




calibmsm: An R package for calibration plots of the transition probabilities from a multistate model

Alexander Pate 
University of Manchester

Matthew Sperrin 
University of Manchester

Richard D. Riley 
University of Birmingham

Ben Van Calster 
Leiden University
Medical Center

Glen P. Martin 
University of Manchester

Abstract

Multistate models, which allow the prediction of complex multistate survival processes such as multimorbidity, or recovery, relapse and death following treatment for cancer, are beginning to be used for clinical prediction. It is paramount to evaluate the calibration (as well as other metrics) of a risk prediction model before implementation of the model. Currently no software exists to aid in assessing the calibration of a multistate model. **calibmsm** has been developed to fill this gap, providing easy to use software for model developers. Calibration of the transition probabilities between given follow up times is made possible through three approaches. The first two utilise calibration techniques for binary and multinomial logistic regression models in combination with inverse probability of censoring weights, whereas the third utilises psuedo-values. All methods are implemented in conjunction with landmarking to allow calibration assessment of predictions made at any time beyond the start of follow up.

This article details the methodology and provides a comprehensive example on how to use **calibmsm** to assess the calibration of a multistate model developed to predict recovery, adverse events, relapse and survival in patients with blood cancer after a transplantation. **calibmsm** could be used to assess the calibration of predicted risks from a range of other models, including: dynamic models and landmark supermodels which utilise information post baseline to update predictions, competing risks models and standard single outcome survival models, where predictions can be made at any landmark time.

Keywords: clinical prediction, calibration, validation, multistate, multi-state, R.

1. Introduction

Risk prediction models enable the prediction of clinical events in either diagnostic or prognostic settings (van Smeden *et al.* 2021) and are used widely to inform clinical practice. A multistate model (Putter *et al.* 2007) may be used when there are multiple outcomes of interest, or when a single outcome of interest may be reached via intermediate states. For example, prediction of death after local recurrence or distant metastasis in patients with breast cancer following surgery (Putter *et al.* 2006); prediction of death following progression of chronic kidney disease (Lintu *et al.* 2022); prediction of non-AIDS events and death in individuals living with HIV (Masia *et al.* 2017). Using a multistate model for prediction is important when the development of an intermediate condition occurring post index date may have an impact on the risk of future outcomes of interest. Risk prediction models developed for use in clinical practice should be evaluated in a relevant cohort, or preferably multiple settings/cohorts, prior to implementation (Steyerberg and Harrell Jr 2016). If the intended use of this model is known, targeted validation in a specific setting may be preferred (Sperrin *et al.* 2022). A key part of the validation process is assessment of the calibration of the model (Van Calster *et al.* 2019). Calibration assesses whether the predicted risks match the observed event rates in the cohort of interest. Ideally calibration curves should be produced, which estimate observed event rates as a function the predicted risks over the entire distribution of predicted risk. This corresponds to a moderate assessment of calibration (Van Calster *et al.* 2016).

The R (R Core Team 2023) package **mstate** (de Wreede *et al.* 2011) provides a comprehensive set of tools to develop a multistate model for a continuously observed multistate survival process. However, currently no software exists to aid researchers in assessing the calibration of a multistate model that has been developed for the purposes of individual risk prediction. The R package **calibmsm** has been developed to enable researchers to estimate calibration curves and scatter plots using three approaches outlined in Pate *et al.* (2023), which focused on assessing the calibration of the transition probabilities out of the starting state. The work in this paper extends the framework to assess the calibration of transition probabilities out of any state j at any time s using landmarking (van Houwelingen 2007; Dafni 2011), provides more details on estimation of the inverse-probability of censoring weights (where relevant), and demonstrates the process for estimating confidence intervals. **calibmsm** is available from the Comprehensive R Archive Network at (<https://CRAN.R-project.org/package=calibmsm>).

de Wreede *et al.* (2011) used data from the European Society for Blood and Marrow Transplantation (EBMT 2023) to showcase how to develop a multistate model for clinical prediction of outcomes after bone marrow transplantation in leukemia patients (Figure 1). In this study, we show how to assess the calibration of a model developed on the same EBMT data as a way of illustrating the syntax and workflows of **calibmsm**. This clinical example also highlights some important differences between the methods in how they deal with informative censoring and computational feasibility, which may impact future uptake of the methods. Details on the methodology are given in section 2. The clinical example, including steps for data preparation and production of calibration plots are given in section 3. Section 4 contains a discussion and summary.

2. Methodology

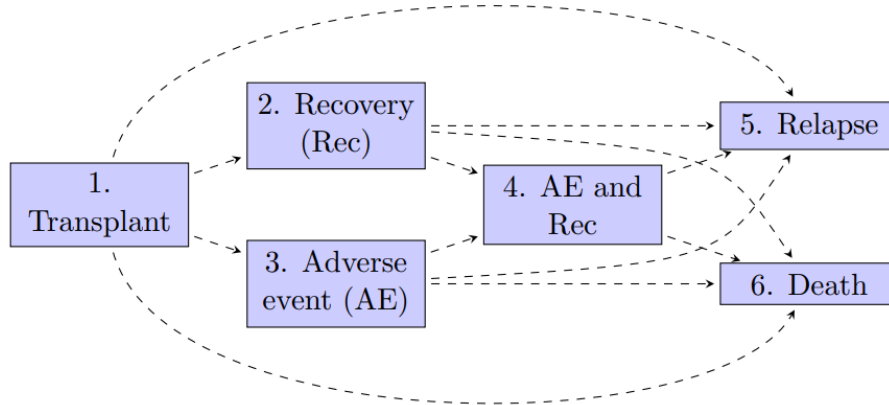


Figure 1: A six-state model for leukemia patients after bone marrow transplantation. Figure taken from (de Wreede *et al.* 2011).

2.1. Setup

Let $X(t) \in \{1, \dots, K\}$ be a multistate survival process with K states. We assume a multistate model has already been developed and we want to assess the calibration of the predicted transition probabilities, $\hat{p}_{j,k}(s, t)$, in a cohort of interest. The transition probabilities are the probability of being in state k at time t , if in state j at time s , where $s < t$. To assess the calibration of the multistate model, we must estimate observed event probabilities:

$$o_{j,k}(s, t) = P[X(t) = k | X(s) = j, \hat{p}_{j,k}(s, t)].$$

In a well calibrated model, the transition probabilities will be equal to the observed event probabilities.

In the absence of censoring, $o_{j,k}(s, t)$ can be estimated using cross sectional calibration techniques in a landmark (van Houwelingen 2007; Dafni 2011) cohort of individuals who are in state j at time s (i.e. methods to assess the calibration of models predicting binary or multinomial outcomes). In the presence of censoring, calibration must be assessed in this landmark cohort of individuals either using these cross sectional techniques in combination with inverse probability of censoring weights, or through pseudo-values. These approaches are detailed in sections 2.2 - 2.6.

2.2. Binary logistic regression with inverse probability of censoring weights (BLR-IPCW) calibration curves

The first approach produces calibration curves using a framework for binary logistic regression models in conjunction with inverse probability of censoring weights to account for informative censoring. Let $I_k(t)$ be an indicator for whether an individual is in state k at time t . $I_k(t)$ is then modeled using a flexible approach with $\hat{p}_{j,k}(s, t)$ as the sole predictor. This model is fit in the landmark cohort (in state j at time s) of individuals who are also still uncensored at time t . This cohort is weighted using inverse probability of censoring weights (see section 2.4). We suggest using a loess smoother (Austin and Steyerberg 2014):

$$I_k(t) = \text{loess}(\hat{p}_{j,k}(s, t)), \quad (1)$$

or a logistic regression model with restricted cubic splines (Harrell 2015):

$$\text{logit}(I_k(t)) = \text{rcs}(\text{logit}(\hat{p}_{j,k}(s, t))). \quad (2)$$

Any flexible model for binary outcomes could be used, but these are the most common and are implemented in this package. Observed event probabilities $\hat{o}_{j,k}(s, t)$ are then estimated as fitted values from these models. The calibration curve is plotted using the set of points $\{\hat{p}_{j,k}(s, t), \hat{o}_{j,k}(s, t)\}$. To obtain unbiased calibration curves, the assumption that each outcome $I_k(t)$ is independent from the censoring mechanism in the reweighted population must hold.

