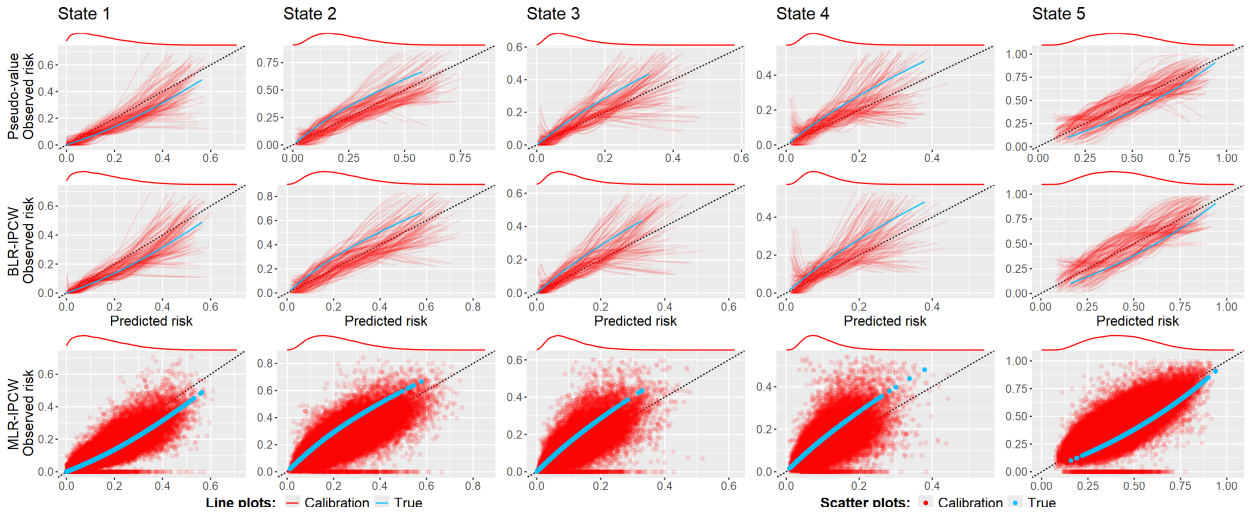


Figure 5.22: *Moderate* calibration plots for miscalibrated predicted calibration probabilities $\hat{p}^i_{k,miscal2}$, under SAC scenario and sample size $n = 700$.

Chapter 6

Application: Ewing sarcoma

In this chapter a real data application of the methodology presented in Chapter 4 will be reported.

Data were collected within the Euro Ewing 2012 (EE2012) trial: an international, phase III, open-label, randomized controlled trial, that involved patients with Ewing sarcoma of bone or soft tissue, as well as Ewing-like sarcomas.

Patients were randomized into two groups of induction chemotherapy regimes: arm A received the VIDE therapy and arm B received the VDC/IE therapy. A detailed comparison of the two chemotherapy regimens can be found in Brennan *et al*[2].

6.1 Ewing data

The Ewing dataset contains observations regarding 640 patients. The baseline observations were collected at the time of induction chemotherapy randomization, which is considered the starting point of the survival analysis.

In total 630 patients were included in this work: eight patients were lost to follow-up and the location of the primary tumor of two others was not reported. The selection process is

shown in Figure 6.1.

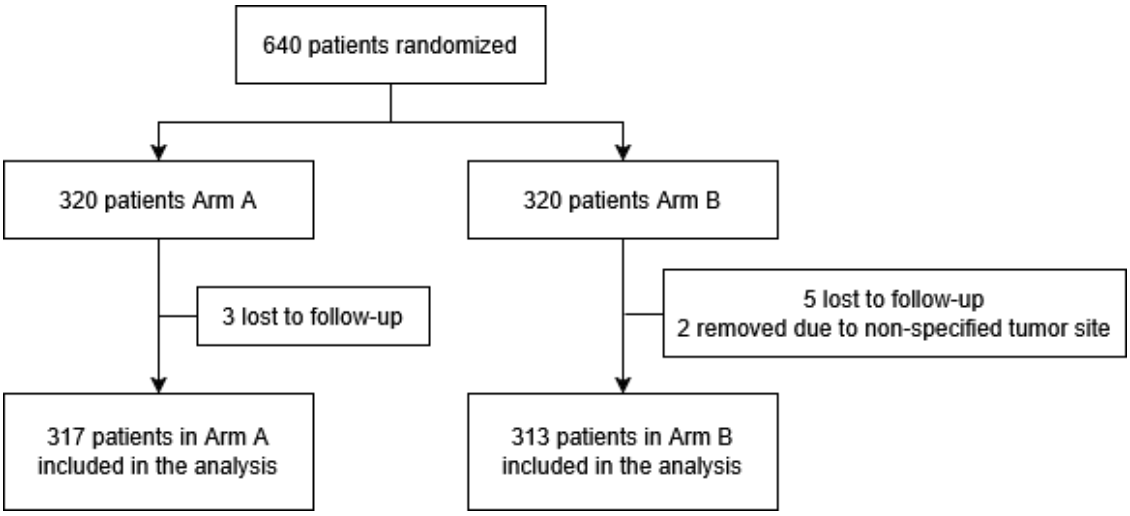


Figure 6.1: Selection process of patients

This work considered five baseline variables: age (in years), sex (*Male*, *Female*), primary tumor site (*Axis*, *Limb)*, volume of the tumor (< 200 mL, ≥ 200 mL), and chemotherapy regime (*VIDE*, *VDC/IE*). Table 6.1 illustrates the distribution of the baseline variables in the entire dataset and in the groups defined by the chemotherapy regime.