2.3. Multinomial logistic regression with inverse probability of censoring weights (MLR-IPCW) calibration scatter plots

The second approach produces calibration scatter plots using a framework for multinomial logistic regression models with inverse probability of censoring weights (MLR-IPCW). Let $I_X(t)$ be a multinomial indicator variable taking values $I_X(t) \in \{1, \dots, K\}$ such that $I_X(t) = k$ if an individual is in state k at time t . The nominal recalibration framework of Van Hoorde *et al.* (2014, 2015) is then applied in the landmark cohort of individuals uncensored at time t , weighted using inverse probability of censoring weights (section 2.4). First calculate the log-ratios of the predicted transition probabilities:

$$\hat{L}P_k = \ln \left(\frac{\hat{p}_{j,k}(s, t)}{\hat{p}_{j,k_{ref}}(s, t)} \right),$$

Then fit the following multinomial logistic regression model:

$$\ln \left(\frac{P[I_X(t) = k]}{P[I_X(t) = k_{ref}]} \right) = \alpha_k + \sum_{h=2}^K \beta_{k,h} * s_k(\hat{L}P_h), \quad (3)$$

where k_{ref} is an arbitrary reference category which can be reached from state j , $k \neq k_{ref}$ takes values in the set of states that can be reached from state j , and where s is a vector spline smoother (Yee 2015). Observed event probabilities $\hat{o}_{j,k}(s, t)$ are then estimated as fitted values from this model. This results in a calibration scatter plot rather than a curve due to all states being modeled simultaneously, as opposed to BLR-IPCW, which is a "one vs all" approach. The scatter occurs because the observed event probabilities for state k vary depending on the predicted transition probabilities of the other states. This is a stronger (Van Calster *et al.* 2016) form of calibration than that evaluated by BLR-IPCW, and will also result in observed event probabilities which sum to 1. In future iterations of **calibmsm** functionality will be added to produce smoothed curves estimated from these scatter plots. To obtain unbiased calibration curves, the assumption that the outcome $I_X(t)$ is independent from the censoring mechanism in the reweighted population must hold.

2.4. Estimation of the inverse probability of censoring weights

The estimand for the weights is $w_j(s, t)$, the inverse of the probability of being uncensored at time t if in state j at time s :

$$w_j(s, t) = \frac{1}{P[t_{cens} > t | t > s, X(s) = j, \mathbf{Z}, \mathbf{X}(t)]},$$

where $\mathbf{X}(t)$ denotes the history of the multistate survival process up to time t , including the transition times, and \mathbf{Z} is a set of baseline predictor variables believed to be predictive of the censoring mechanism. Note that \mathbf{Z} may be the same as, but is not restricted to, the variables used for prediction when developing the multistate model. First the estimator $\hat{P}[t_{cens} > t | t > s, X(s) = j, \mathbf{Z}]$ is calculated by developing an appropriate survival model. The outcome in this model is the time until censoring occurs. Moving into an absorbing state prevents censoring from happening and is treated as a censoring mechanism in this model (i.e. a competing risks approach is not taken when fitting this model). $\mathbf{X}(t)$ is explicitly conditioned on when defining $w_j(s, t)$ because the weights must reflect that censoring can no longer be observed for an individual if they enter an absorbing state at some time $s < t_{abs} < t$. Therefore

$$\hat{P}[t_{cens} > t | t > s, X(s) = j, \mathbf{Z}, \mathbf{X}(t)] = \hat{P}[t_{cens} > \min\{t, t_{abs}\} | t > s, X(s) = j, \mathbf{Z}]$$

In **calibmsm**, unless otherwise specified, $\hat{P}[t_{cens} > t | t > s, X(s) = j, \mathbf{Z}]$ is estimated using a cox proportional hazards model where all predictors \mathbf{Z} are assumed to have a linear effect on the log-hazard. This is highly restrictive, users can therefore also input their own vector of weights, which is strongly recommended. Given the BLR-IPCW and MLR-IPCW approaches are both reliant on correct estimation of the weights, we encourage users to take the time to carefully estimate the inverse probability of censoring weights using a well specified model. The limitations of using the **calibmsm** internal functions for estimating the weights in this clinical example (section ??) are discussed in more detail later, and explored in [vignette-Evaluation-of-estimation-of-IPCWs](#).

Stabilised weights can be estimated by multiplying by the weights $w_j(s, t)$ by the mean probability of being uncensored:

$$w_j^{stab}(s, t) = \frac{P[t_{cens} > t | t > s, X(s) = j]}{P[t_{cens} > t | t > s, X(s) = j, \mathbf{Z}, \mathbf{X}(t)]}.$$

The numerator can be estimated using an intercept only model, and note there is no dependence on $\mathbf{X}(t)$.

Another option is to estimate $w(s, t)$, which is the inverse of the probability of being uncensored at time t if uncensored at time s :

$$w(s, t) = \frac{1}{P[t_{cens} > t | t > s, \mathbf{Z}, \mathbf{X}(t)]}.$$

This can be estimated using the same approach as for $w_j(s, t)$, except there is no requirement to be in state j when landmarking at time s . If the censoring mechanism is non-informative after conditioning on \mathbf{Z} , then $w(s, t) = w_j(s, t)$, and any consistent estimator for $w(s, t)$ will be a consistent estimator of $w_j(s, t)$. The advantage is that $\hat{w}(s, t)$ is calculated by developing a model in the cohort of individuals uncensored at time s , which is a larger cohort than those

uncensored and in state j at time s . Therefore $\hat{w}(s, t)$ will be a more precise estimator than $\hat{w}_j(s, t)$. On the contrary, if the assumption of non-informative censoring after conditioning on \mathbf{Z} does not hold, there is a risk of bias in estimation of the weights. We therefore recommend using the estimator $w_j(s, t)$ unless sample size (number of individuals in state j at time s) is low, which may be assessed using sample size formula for prediction models with time-to-event outcomes (Riley *et al.* 2019). If the sample size is deemed insufficient, one may consider using $w(s, t)$, but the risk of bias associated with this estimator must be carefully considered. Finally, we state the importance of using inverse probability of censoring weights, even if the censoring mechanism is believed to be completely non-informative (i.e. happens at random). All multistate models must have an absorbing state, entry into which prevents censoring from happening. This induces a dependence between the outcome and the censoring mechanism which must be adjusted for using inverse probability of censoring weights. This issue was highlighted in the supplementary material of previous work (Pate *et al.* 2023)

2.5. Pseudo-value calibration plots

The third approach produces calibration curves using pseudo-values (Andersen and Pohar Perme 2010; Andersen *et al.* 2022). Pseudo-values can be used in place of the outcome of interest in a regression model if some outcomes are not observed due to right censoring. This is the case in models (1) and (2). For certain estimators $\hat{\theta}$ (where θ estimates the expectation of the outcome it is replacing), the pseudo-value for individual i is defined as:

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

where $\hat{\theta}^{-i}$ is equal to $\hat{\theta}$ estimated in a cohort without individual i . One such estimator for the outcomes in models (1) and (2) given the underlying multistate survival process, is the Landmark Aalen-Johansen estimator (Putter and Spitoni 2018), which estimates the expectation of $I_k(t)$ in the landmark cohort of individuals in which calibration is being assessed. The resulting pseudo-values are a vector with K elements, one for each possible transition, for every individual i . These pseudo-values can replace the outcome $I_k(t)$ in equations (1) and (2) in order to estimate $o_{j,k}(s, t)$.

Pseudo-values are based on the same assumptions as the underlying estimator $\hat{\theta}$. The Landmark Aalen-Johansen estimator is valid for both Markov and non-Markov multistate models. However, it does make the assumption that the multistate survival process and the censoring distribution are independent (uninformative censoring). The approach to alleviate this is to estimate the pseudo-values within sub-groups of individuals, now making the assumption that censoring is non-informative within the specified subgroups. This can be done by calculating the pseudo-values within subgroups defined by baseline predictors, or subgroups defined by the predicted transition probabilities $\hat{p}_{j,k}(s, t)$. Both options are implemented in this package. When pseudo-values are calculated within subgroups, they are still used as the outcome in models (1) and (2) in the same way. Note that the pseudo-values $\hat{\theta}^i$ are continuous, as opposed to binary $I_k(t)$, but the link function in model (2) remains the same to ensure $\hat{o}_{j,k}(s, t)$ are between zero and one.

2.6. Estimation of confidence intervals

Confidence intervals for both BLR-IPCW and pseudo-value calibration curves can be estimated using bootstrapping. While theoretically feasible, it is currently unclear how to present confidence intervals for each data point in the calibration scatter plots produced by MLR-IPCW, and therefore these are omitted. A process for estimating the confidence intervals around the BLR-IPCW calibration curves is as follows:

1. Resample validation dataset with replacement
2. Landmark the dataset for assessment of calibration
3. Calculate inverse probability of censoring weights
4. Fit the preferred calibration model in the landmarked dataset (restricted cubic splines or loess smoother)
5. Generate observed event probabilities for a fixed vector of predicted transition probabilities (specifically the predicted transition probabilities from the non-bootstrapped landmark validation dataset)

This will produce a number of bootstrapped calibration curves, all plotted over the same vectors of predicted transition probabilities. Taking the $\frac{\alpha}{2}$ and $(1 - \frac{\alpha}{2})$ percentiles of the observed event probabilities for each predicted transition probability gives the required $1 - \alpha$ confidence interval around the estimated calibration curve. To estimate confidence intervals for the pseudo-value calibration curves using bootstrapping, the same procedure is applied except the third step is replaced with 'calculate the pseudo-values within the landmarked bootstrapped dataset'. This will be highly computationally demanding as the pseudo-values must be estimated in every bootstrap dataset.

If using the pseudo-value method, confidence intervals can however be calculated using closed form estimates of the standard error when making predictions of the observed event probabilities (i.e. when obtaining fitted values from models (1) or (2)). We recommended this due to the computational burden of bootstrapping the confidence intervals around the pseudo-value calibration curves. There are a number of issues with estimating parametric confidence intervals for the BLR-IPCW calibration curves. Firstly, a robust sandwich-type estimator should be used to estimate the standard error [Hernan and Robins \(2020\)](#), which are known to result in conservative confidence intervals, i.e. too large [Hernan and Robins \(2020\)](#); [Austin et al. \(2020\)](#). On the contrary, the size of the confidence interval will be underestimated as uncertainty in estimation of the weights is not considered. Due to the impact of these two factors, we recommend using bootstrapping to estimate the confidence intervals for BLR-IPCW calibration curves.

3. Clinical example

3.1. Clinical setting and data preparation

We utilise data from the European Society for Blood and Marrow Transplantation ([EBMT 2023](#)), containing multistate survival data after a transplant for patients with blood cancer. The start of follow up is the day of the transplant and the initial state is alive and in remission.

There are three intermediate events (2: recovery, 3: adverse event, or 4: recovery + adverse event), and two absorbing states (5: relapse and 6: death). This data is available from the **mstate** package (de Wreede *et al.* 2011). We assume the user of **calibmsm** has experience with handling the type of data used to develop a multistate model as outlined by de Wreede *et al.* (2011).

Four datasets are provided to enable assessment of a multistate model fitted to these data. The code for deriving all these datasets is provided in the source code for **calibmsm**. The first is **ebmtcal**, which is the same as the **ebmt** dataset provided in **mstate**, with two extra variables derived: time until censoring (**dtcens**) and an indicator for whether censoring was observed (**dtcens.s** = 1) or an absorbing state was entered (**dtcens.s** = 0). This dataset contains baseline information on year of transplant (**year**), age at transplant (**age**), prophylaxis given (**proph**), and whether the donor was gender matched (**match**). The second dataset provided is **msebmcal**, which is the **ebmt** dataset converted into a dataset of class **msdata** using the processes and functions in the package **mstate** (de Wreede *et al.* 2011). It contains all transition times, an event indicator for each transition, as well as a **trans** attribute containing the transition matrix.

```
R> library("calibmsm")
R> data("ebmtcal")
R> head(ebmtcal)
```

	id	rec	rec.s	ae	ae.s	recae	recae.s	rel	rel.s	srv	srv.s	
	1	1	22	1	995	0	995	0	995	0	995	0
	2	2	29	1	12	1	29	1	422	1	579	1
	3	3	1264	0	27	1	1264	0	1264	0	1264	0
	4	4	50	1	42	1	50	1	84	1	117	1
	5	5	22	1	1133	0	1133	0	114	1	1133	0
	6	6	33	1	27	1	33	1	1427	0	1427	0

	year	agecl	proph	match	dtcens	dtcens.s
1	1995-1998	20-40	no no	gender mismatch	995	1
2	1995-1998	20-40	no no	gender mismatch	422	0
3	1995-1998	20-40	no no	gender mismatch	1264	1
4	1995-1998	20-40	no	gender mismatch	84	0
5	1995-1998	>40	no	gender mismatch	114	0
6	1995-1998	20-40	no no	gender mismatch	1427	1

```
R> data("msebmcal")
R> subset(msebmcal, id %in% c(1,2,3))
```

	id	from	to	trans	Tstart	Tstop	time	status
1	1	1	2	1	0	22	22	1
2	1	1	3	2	0	22	22	0
3	1	1	5	3	0	22	22	0
4	1	1	6	4	0	22	22	0
5	1	2	4	5	22	995	973	0
6	1	2	5	6	22	995	973	0

7	1	2	6	7	22	995	973	0
8	2	1	2	1	0	12	12	0
9	2	1	3	2	0	12	12	1
10	2	1	5	3	0	12	12	0
11	2	1	6	4	0	12	12	0
12	2	3	4	8	12	29	17	1
13	2	3	5	9	12	29	17	0
14	2	3	6	10	12	29	17	0
15	2	4	5	11	29	422	393	1
16	2	4	6	12	29	422	393	0
17	3	1	2	1	0	27	27	0
18	3	1	3	2	0	27	27	1
19	3	1	5	3	0	27	27	0
20	3	1	6	4	0	27	27	0
21	3	3	4	8	27	1264	1237	0
22	3	3	5	9	27	1264	1237	0
23	3	3	6	10	27	1264	1237	0

In the work of [de Wreede *et al.* \(2011\)](#), the focus is on predicting transition probabilities made at times $s = 0$ and $s = 100$ days, across a range of follow up times t , and comparing prognosis for patients in different states j . In this study we also focus on assessing the calibration of the transition probabilities made at these times. We assess calibration of the transition probabilities at $t = 5$ years, a common follow up time for cancer prognosis, but calibration of the model may vary for other values of t . We estimate transition probabilities for each individual by developing a model as demonstrated in [de Wreede *et al.* \(2011\)](#), following the theory of [Putter *et al.* \(2007\)](#).

The predicted transitions probabilities from each state j at times $s = 0$ and $s = 100$ are contained in stacked datasets `tps0` and `tps100` respectively. A leave-one-out approach was used when estimating these transition probabilities. This means each individual was removed from the development dataset when fitting the multistate model to estimate their transition probabilities. This approach allows validation to be assessed in the same dataset that the model was developed with minimal levels of in-sample optimism. Note that for `tps100` the predicted probabilities for some states k are equal to 0. This is because no individuals in state $j = 1$ at time $s = 100$ transition into states 3 or 4. This may be due to the definition of an adverse event having to occur within a certain number of days post transplant.

```
R> data("tps0")
R> head(tps0)
```

	id	pstate1	pstate2	pstate3	pstate4	pstate5	pstate6
1	1	0.1139726	0.2295006	0.08450376	0.2326861	0.1504855	0.1888514
2	2	0.1140189	0.2316569	0.08442692	0.2328398	0.1481977	0.1888598
3	3	0.1136646	0.2317636	0.08274331	0.2325663	0.1504787	0.1887834
4	4	0.1383878	0.1836189	0.07579429	0.2179331	0.1538475	0.2304185
5	5	0.1233226	0.1609740	0.05508100	0.1828176	0.1425950	0.3352099
6	6	0.1136646	0.2317636	0.08462424	0.2305854	0.1505534	0.1888087

```

      se1      se2      se3      se4      se5      se6 j
1 0.01291133 0.02369584 0.01257251 0.02323376 0.01648630 0.01601795 1
2 0.01291552 0.02374329 0.01256056 0.02324869 0.01632797 0.01603703 1
3 0.01289444 0.02375770 0.01245752 0.02322375 0.01647890 0.01601525 1
4 0.01857439 0.03004447 0.01462570 0.03018673 0.02124071 0.02416121 1
5 0.01944967 0.03419721 0.01367768 0.03423941 0.02329644 0.03688586 1
6 0.01289444 0.02375770 0.01257276 0.02317348 0.01649531 0.01602438 1

```

```
R> data("tps100")
```

```
R> head(tps100)
```

```

  id  pstate1  pstate2 pstate3 pstate4  pstate5  pstate6
1  1 0.7013881 0.05239271      0      0 0.1408120 0.1054072
2  2 0.7012745 0.05261136      0      0 0.1407625 0.1053516
3  3 0.7011368 0.05270176      0      0 0.1407628 0.1053987
4  4 0.6840325 0.04139266      0      0 0.1700565 0.1045183
5  5 0.6804049 0.04308434      0      0 0.1500344 0.1264764
6  6 0.7011368 0.05270176      0      0 0.1407628 0.1053987
      se1      se2 se3 se4      se5      se6 j
1 0.04691168 0.02077138  0  0 0.03457006 0.03081258 1
2 0.04691218 0.02082871  0  0 0.03456448 0.03079617 1
3 0.04693068 0.02086917  0  0 0.03456101 0.03081033 1
4 0.05885230 0.02161973  0  0 0.04710517 0.03673242 1
5 0.06694739 0.02484634  0  0 0.04905043 0.04628088 1
6 0.04693068 0.02086917  0  0 0.03456101 0.03081033 1

```

The procedure for producing calibration plots requires the use of two functions. The first function, `calib_blr`, `calib_pv` or `calib_mlr`, calculates the data for the calibration plot using the methods described in section 2. The second function, `plot`, produces the plots. `plot` is an S3 generic written for objects of class `calib_blr`, `calib_mlr` or `calib_pv`, and produces the calibration plots using `ggplot2` (Wickham 2016). Separating these processes allows users to manually estimate bootstrapped calibration curves (see [vignette-BLR-IPCW-manual-bootstrap](#)) using the output from `calib_blr`, `calib_pv` or `calib_mlr`. It also allows users the flexibility of producing their own plots utilising the full functionality of `ggplot2`, rather than being reliant on the S3 generics provided.