	N(%)	VIDE(%)	VDC(%)
Tot	630	317	313
Age, years			
< 14	263(42%)	133(42%)	130(42%)
≥ 14	367(58%)	184(58%)	183(58%)
Sex			
Male	364(58%)	179(56%)	185(59%)
Female	266(42%)	138(44%)	128(41%)
Tumour volume			
< 200 mL	358(57%)	187(59%)	171(55%)
≥ 200 mL	272(43%)	130(41%)	142(45%)
Primary tumour site			
Axis	383(61%)	195(62%)	188(60%)
Limb	247(39%)	122(38%)	125(40%)

Table 6.1: Baseline clinical characteristics

The two categories *Axis* and *Limb* for tumor site were obtained by grouping the originally categories reported in the dataset as follows: 'Upper extremity' and 'Lower extremity' were

grouped into *Limb*, while 'Pelvis', 'Abdomen', 'Spine', 'Chest', 'Head and Neck' were grouped into *Axis*. This division formed two balanced groups of patients, preventing convergence problems of the model.

6.2 Models and small-sample limitations

The existing literature on Ewing sarcoma often considers as events of interest in survival analysis the death of patients and three possible types of disease progression: local recurrence (LR), distant pulmonary metastasis (DMP), and other distant metastasis (DMO).

The dataset used in this thesis contained a sufficient number of patients and transitions to perform a survival analysis on the 5-state model shown in Figure 6.2.

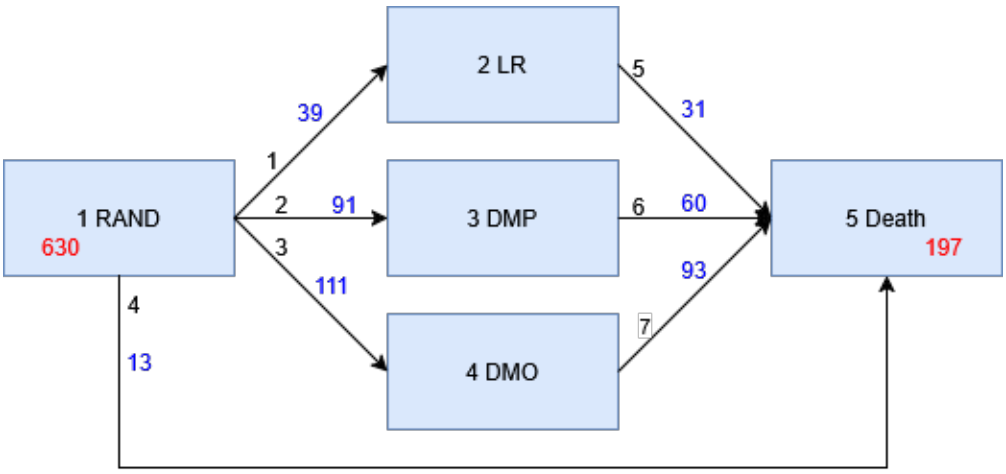


Figure 6.2: Multi-state model with 5 states {*1 RAND: randomization; 2 LR: local recurrence; 3 DMP: distant pulmonary metastasis; 4 DMO: other distant metastasis; 5 Death*} and all patients starting in the randomization state. State and transition number are written in black, number of transitions in blue and number of patients starting at time of randomization and entering the death state in red.

However, the assessment of calibration required splitting the dataset into a development set and a validation set. With this division, a model with five states and seven transitions was no longer compatible with the smaller sample sizes of the development and validation set, due to the low counts observed for some transitions. Thus, a reduced version of the model

was evaluated. As shown in Figure 6.3, the transition from Randomization to Death was removed, and the LR and DMP states were merged to increase the number of patients for each transition.

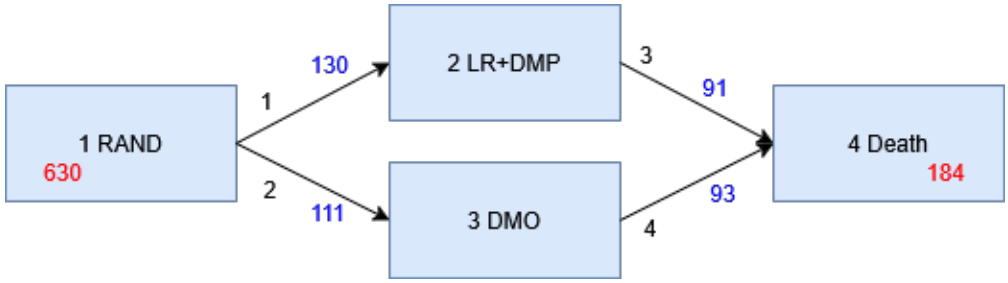


Figure 6.3: Multi-state model with 4 states {*1 RAND: randomization; 2 LR+DMP: local recurrence and distant pulmonary metastasis; 3 DMO: other distant metastasis; 4 Death*} and all patients starting in the randomization state. State and transition number are written in black, number of transitions in blue and number of patients starting at time of randomization and entering the death state in red.

Once the 4-state model structure was determined, a development set of 300 patients was used to estimate a Cox model that provided the regression coefficients used in the transition probabilities estimates on the validation set of 330 patients. The corresponding Figures 6.4 and 6.5 display the transition counts of the 4-state model for the two datasets, illustrating the distribution of observed transitions in the development and validation sets, respectively.