The validation cohort must be provided to functions `calib_blr`, `calib_pv` and `calib_mlr` in two different formats. The `data.raw` argument requires a `data.frame` (one observation per individual) and is used to fit the calibration models. For methods BLR-IPCW and MLR-IPCW, `data.raw` should contain variables `dtcens` (censoring time) and `dtcens.s` (censoring indicator, `dtcens.s = 1` if the individual is censored at time `dtcens`, `dtcens.s = 0` otherwise), plus any baseline predictors \mathbf{Z} used to estimate the weights. For the pseudo-value approach, this dataset should contain any baseline predictors \mathbf{Z} which variables will be grouped by before calculating the pseudo-values. The `data.mstate` argument requires a dataset of class `mstate`, which is used to implement the landmarking and estimate the Aalen-Johansen estimator for the pseudo-value approach. A dataset of this class must be produced using the package `mstate` (de Wreede *et al.* 2011). Both `data.mstate` and `data.raw` should contain

corresponding patient ID variables `id`. The predicted transition probabilities out of state j at time s must then be specified through the `tp.pred` argument, which must contain a column for each transition k , even if the transition from j to k has zero probability. The rows in `tp.pred` must be ordered in the same way as those in `data.raw`. The datasets described in section 3.1 meet these criteria.

3.2. Calibration plots for the transition probabilities out of state $j = 1$ at time $s = 0$

We start by producing calibration curves for the predicted transition probabilities out of state $j = 1$ at time $s = 0$. Given all individuals start in state 1, there is no need to consider the transition probabilities out of states $j \neq 1$ at $s = 0$. Calibration is assessed at follow up time ($t = 1826$ days). We start by extracting the predicted transition probabilities from state $j = 1$ at time $s = 0$ from the object `tps0`. These are the transition probabilities we aim to assess the calibration of.

```
R> tp.pred.s0 <- tps0 |>
+   dplyr::filter(j == 1) |>
+   dplyr::select(any_of(paste("pstate", 1:6, sep = "")))
```

We first evaluate calibration using the BLR-IPCW approach, implemented through the function `calib_blr`. We choose to estimate the calibration curves using restricted cubic splines, although the use of loess smoothers would be equally valid. When using restricted cubic splines the number of knots must always be specified by the user, and 3 knots are chosen here given the reasonably small size of the dataset. Calibration curves could be estimated using the internal estimation procedure and the predictor variables `year`, `agec1`, `proph` and `match`. The `w.landmark.type` argument assigns whether weights are estimated using all individuals uncensored at time s , or only those uncensored and in state j at time s , as discussed in section 2.4. The maximum weight (`w.max = 10`) and stabilisation of weights (`stabilised = TRUE`) are left as default. Weights can also be manually specified using the `weights` argument. We request 95% confidence intervals for the calibration curves calculated through bootstrapping with 200 bootstrap replicates.

```
R> t.eval <- 1826
R> dat.calib.blr <-
+   calib_blr(data.mstate = msebmtcal,
+             data.raw = ebmtcal,
+             j=1,
+             s=0,
+             t = t.eval,
+             tp.pred = tp.pred.s0,
+             curve.type = "rcs",
+             rcs.nk = 3,
+             w.covs = c("year", "agec1", "proph", "match"),
+             CI = 95,
+             CI.R.boot = 200)
```

The first element of `dat.calib.blr` (named `plotdata`) contains 6 data frames for the calibration curves of the transition probabilities into each of the six states, $k \in \{1, 2, 3, 4, 5, 6\}$. Each data frame contains three columns, `id`: the identifier of each individual; `pred`: the predicted transition probabilities; `obs`: the observed event probabilities. The second element (named `metadata`) is a metadata argument containing information about the data and chosen calibration analysis.

```
R> str(dat.calib.blr[["plotdata"]])
```

List of 6

```
$ state1:'data.frame':      1778 obs. of  5 variables:
..$ id      : int [1:1778] 2 4 5 7 10 13 14 16 18 19 ...
..$ pred    : num [1:1778] 0.114 0.1384 0.1233 0.0974 0.1137 ...
..$ obs     : num [1:1778] 0.11 0.104 0.105 0.124 0.11 ...
..$ obs.lower: num [1:1778] 0.0908 0.0849 0.0892 0.0886 0.0909 ...
..$ obs.upper: num [1:1778] 0.133 0.13 0.125 0.165 0.133 ...
$ state2:'data.frame':      1778 obs. of  5 variables:
..$ id      : int [1:1778] 2 4 5 7 10 13 14 16 18 19 ...
..$ pred    : num [1:1778] 0.232 0.184 0.161 0.212 0.232 ...
..$ obs     : num [1:1778] 0.17 0.186 0.176 0.179 0.17 ...
..$ obs.lower: num [1:1778] 0.121 0.155 0.145 0.148 0.121 ...
..$ obs.upper: num [1:1778] 0.228 0.226 0.222 0.212 0.228 ...
$ state3:'data.frame':      1778 obs. of  5 variables:
..$ id      : int [1:1778] 2 4 5 7 10 13 14 16 18 19 ...
..$ pred    : num [1:1778] 0.0844 0.0758 0.0551 0.0615 0.0844 ...
..$ obs     : num [1:1778] 0.1249 0.1167 0.0919 0.1001 0.1248 ...
..$ obs.lower: num [1:1778] 0.0951 0.0858 0.0545 0.0666 0.0951 ...
..$ obs.upper: num [1:1778] 0.151 0.143 0.136 0.134 0.151 ...
$ state4:'data.frame':      1778 obs. of  5 variables:
..$ id      : int [1:1778] 2 4 5 7 10 13 14 16 18 19 ...
..$ pred    : num [1:1778] 0.233 0.218 0.183 0.221 0.233 ...
..$ obs     : num [1:1778] 0.243 0.224 0.185 0.228 0.243 ...
..$ obs.lower: num [1:1778] 0.195 0.191 0.159 0.191 0.195 ...
..$ obs.upper: num [1:1778] 0.284 0.256 0.22 0.261 0.284 ...
$ state5:'data.frame':      1778 obs. of  5 variables:
..$ id      : int [1:1778] 2 4 5 7 10 13 14 16 18 19 ...
..$ pred    : num [1:1778] 0.148 0.154 0.143 0.144 0.149 ...
..$ obs     : num [1:1778] 0.191 0.165 0.222 0.212 0.188 ...
..$ obs.lower: num [1:1778] 0.164 0.146 0.181 0.175 0.163 ...
..$ obs.upper: num [1:1778] 0.215 0.181 0.261 0.245 0.21 ...
$ state6:'data.frame':      1778 obs. of  5 variables:
..$ id      : int [1:1778] 2 4 5 7 10 13 14 16 18 19 ...
..$ pred    : num [1:1778] 0.189 0.23 0.335 0.264 0.189 ...
..$ obs     : num [1:1778] 0.207 0.254 0.316 0.28 0.207 ...
..$ obs.lower: num [1:1778] 0.187 0.232 0.284 0.257 0.187 ...
..$ obs.upper: num [1:1778] 0.229 0.28 0.351 0.304 0.229 ...
```

```
R> str(dat.calib.blr[["metadata"]])
```

```
R> plot(dat.calib.blr, combine = TRUE, nrow = 2, ncol = 3)
```

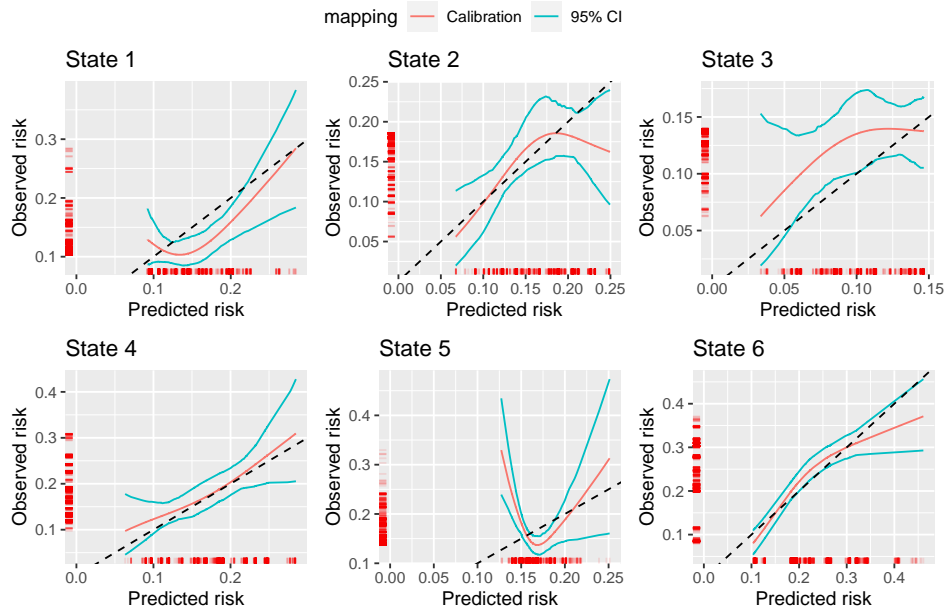


Figure 2: BLR-IPCW calibration curves out of state $j = 1$ at time $s = 0$.

List of 8

```
$ valid.transitions : num [1:6] 1 2 3 4 5 6
$ assessed.transitions: num [1:6] 1 2 3 4 5 6
$ CI : num 95
$ CI.R.boot : num 200
$ curve.type : chr "rcs"
$ j : num 1
$ s : num 0
$ t : num 1826
```

Calibration curves can then be generated using `plot`. The calibration curves (Figure 2) indicate the level of calibration is different for the transition probabilities into each of the different states. The calibration into states 4 and 6 looks the best. State 2 has good calibration over the majority of the predicted risks but over predicts for individuals with the highest predicted risks. Transition probabilities into states 1 and 3 are over and under predicted respectively over most of the range of predicted risks. Importantly the calibration of the transition probabilities into state 5 (Relapse), a key clinical outcome in this clinical setting, is extremely poor. This could be driven by errors in any of the intermediate competing risks models out of states 1, 2, 3 and 4, which all contribute to the predicted transition probabilities into state 5. Further methodological development is required in order to pin down which of the competing risk sub-models may be driving poor calibration in the transition probabilities from a multistate model.

Next we use the pseudo-value approach to assess calibration, implemented through the function `calib_pv`. Instead of specifying how the weights are estimated, we now specify variables

to define groups within which pseudo-values will be calculated (see section 2.5). The goal is to induce uninformative censoring within the chosen subgroups. We chose to calculate pseudo-values in individuals with the same year of transplant (`group.vars = c("year")`), and then split individuals into a further three groups defined by their predicted risk (`n.pctls = 3`). The number of percentiles should be increased in bigger validation datasets, although guidance on specific numbers is currently lacking. Year of transplant was identified as a subgrouping variable because a later transplant resulted in a shorter possible follow up, an earlier administrative censoring time, and it was therefore highly predictive of being censored. Your data should be explored to identify appropriate variables for subgrouping (see [vignette-Evaluation-of-estimation-of-IPCWs](#)). A parametric confidence interval is estimated as recommended in section 2.6.

```
R> dat.calib.pv <-
+   calib_pv(data.mstate = msebmtcal,
+           data.raw = ebmtcal,
+           j=1,
+           s=0,
+           t = t.eval,
+           tp.pred = tp.pred.s0,
+           curve.type = "rcs",
+           rcs.nk = 3,
+           group.vars = c("year"),
+           n.pctls = 3,
+           CI = 95,
+           CI.type = "parametric")
```

Calibration curves were then generated using `plot`. The pseudo-value calibration curves (Figure 3) are largely similar to the BLR-IPCW calibration curves (Figure 2). The agreement in the calibration curves from two completely distinct methods provides reassurance the assessment of calibration is correct. This is with the exception of state $k = 3$, where the pseudo-value calibration plot indicates the transition probabilities are well calibrated, but the BLR-IPCW calibration plot indicates the transition probabilities under predict. In a situation like this, we recommend testing the assumptions made by each of the methods to try and diagnose which are most likely to hold, and what may be driving the difference, and . In this particular example, we hypothesised that the model for estimating the inverse probability of censoring weights may be misspecified due to the strong effect of year of transplant on the censoring mechanism. We explored this theory in more detail (see [vignette-Evaluation-of-estimation-of-IPCWs](#)), and concluded that the BLR-IPCW calibration curves may be biased in this particular clinical example due to incorrect estimation of the weights.

Next we use the MLR-IPCW to evaluate calibration which produces a calibration scatter plot. This is done using the `calib_mlr` function, which has the same inputs as `calib_blr`.

```
R> dat.calib.mlr <-
+   calib_mlr(data.mstate = msebmtcal,
+           data.raw = ebmtcal,
+           j=1,
```

```
R> plot(dat.calib.pv, combine = TRUE, nrow = 2, ncol = 3)
```

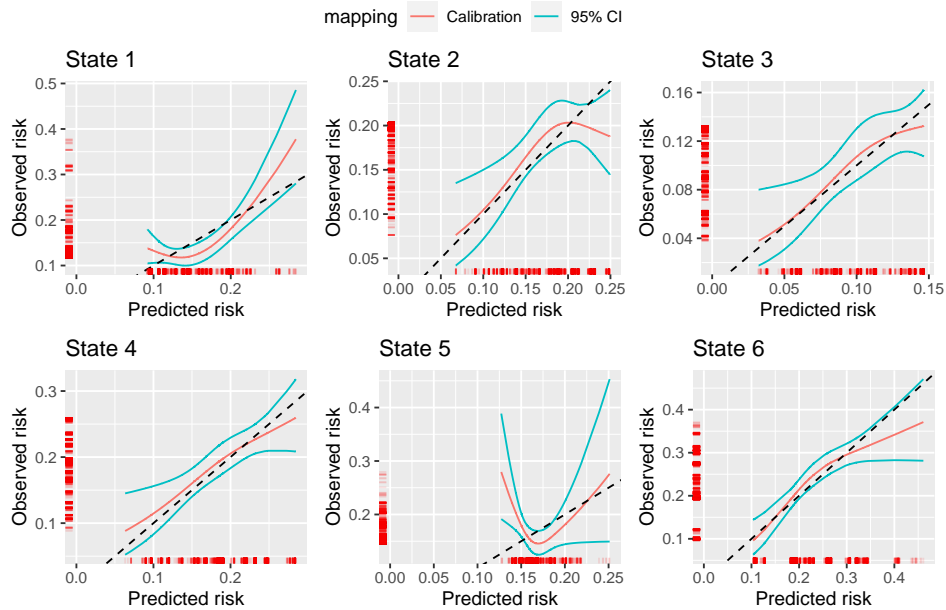


Figure 3: Pseudo-value calibration curves out of state $j = 1$ at time $s = 0$.

```
+           s=0,
+           t = 1826,
+           tp.pred = tp.pred.s0,
+           w.covs = c("year", "agecl", "proph", "match"))
```

The MLR-IPCW calibration scatter plots, produced using `plot` are contained in Figure 4. Within each plot for state k , there is a large amount of variation in calibration of the transition probabilities depending on the predicted transition probabilities into states $\neq k$. One valuable insight from these plots is that the variance in the calibration of the transition probabilities into state 6, is considerably smaller than that of state 4, despite these two states both having good calibration according to the BLR-IPCW plots (arguably state 4 looked better calibrated). This means the calibration of the transition probabilities into state 6 remains reasonably consistent, irrespective of the risks of the other states. On the contrary, the calibration of the predicted transition probabilities into state 4 is more highly dependent on the predicted transition probabilities of the other states. This insight can be gained because MLR-IPCW is a stronger (Van Calster *et al.* 2016) form of calibration assessment than the BLR-IPCW and pseudo-value approaches.

3.3. Calibration plots for the transition probabilities out of states $j = 1$ and 3 at time $s = 100$

In the work of de Wreede *et al.* (2011) focus then shifts to comparing transition probabilities when $s = 100$ depending on whether an individual has had an adverse event (state 3) or remains in state 1 (post transplant). Our focus therefore now shifts to assessing the calibration of these transition probabilities. This is done through landmarking as described in section 2.

```
R> plot(dat.calib.mlr, combine = TRUE)
```

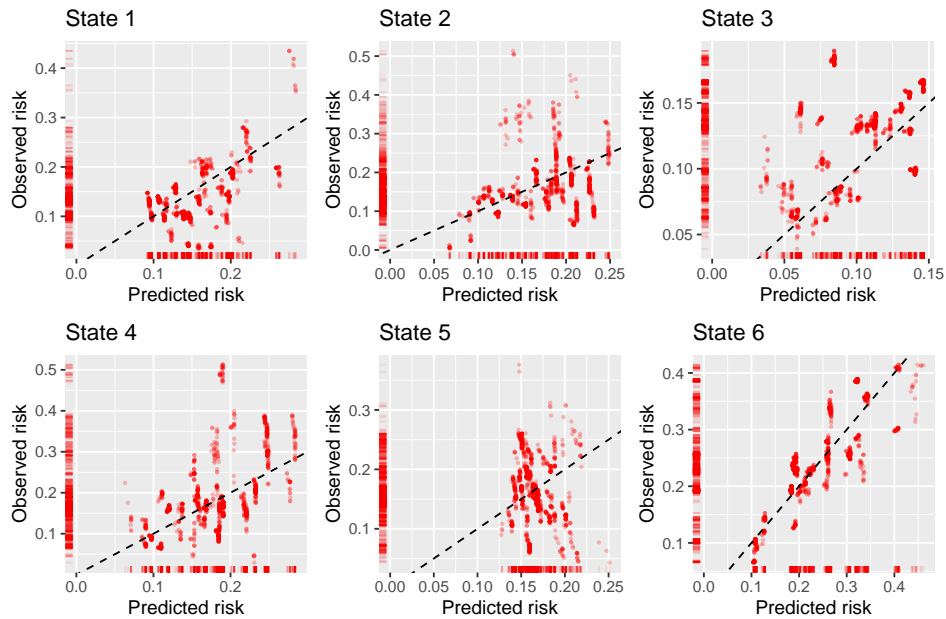


Figure 4: MLR-IPCW calibration scatter plots out of state $j = 1$ at time $s = 0$.