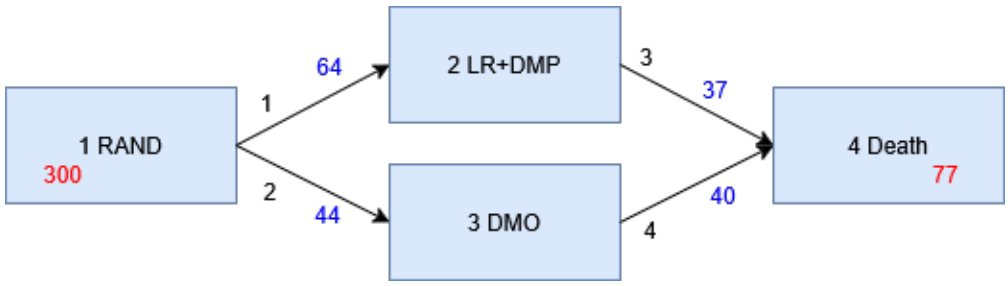


Figure 6.4: Observed state transitions in the development set with 300 patients for the 4-state model with states {*1 RAND; 2 LR+DMP; 3 DMO; 4 Death*}, in which all patients start in the randomization state. State and transition number are written in black, number of transitions in blue and number of patients starting at time of randomization and entering the death state in red.

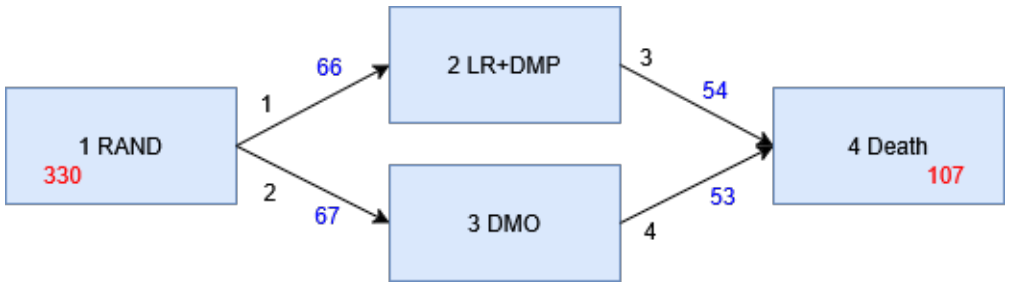


Figure 6.5: Observed state transitions in the validation set of 330 patients for the 4-state model with states {*1 RAND; 2 LR+DMP; 3 DMO; 4 Death*}, in which all patients starting in the randomization state. State and transition number are written in black, number of transitions in blue and number of patients starting at time of randomization and entering the death state in red.

Once reached this stage, the calibration should have been assessed by using the `calibmsm` package[3], but problems arose when using the `calib_msm()` function. This was due to the absence of evaluation time points at which at least 30 were present patients in each state. This scarcity of events induced a lack of convergence that produced `NA` values and thus it was not possible to obtain calibration plots. Indeed, a substantial proportion of patients died shortly after experiencing a disease progression. Consequently, in the validation set there were never 30 patients in the *LR+DMP* and *DMO* states at the same time: most of them remained in Randomization state or transitioned to Death soon after enetering the intermediate events.

To overcome this issue and enable calibration assessment on the Ewing data, it was necessary to further simplify the model, obtaining the 3-state model structure shown in Figure 6.6, where *LR+DMP* and *DMO* states were merged into a single progression state *Progr*. Figures 6.4 and 6.5 display the transition counts of the 3-state model for the development and validation sets, showing the distribution of observed transitions in the datasets. By applying this 3-state model structure to both the development and validation sets, it was finally possible to obtain calibration plots. Further details and results are reported in Section 6.4.1.

Before proceeding with the calibration of the simplified 3-state model, the 5-state model estimated on the complete dataset of 630 patients is presented in the next section, since it

was more relevant from a medical point of view.

Figure 6.6: Model with 3 states {*1 RAND: randomization; 2 Progr: disease progression (LR+DMP+DMO); 3 DMO; 4 Death*} and all patients starting in the randomization state. State and transition number are written in black, number of transitions in blue and number of patients starting at time of randomization and entering the death state in red.

Figure 6.7: Observed state transitions in the development set with 300 patients for the 3-state model with states {*1 RAND; 2 Progr; 3 Death*}, in which and all patients starting in the randomization state. State and transition number are written in black, number of transitions in blue and number of patients starting at time of randomization and entering the death state in red.

Figure 6.8: Observed state transitions in the validation set with 330 patients for the 3-state model with states {*1 RAND; 2 Progr; 3 Death*}, in which all patients starting in the randomization state. State and transition number are written in black, number of transitions in blue and number of patients starting at time of randomization and entering the death state in red.

6.3 Multistate model with five states

The multistate model estimated in this section (Figure 6.2) comprises five states {*1 RAND: randomization; 2 LR: local recurrence; 3 DMP: distant pulmonary metastasis; 4 DMO: other distant metastasis; 5 Death*} and seven transitions. The initial state for all patients is Randomization (state 1), while Death (state 5) is the sole absorbing state. Overall survival (OS) was measured from the date of randomization until the patient's death or end of follow-up.

Using a 'clock-forward' approach and assuming a Markov model, the transition-specific coefficients of a Cox model were estimated. Table 6.2 shows the estimated log hazard ratios for transition-specific covariates. Results show significant statistical influence of tumor volume on the transitions from randomization to every progression event, with a volume greater than 200 mL increasing the risk of experiencing the events. Age also influences the transitions from randomization to development of metastasis. In this case a patient being younger than 14 decreases the risk of experiencing the events. The last noticeable effect is that of being assigned to Arm A, following the VIDE therapy, which increases the hazard of transitioning from randomization to other distant metastasis.