We start by extracting the predicted transition probabilities from state $j = 1$ and 3 at time $s = 100$ from the object `tps100`. These are the transition probabilities we aim to assess the calibration of.

```
R> tp.pred.j1s100 <- tps100 |>
+   dplyr::filter(j == 1) |>
+   dplyr::select(any_of(paste("pstate", 1:6, sep = "")))
R> tp.pred.j3s100 <- tps100 |>
+   dplyr::filter(j == 3) |>
+   dplyr::select(any_of(paste("pstate", 1:6, sep = "")))
```

The process for estimating the calibration curves remains the same, changing the inputted values j and s , and specifying the appropriate predicted transition probabilities to the argument `tp.pred`. We start by producing the calibration plots for $j = 1$ and $s = 100$ using the BLR-IPCW (Figure 5) and pseudo-value (Figure 6) methods. Given the small number of data points in this analysis induced by landmarking, we do not produce calibration scatter plots using MLR-IPCW, which may be misleading given the lack of confidence intervals.

```
R> ### Calibration using BLR-IPCW
R> dat.calib.blr.j1.s100 <-
+   calib_blr(data.mstate = msebmtcal,
+             data.raw = ebmtcal,
+             j=1,
+             s=100,
```



```

+           t = t.eval,
+           tp.pred = tp.pred.j1s100,
+           curve.type = "rcs",
+           rcs.nk = 3,
+           w.covs = c("year", "agecl", "proph", "match"),
+           CI = 95,
+           CI.R.boot = 200)
R> ### Calibration using pseudo-values
R> dat.calib.pv.j1.s100 <-
+   calib_pv(data.mstate = msebmtcal,
+            data.raw = ebmtcal,
+            j=1,
+            s=100,
+            t = t.eval,
+            tp.pred = tp.pred.j1s100,
+            curve.type = "rcs",
+            rcs.nk = 3,
+            group.vars = c("year"),
+            CI = 95,
+            CI.type = "parametric")

```

There are only four calibration plots because no individuals in state $j = 1$ at time $s = 100$ are in states $k = 3$ (adverse event) or $k = 4$ (recovery + adverse event) after $t = 1826$ days. We believe this is due to the definition of an adverse event occurring within 100 days, but as secondary users of the data, cannot be sure about this. The calibration of the predicted transition probabilities is very poor. Only for state $k = 6$ is the observed risk a monotonically increasing function of the predicted transition probabilities. We follow this up with the pseudo-value calibration plots (Figure 6) which leads to similar conclusions, as again only state $k = 6$ has a monotonically increasing calibration curve. The confidence intervals are very large. For states $k = 2$ and $k = 5$, we cannot rule out that the poor calibration is a result of sampling variation as opposed to a poorly performing prediction model. A larger validation dataset would be required to get to the bottom of this. There is a major issue with the calibration of the transition probabilities of staying in state 1, as the predicted risk is inversely proportional to the observed event rate.

Next we produce calibration plots for $j = 3$ and $s = 100$ using the BLR-IPCW (Figure 7) and pseudo-value (Figure 8) methods.

```

R> ### Calibration using BLR-IPCW
R> dat.calib.blr.j3.s100 <-
+   calib_blr(data.mstate = msebmtcal,
+            data.raw = ebmtcal,
+            j=3,
+            s=100,
+            t = t.eval,
+            tp.pred = tp.pred.j3s100,
+            curve.type = "rcs",
+            rcs.nk = 3,

```

```
R> plot(dat.calib.blr.j1.s100, combine = TRUE, nrow = 2, ncol = 2)
```

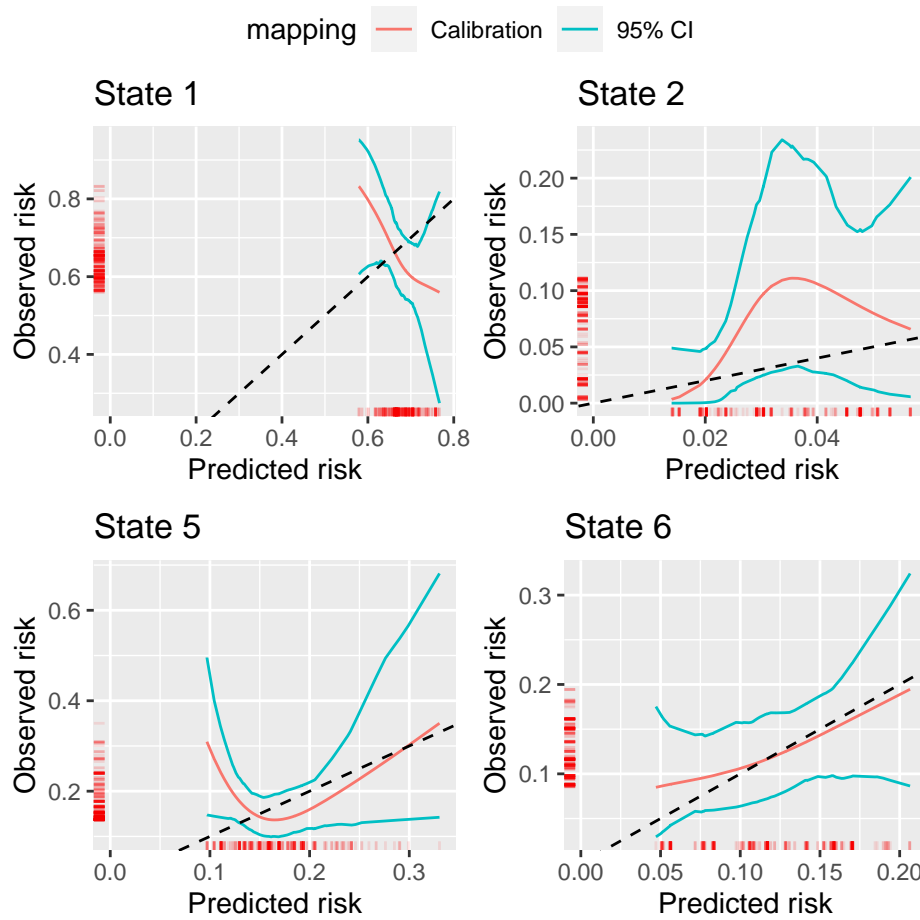


Figure 5: BLR-IPCW calibration curves out of state $j = 1$ at time $s = 100$.

```
R> plot(dat.calib.pv.j1.s100, combine = TRUE, nrow = 2, ncol = 2)
```