Log-HR estimates for all prognostic factors and the different transitions in the 5-state model							
	Rand→LR	Rand→DMP	Rand→DMO	Rand→Death	LR→Death	DMP→Death	DMO→Death
Predictor	HR (p-value)	HR (p-value)	HR (p-value)	HR (p-value)	HR (p-value)	HR (p-value)	HR (p-value)
Sex, *female*		-0.12 (0.591)	-0.26 (0.193)		-0.79 (0.062)	-0.05 (0.870)	0.03 (0.876)
Volume, $\geq$ 200mL	0.85 (0.011)	0.62 (0.004)	0.64 (0.001)		0.33 (0.420)	0.10 (0.731)	-0.14 (0.535)
Age, $<$ 14		-0.71 (0.003)	-0.45 (0.028)	-0.97 (0.139)		-0.14 (0.673)	-0.19 (0.421)
Site, *Limb*	-0.68 (0.076)	0.16 (0.475)	-0.25 (0.226)		-0.23 (0.653)	-0.35 (0.225)	0.22 (0.355)
Chemo, *Arm A*	0.40 (0.223)	0.14 (0.499)	0.41 (0.032)			-0.03 (0.923)	-0.20 (0.362)

Table 6.2: Estimated log hazard ratios for transition-specific covariates using the complete set (630 subjects). See Figure 6.2 for the corresponding observed transition counts.

Then, the occupation probabilities (from time 0) in each state were obtained through the `probtrans()` function of the `mstate` package[23] in R version 4.5.0. Figure 6.9 reports the results in a six-year period. As expected from the number of transitions, the occupation probability of local recurrence is the lowest, while the occupation probability of randomization state remains the highest in the entire period. The probability of developing metastasis, pulmonary and others, is highest between 2 and 3 years after randomization, while the probability of death increases over time.

Finally, Figure 6.10 shows a comparison between two patients with the same characteristics but different tumor volume. Although this covariate was particularly significant for the

model, there is no noticeable difference between the state occupation probabilities of the two patients, which remain very similar to the ones in Figure 6.9.

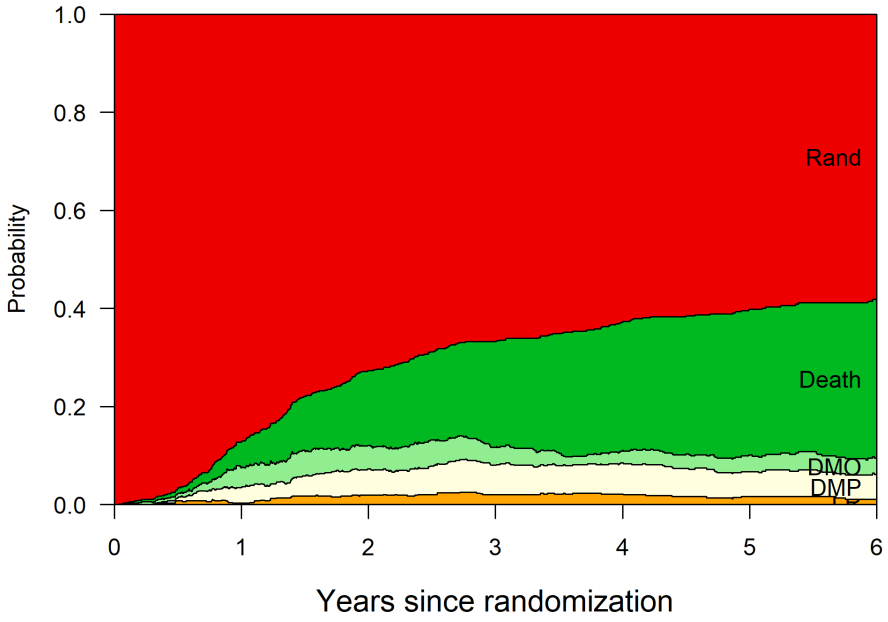
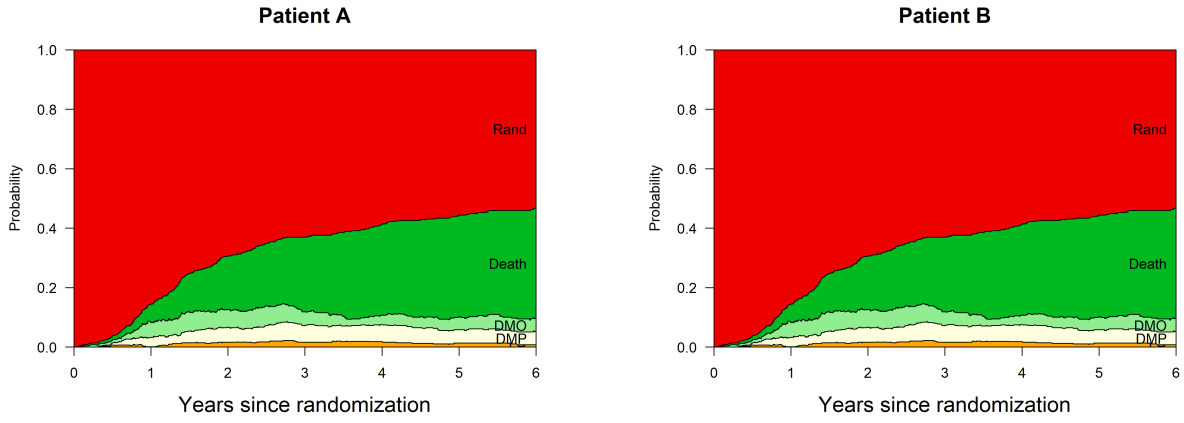


Figure 6.9: State occupation probabilities from randomization



(a) Patient A: tumor volume < 200mL

(b) Patient B: tumor volume ≥ 200mL

Figure 6.10: State occupation probabilities from randomization for two patients with the same characteristics but different tumor volume: (a) < 200 mL, (b) ≥ 200 mL.

6.4 Multistate model with three states

As explained in Section 6.2, the 5-state model could not be calibrated due to the insufficient number of patients in some transition states when splitting the data into development and validation sets, which prevented convergence of the `calib_msm()` function. Therefore, calibration was assessed only for the simplified 3-state model with Randomization, Progression, and Death states in Figure 3.2.

For the 3-state structure, calibration was evaluated using the `calibmsm` package[3] by estimating transition probabilities in the development set and comparing them with the observed probabilities in the validation set. The detailed *mean* and *moderate* calibration results and the corresponding plots are presented in Section 6.4.1.

6.4.1 Calibration

First, a Cox model was estimated on a development set of 300 patients. Table 6.7 shows the estimated log hazard ratios for transition specific covariates in the 3-states model.

Log-HR estimates for all prognostic factors and different transitions in the 3-state model		
	Rand →Progr	Progr → Death
Predictor	HR (p-value)	HR (p-value)
Sex, *female*	0.12 (0.550)	-0.22 (0.338)
Volume, $\geq$ 200mL	0.56 (0.005)	-0.04 (0.882)
Age, $<$ 14	-0.52 (0.011)	0.14 (0.588)
Site, *Limb*	-0.08 (0.711)	-0.26 (0.302)
Chemo, *Arm A (VIDE)*	0.37 (0.063)	0.04 (0.875)

Table 6.3: Estimated log hazard ratios for transition specific covariates in a 3-states model using the development set (300 subjects). See Figure 6.4 for the corresponding observed transition counts.

Then, using `probtrans()`, the state occupation probabilities were estimated in the validation set of 330 patients. The time of evaluation for the calibration was set to 2.7 years and the corresponding occupation probabilities were saved into the validation dataset in wide format.

The calibration was finally assessed using the same methods employed in the simulation

study (Section 5.5). Table 6.4 reports the values of *mean* calibration, which are very similar for all methods. The Aalen-Johansen (AJ) method performs slightly better than the BLR and the MLR methods for state 1, but worse for state 2 and 3, while the difference between BLR and MLR is negligible. In general, all methods show good results in assessing mean calibration.

Mean calibration			
Method	State 1 (Rand)	State 2 (Progr)	State 3 (Death)
BLR	-0.02276780	-0.03349228	0.05618954
MLR	-0.02271558	-0.03346632	0.05618190
AJ	-0.02097490	-0.03602439	0.05680707

Table 6.4: Mean calibration calculated using the BLR-IPCW, MLR-IPCW and AJ methods.

Moderate calibration was evaluated using calibration plots: smoothed curves are shown for pseudo-value (Figure 6.11) and BLR-IPCW (Figure 6.12) methods, while scatter plots are used for MLR-IPCW (Figure 6.13). All methods performed better in state 1, while the worst results were registered in state 2. This is consistent with the number of patients occupying the states, since state 1 always contains the higher number of patients, while at the time of evaluation state 2 has the lowest number and the accuracy of predictions grows with the number of observations. Regarding the differences between methods, there are only minor changes between the graphs of BLR-IPCW and pseudo-values in state 2 and 3. In state 3 in particular, the pseudo-values method produces a wider confidence interval for low values of predicted risk.

6.4.2 Discussion

This real data example highlighted the limitations of assessing calibration in small samples: the critical factor is the number of patients occupying each state at the evaluation time. *Mean* and *moderate* calibration can still be evaluated in samples of approximately 330 patients, provided that the number of transitions is sufficiently high and that they are distributed over the follow-up period such that a time point can be identified at which each state includes at

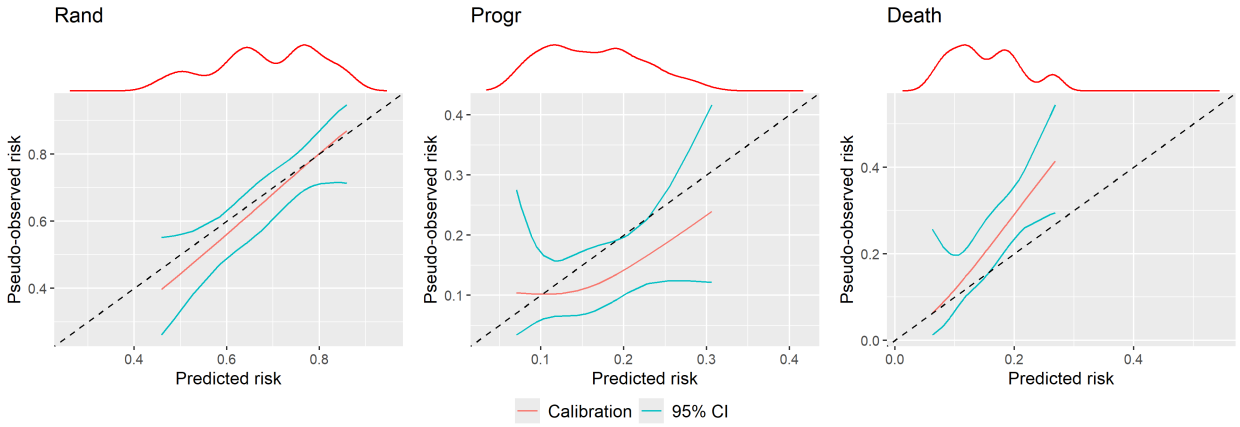


Figure 6.11: *Moderate* calibration obtained from pseudo-values method for each state.