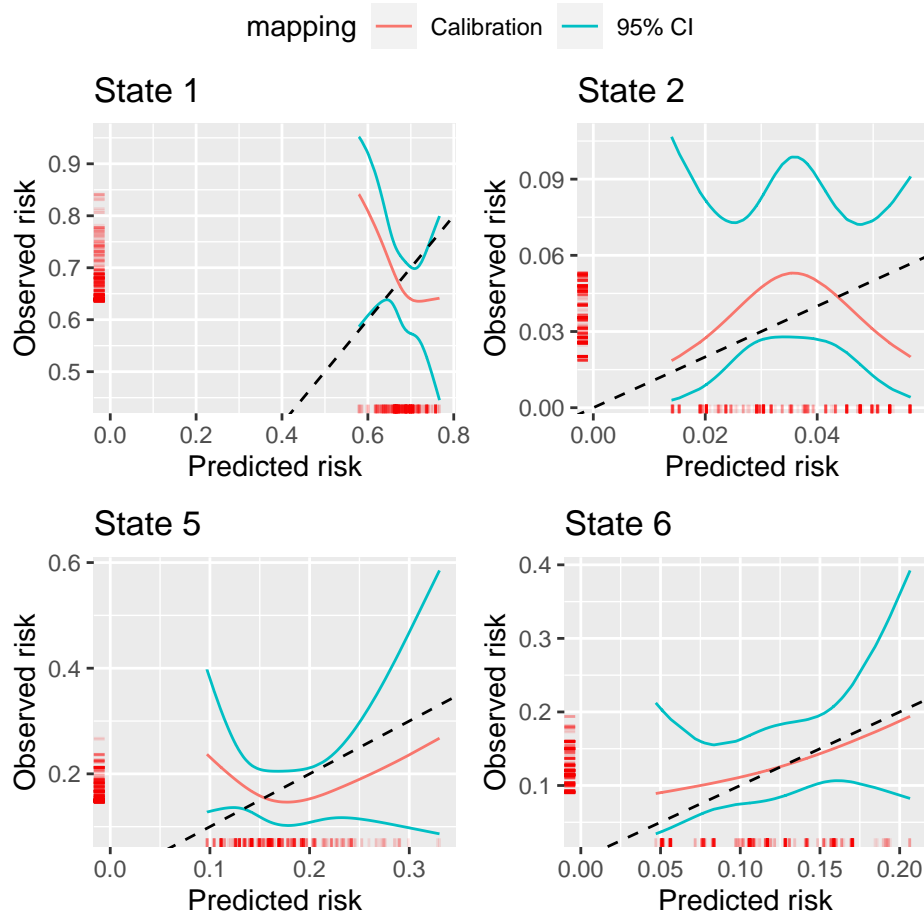


Figure 6: Pseudo-value calibration curves out of state $j = 1$ at time $s = 100$.

```

+           w.covs = c("year", "agecl", "proph", "match"),
+           CI = 95,
+           CI.R.boot = 200)
R> ### Calibration using pseudo-values
R> dat.calib.pv.j3.s100 <-
+   calib_pv(data.mstate = msebmtcal,
+           data.raw = ebmtcal,
+           j=3,
+           s=100,
+           t = t.eval,
+           tp.pred = tp.pred.j3s100,
+           curve.type = "rcs",
+           rcs.nk = 3,
+           group.vars = c("year"),
+           CI = 95,
+           CI.type = "parametric")

```

Again there are only four possible states that an individual may transition into, although this includes states 3 (adverse event) and 4 (recovery + adverse event), instead of 1 (post transplant) and 2 (recovery). This is because once an individual has entered state 3, they cannot move backwards into states 1 or 2. The calibration plots are better than for $j = 1$. For transitions into states $k = 3, 4$ and 6, the calibration curves are monotonically increasing and comparatively close to the line of perfect calibration, although the confidence intervals are still quite large. This is true when calibration is assessed using BLR-IPCW or pseudo-values. Again the calibration of state 5 is very poor. This makes it difficult to base any clinical decisions on the predicted transition probabilities for relapse out of states $j = 1$ or 3 at time $s = 100$, whereas making clinical decisions based on the risk of death ($k = 6$) after survival for 100 days is more viable, as this was well calibrated for both $j = 1$ and $j = 3$. With the exception of the transition probabilities from $j = 1$ into state $k = 3$ made at time $s = 0$, there has been broad agreement between the calibration curves estimated using the BLR-IPCW and pseudo-value approaches. This provides some reassurance about the assessment of calibration, and that the assumptions on which each method is based are satisfied.

4. Discussion

Multistate models are a unique tool for prediction, handling both competing risks and the occurrence of intermediate health states in the same model. Development of multistate models for prediction is becoming more common, yet validation of such models is still very uncommon. A major barrier to implementation of statistical techniques is often the availability of software (Pullenayegum *et al.* 2016). **calibmsm** has been developed to aid in the implementation of techniques to assess the calibration of the transition probabilities from a multistate model. This paper has extended previously proposed methods for assessing the calibration of the transition probabilities out of the initial state (Pate *et al.* 2023), to the transition probabilities out of any state j at any time s . While package development has focused on multistate models, **calibmsm** could be used to assess the calibration of predicted risks from a range of other models, including: any model which utilises information post baseline to update

```
R> plot(dat.calib.blr.j3.s100, combine = TRUE, nrow = 2, ncol = 2)
```

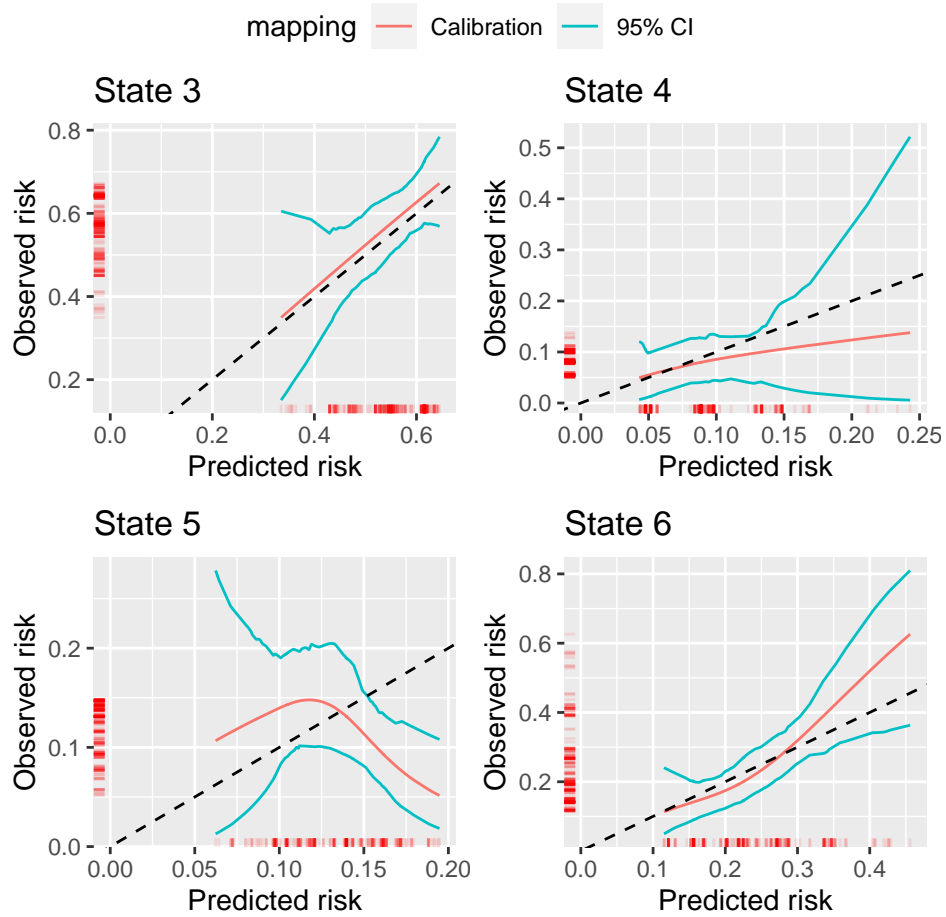


Figure 7: BLR-IPCW calibration curves out of state $j = 3$ at time $s = 100$.

```
R> plot(dat.calib.pv.j3.s100, combine = TRUE, nrow = 2, ncol = 2)
```

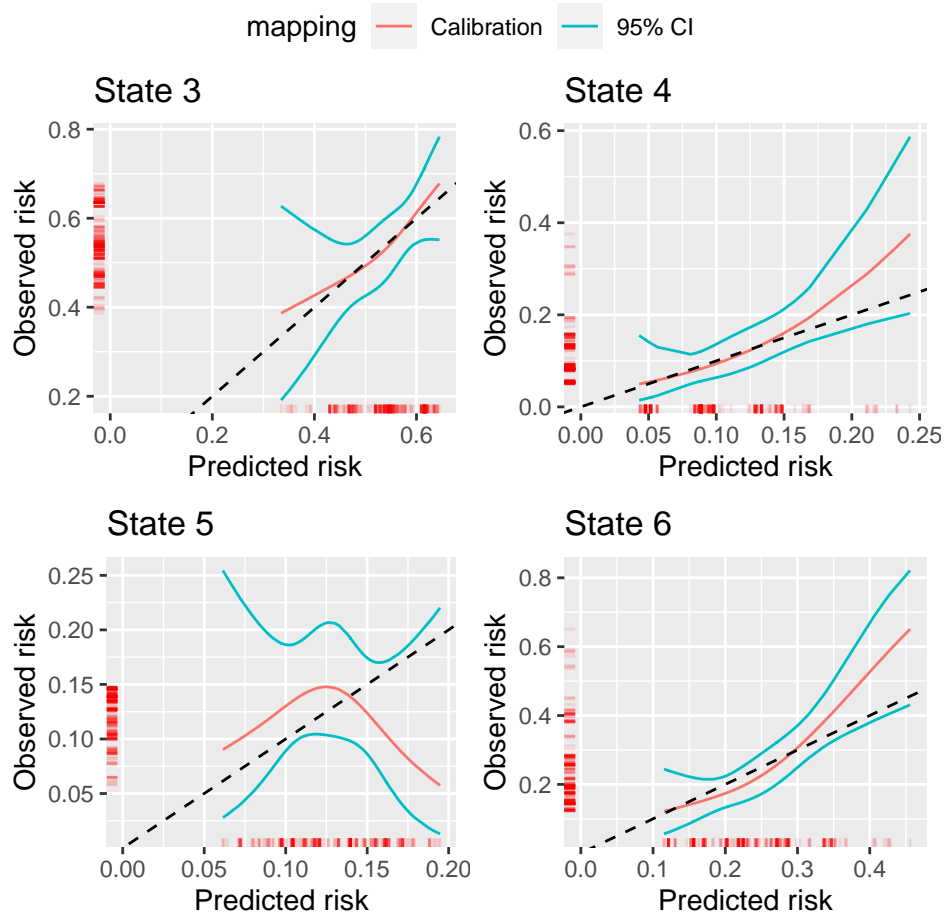


Figure 8: Pseudo-value calibration curves out of state $j = 3$ at time $s = 100$.

predictions (Bull *et al.* 2020), dynamic models (van Houwelingen 2007; ?; Grand *et al.* 2018), competing risks models (Putter *et al.* 2006) and standard single outcome survival models, where predictions can be made at any landmark time.