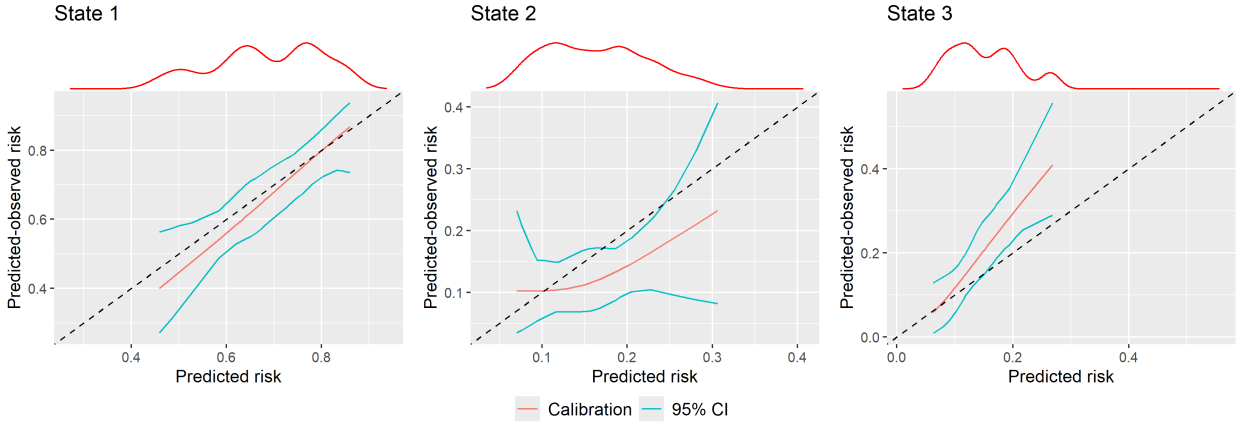


Figure 6.12: *Moderate* calibration obtained from BLR-IPCW method for each state.

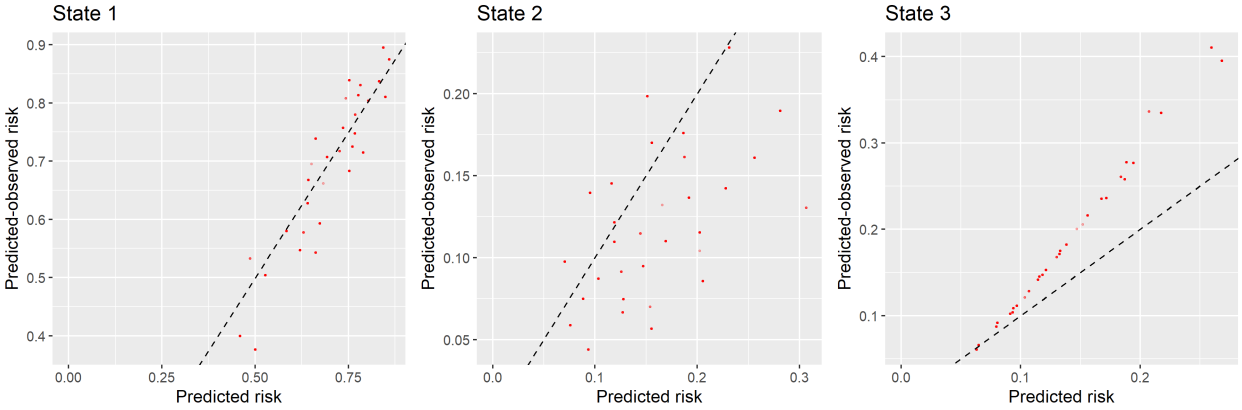


Figure 6.13: *Moderate* calibration obtained from MLR-IPCW method for each state.

least 30 individuals. This threshold was arbitrarily chosen by Pate *et al* in the `calibmsm`[3] package and determines the convergence of the functions used to compute the calibration.

Regardless of the method used, the precision of the calibration depends on the number of patients in each state, as illustrated by the differing performance observed across the three states.

Chapter 7

Discussion

The simulation study and the real data example showed very similar performances across all the methods considered, with no substantial differences in the assessment of *mean* calibration.

For *moderate* calibration, the only distinction between the pseudo-values and binary logistic recalibration with inverse censoring probability weights (BLR-IPCW) methods was the substantially longer computational time required by the pseudo-values approach. This resulted in lengthy simulation runtimes, even when parallel computing was employed.

However, when analyzing single small samples of a few hundred individuals, the computational runtime of the pseudo-values method remained reasonable. Therefore, for producing calibration plots with smoothed lines, either the pseudo-values or the BLR-IPCW approach may be used, as neither method offers a substantial advantage over the other.

It is recommended to combine smoothed calibration curves with scatter plots obtained from multinomial logistic regression with inverse censoring probability weights (MLR-IPCW), as this combination provides a more complete understanding of the model's calibration.

Finally, the required sample size depends on the number of transitions and their distribution over the follow-up period. Even the smallest datasets considered in both the simulation study

and the real data example showed good performance for *moderate* calibration in states with a large number of patients. In addition, the number of patients in each state at the time of evaluation must be sufficiently high, making the choice of the evaluation time particularly important in small samples.

As illustrated in the real data example, a simple model allows the calibration in smaller samples, but at the cost of losing potential relevant information. Future work may therefore explore the use of time-dependent covariates, which could allow the model to be simplified, while preserving important information.

Future developments could focus on extending the present work by incorporating the assessment of *weak* calibration, which represents an alternative perspective to the *mean* and *moderate* calibration examined in this thesis and would allow for a more complete evaluation of model adequacy. In addition, applying these calibration methods to a broader range of real-world small-sample datasets would help determine whether consistent patterns emerge across different datasets. Such applications may ultimately support the development of practical guidelines for performing calibration in small-sample multistate models, clarifying when calibration is feasible, how to select appropriate evaluation times, and which methods are most reliable under realistic data constraints.

Bibliography

[1] Alexander Pate, Matthew Sperrin, Richard D. Riley, Niels Peek, Tjeerd Van Staa, Jamie C. Sergeant, Mamas A. Mamas, Gregory Y. H. Lip, Martin O'Flaherty, Michael Barrowman, Iain Buchan, and Glen P. Martin. Calibration plots for multistate risk prediction models. *Statistics in Medicine*, 43(14):2830–2852, 2024.

[2] Bernadette Brennan, Laura Kirton, Perrine Marec-Bérard, Nathalie Gaspar, Valerie Laurence, Javier Martín-Broto, Ana Sastre, Hans Gelderblom, Cormac Owens, Nicola Fenwick, Sandra Strauss, Veronica Moroz, Jeremy Whelan, and Keith Wheatley. Comparison of two chemotherapy regimens in patients with newly diagnosed ewing sarcoma (ee2012): an open-label, randomised, phase 3 trial. *The Lancet*, 400(10362):1513–1521, 2022.

[3] Martin GP Pate A. calibmsm. https://alexpate30.github.io/calibmsm/, 2024.

[4] Morteza Aalabaf-Sabaghi. Handbook of survival analysis. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 185(Supplement$_2$) : $S775 - -S775, 102022$.

[5] Patricia M. Grambsch and Terry M. Therneau. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81(3):515–526, 09 1994.

[6] Peter C. Austin and Ewout W. Steyerberg. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. *Statistics in Medicine*, 33(3):517–535, 2014.

[7] Andreas Ziegler. Clinical prediction models: A practical approach to development, validation, and updating. *Biometrical Journal*, 62(4):1122–1123, 2020.

[8] Frank E. Harrell. *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis*. Springer International Publishing, Cham, Switzerland, 2nd edition, 2015.

[9] Ben Van Calster, David Mclernon, Maarten van Smeden, Laure Wynants, and Ewout Steyerberg. Calibration: The achilles heel of predictive analytics. *BMC Medicine*, 17, 12 2019.

[10] Ben Van Calster, Daan Nieboer, Yvonne Vergouwe, Bavo De Cock, Michael J. Pencina, and Ewout W. Steyerberg. A calibration hierarchy for risk models was defined: from utopia to empirical data. *Journal of Clinical Epidemiology*, 74:167–176, 2016.

[11] Kirsten Van Hoorde, Yvonne Vergouwe, Dirk Timmerman, Sabine Van Huffel, Ewout W. Steyerberg, and Ben Van Calster. Assessing calibration of multinomial risk prediction models. *Statistics in Medicine*, 33(15):2585–2596, 2014.

[12] K. Van Hoorde, S. Van Huffel, D. Timmerman, T. Bourne, and B. Van Calster. A spline-based tool to assess and visualize the calibration of multiclass risk predictions. *Journal of Biomedical Informatics*, 54:283–293, 2015.

[13] Peter C. Austin, Frank E. Harrell Jr, and David van Klaveren. Graphical calibration curves and the integrated calibration index (ici) for survival models. *Statistics in Medicine*, 39(21):2714–2742, 2020.

[14] Patricia Grambsch. Survival and event history analysis: A process point of view by aalen, o. o., borgan, o., and gjessing, h. k. *Biometrics*, 65(2):663–665, 2009.

[15] Odd O. Aalen and Søren Johansen. An empirical transition matrix for non-homogeneous markov chains based on censored observations. *Scandinavian Journal of Statistics*, 5(3):141–150, 1978.

[16] Somnath Datta and Glen A. Satten. Validity of the aalen–johansen estimators of stage occupation probabilities and nelson–aalen estimators of integrated transition hazards for non-markov models. *Statistics Probability Letters*, 55(4):403–411, 2001.

[17] David V. Glidden. Robust inference for event probabilities with non-markov event data. *Biometrics*, 58(2):361–368, 2002.

[18] Per Kragh Andersen and Maja Pohar Perme. Pseudo-observations in survival analysis. *Statistical Methods in Medical Research*, 19(1):71–99, 2010. Keywords: Adult; Bone Marrow Transplantation; Female; Humans; Leukemia, Myeloid, Acute; Male; Middle Aged; Precursor Cell Lymphoblastic Leukemia-Lymphoma; Proportional Hazards Models; Regression Analysis; Risk Assessment; Survival Analysis; Survival Rate; Young Adult.

[19] Tim P. Morris, Ian R. White, and Michael J. Crowther. Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11):2074–2102, 2019.

[20] Rosetta L. Github repository. lararst. assessment-of-calibration-in-multi-state-models. https://github.com/lararst/Assessment-of-calibration-in-multi-state-models/tree/main/simulation, 2026.

[21] Pate A. Github repository. manchester predictive healthcare group. mrc-multi-outcome-project-6-calibration-plots-for-multistate- risk-prediction-models. https://github.com/manchester-predictive-healthcare-group/CHI-MRC-multi-outcome/blob/main/Project

[22] Hein Putter, Jos van der Hage, Geertruida Bock, and Rachid Elgalta. Estimation and prediction in a multistate model for breast cancer. *Biometrical Journal - BIOM J*, 48:366–380, 06 2006.

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

Appendix A

True transition probabilities

Denote by $h_{ij}(t)$, $h_i(t)$, $H_{ij}(t)$, $H_i(t)$ and $S_i(t)$ the transition hazards, the total hazard out of state i, the cumulative hazards, the total cumulative hazard and the survival function respectively.

Let $P_{ij}(u,t)$ be the probability of being in state j at t, conditioned by being in state i at u and $P_{ij}^{route}(u,t)$ the probability of being in state j at t, conditioned by being in state i at u and going from i to j through the states indicated in the *route*.

The true transition probabilities $P_{1k}(0, t.eval)$, $k = 2, ..., 5$ can be calculated by based on the well known relations between survival and cumulative hazard

$$S_i(t) = e^{-H_i(t)}$$

$$P_{ii}(u,t) = e^{-H_i(u,t)} = e^{-(H_i(t)-H_i(u))} = \frac{e^{-H_i(t)}}{e^{-H_i(u)}} = \frac{S_i(t)}{S_i(u)}$$

The probability of being in state j at t after entering an intermediate state k, conditioned by being in state i at u is as follows

$$P_{ij}(u,t) = \int_u^t h_{ik}(r)\frac{S_i(r)}{S_i(u)}P_{kj}(r,t)dr$$

In the next sections transitions probabilities out of states 1,2,3 and 4 are provided.

A.1 Transition probabilities out of state 4

$P_{44}(u,t) = \frac{S_4(t)}{S_4(u)}$

$P_{45}(u,t) = 1 - \frac{S_4(t)}{S_4(u)}$

A.2 Transition probabilities out of state 2

$P_{22}(u,t) = \frac{S_2(t)}{S_2(u)}$

$P_{25}^{direct}(u,t) = \int_u^t h_{25}(r)\frac{S_2(r)}{S_2(u)}dr$

$P_{25}^{4}(u,t) = \int_u^t h_{24}(r)\frac{S_2(r)}{S_2(u)}P_{45}(r,t)dr$

$P_{24}(u,t) = \int_u^t h_{24}(r)\frac{S_2(r)}{S_2(u)}P_{44}(r,t)dr$

A.3 Transition probabilities out of state 3

$P_{33}(u,t) = \frac{S_3(t)}{S_3(u)}$

$P_{35}^{direct}(u,t) = \int_u^t h_{35}(r)\frac{S_3(r)}{S_3(u)}dr$

$P_{35}^{4}(u,t) = \int_u^t h_{34}(r)\frac{S_3(r)}{S_3(u)}P_{45}(r,t)dr$

$P_{34}(u,t) = \int_u^t h_{34}(r)\frac{S_3(r)}{S_3(u)}P_{44}(r,t)dr$

A.4 Transition probabilities out of state 1

$P_{11}(u,t) = \frac{S_1(t)}{S_1(u)}$

$P_{12}(u,t) = \int_u^t h_{12}(r)\frac{S_1(r)}{S_1(u)}P_{22}(r,t)dr$

$P_{13}(u,t) = \int_u^t h_{13}(r)\frac{S_1(r)}{S_1(u)}P_{33}(r,t)dr$

$P_{14}^{2}(u,t) = \int_u^t h_{12}(r)\frac{S_1(r)}{S_1(u)}P_{24}(r,t)dr$

$P_{14}^{3}(u,t) = \int_u^t h_{13}(r)\frac{S_1(r)}{S_1(u)}P_{34}(r,t)dr$

$P_{14}(u,t) = P_{14}^{2}(u,t) + P_{14}^{3}(u,t)$

$P_{15}^{direct}(u,t) = \int_u^t h_{15}(r)\frac{S_1(r)}{S_1(u)}dr$

$P_{15}^{2}(u,t) = \int_u^t h_{12}(r)\frac{S_1(r)}{S_1(u)}P_{25}^{direct}(r,t)dr$

$P_{15}^{24}(u,t) = \int_u^t h_{12}(r)\frac{S_1(r)}{S_1(u)}P_{25}^{4}(r,t)dr$

$P_{15}^{3}(u,t) = \int_u^t h_{13}(r)\frac{S_1(r)}{S_1(u)}P_{35}^{direct}(r,t)dr$

$P_{15}^{34}(u,t) = \int_u^t h_{13}(r)\frac{S_1(r)}{S_1(u)}P_{35}^{4}(r,t)dr$

$P_{15}(u,t) = P_{15}^{direct}(u,t) + P_{15}^{2}(u,t) + P_{15}^{24}(u,t) + P_{15}^{3}(u,t) + P_{15}^{34}(u,t)$

The true transition probabilities of interest in the model are $P_{11}(0, t.eval)$, $P_{12}(0, t.eval)$, $P_{13}(0, t.eval)$, $P_{14}(0, t.eval)$, $P_{15}(0, t.eval)$.