All three methods (BLR-IPCW, MLR-IPCW and pseudo-value) have been shown to give an unbiased assessment of calibration under non-informative censoring mechanisms, and a predominately unbiased assessment of calibration under strongly informative censoring (Pate *et al.* 2023). This paper found broadly similar evaluation of calibration when using the BLR-IPCW and pseudo-value methods, however there were discrepancies in the evaluation of calibration of the transition probabilities into state $k = 3$. In situations like this, we recommend testing the assumptions of each method as was done in [vignette-Evaluation-of-estimation-of-IPCWs](#). While we concluded that the BLR-IPCW was likely to be biased in this particular example, this is not a general finding. Further research evaluating each methods performance in a wider range of simulation scenarios, and by a different research group (Boulesteix *et al.* 2013), would be highly valuable (Heinze *et al.* 2022).

Given it is possible to use `calibmsm` to validate a standard competing risks model, we carried out a sensitivity analysis to compare the approaches described in this paper with the 'graphical calibration curves' of Austin *et al.* (2022), which already exist for this purpose [vignette-Comparison-with-graphical-calibration-curves-in-competing-risks-setting](#). BLR-IPCW, pseudo-values, and graphical calibration curves (MLR-IPCW excluded for not producing a calibration curve) all resulted in similar calibration curves. This is with the exception of BLR-IPCW for state $k = 3$, which has been previously discussed. The three methods take completely different approaches to assessing the calibration of a competing risks model. Therefore finding agreement between these assessments of calibration can provide reassurance that the calibration plots are correct, and is an exercise that could be repeated in practice. Despite this, the relative performance of each method in a wider range of competing risks scenarios remains unknown. A comparison of these methods in a simulation when the assumptions of each method do and do not hold, and under a range of sample sizes and multistate model structures, would be therefore valuable (Heinze *et al.* 2022).

The BLR-IPCW, MLR-IPCW and pseudo-value approaches have different computational burdens. A calibration curve can be obtained reasonably quickly using the BLR-IPCW or MLR-IPCW approaches, however estimation of confidence intervals for BLR-IPCW using bootstrapping (the recommend method in section 2.6) will result in a high computational time in large validation datasets. On the contrary, obtaining the calibration curve itself using the pseudo-value approach has a high computational burden due to estimation of the pseudo-values. Once these have been calculated, a calibration curve and confidence interval can be estimated quickly using parametric techniques, meaning estimation of the confidence interval adds minimal computational burden. We plan to extend the package to allow users to estimate the pseudo-values for each individual separately before estimating the calibration curve. This will allow the first part of the process to be parallelised and will make estimation of calibration curves using the pseudo-value approach more feasible in large datasets.

Estimation of the weights is clearly of high importance for the BLR-IPCW and MLR-IPCW approaches. If the model to do so is misspecified, this could lead to incorrect evaluation of the calibration. It is possible this is what is causing the difference between the BLR-IPCW and pseudo-value approaches for the calibration of transition probabilities from state $j = 1$ at time $s = 0$ into state $k = 3$, as was explored in [vignette-Evaluation-of-estimation-of-IPCWs](#). This package is focused on creation of calibration curves, but is not a dedicated package

for estimating inverse probability of censoring weights. We encourage users to create a well specified model for the weights. Custom functions for estimating the weights can be specified through the `w.function` argument in both `calib_blr` and `calib_mlr`. Alternatively, weights can be estimated externally and then specified through the `weights` argument. In this latter case, the internal bootstrapping procedure will not work, as the weights need to be re-estimated in each bootstrap dataset. We have provided a more detailed vignette about how to estimate calibration curves and confidence intervals using bootstrapping when defining your own function to estimate the weights ([vignette-BLR-IPCW-manual-bootstrap](#)).

In summary, **calibmsm** provides tools to assess the calibration of the transition probabilities of a multistate model or competing risks model using three approaches (BLR-IPCW, MLR-IPCW and pseudo-values). Further comparison of these approaches in targeted simulations to establish their performance under different censoring mechanisms and assumptions would be valuable.

Computational details

The results in this paper were obtained using R 4.3.1 with the **dplyr** 1.1.2, **tidyr** 1.3.0, **ggplot2** 3.4.2, **ggpubr** 0.6.0, **Hmisc** 5.1.0, **rms** 6.7.0, **VGAM** 1.1.8, **boot** 1.3.28.1, **survival** 3.5.5, **stats** 4.3.1, **magrittr** 2.0.3. R itself and all packages used are available from the Comprehensive R Archive Network (CRAN) at <https://CRAN.R-project.org/>.

Acknowledgments

Thank you to Thomas Yee for helping to debug an issue with implementing vector spline smoothers from the **VGAM** package within **calibmsm**.

References

- Andersen PK, Pohar Perme M (2010). “Pseudo-observations in survival analysis.” *Statistical Methods in Medical Research*, **19**(1), 71–99. ISSN 09622802. doi:10.1177/0962280209105020.
- Andersen PK, Wandall ENS, Pohar Perme M (2022). “Inference for transition probabilities in non-Markov multi-state models.” *Lifetime Data Analysis*, **28**(4), 585–604. ISSN 15729249. doi:10.1007/s10985-022-09560-w. URL <https://doi.org/10.1007/s10985-022-09560-w>.
- Austin PC, Harrell FE, van Klaveren D (2020). “Graphical calibration curves and the integrated calibration index (ICI) for survival models.” *Statistics in Medicine*, **39**(21), 2714–2742. ISSN 10970258. doi:10.1002/sim.8570.
- Austin PC, Putter H, Giardiello D, van Klaveren D (2022). “Graphical calibration curves and the integrated calibration index (ICI) for competing risk models.” *Diagnostic and Prognostic Research*, **6**(1). doi:10.1186/s41512-021-00114-6.

- Austin PC, Steyerberg EW (2014). “Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers.” *Statistics in Medicine*, **33**(3), 517–535. ISSN 02776715. doi:10.1002/sim.5941.
- Boulesteix AL, Lauer S, Eugster MJ (2013). “A Plea for Neutral Comparison Studies in Computational Sciences.” *PLoS ONE*, **8**(4). ISSN 19326203. doi:10.1371/journal.pone.0061562.
- Bull LM, Lunt M, Martin GP, Hyrich K, Sergeant JC (2020). “Harnessing repeated measurements of predictor variables for clinical risk prediction: a review of existing methods.” *Diagnostic and Prognostic Research*, **4**(1). doi:10.1186/s41512-020-00078-z.
- Dafni U (2011). “Landmark analysis at the 25-year landmark point.” *Circulation: Cardiovascular Quality and Outcomes*, **4**(3), 363–371. ISSN 19417713. doi:10.1161/CIRCOUTCOMES.110.957951.
- de Wreede LC, Fiocco M, Putter H (2011). “mstate: An R Package for the Analysis of Competing Risks and Multi-State Models.” *Journal of Statistical Software*, **38**(7). URL <https://cran.r-project.org/package=mstate>.
- EBMT (2023). “Data from the European Society for Blood and Marrow Transplantation.” URL <https://search.r-project.org/CRAN/refmans/mstate/html/EBMT-data.html>.
- Grand MK, de Witte TJM, Putter H (2018). “Dynamic prediction of cumulative incidence functions by direct binomial regression.” *Biometrical Journal*, **60**(4), 737–747. doi:10.1002/bimj.201700194.
- Harrell FE (2015). *Regression Modeling Strategies*. Springer s edition. Springer, Cham.
- Heinze G, Boulesteix AL, Kammer M, Morris TP, White IR (2022). “Phases of methodological research in biostatistics - building the evidence base for new methods.” *Biometrical Journal*, **Early View**. ISSN 15214036. doi:10.1002/bimj.202200222. 2209.13358, URL <http://arxiv.org/abs/2209.13358>.
- Hernan M, Robins J (2020). “12.2 Estimating IP weights via modeling.” In *Causal Inference: What If*, chapter 12.2. Chapman Hall/CRC, Boca Raton.
- Lintu MK, Shreyas KM, Kamath A (2022). “A multi-state model for kidney disease progression.” *Clinical Epidemiology and Global Health*, **13**(December 2021), 100946. ISSN 22133984. doi:10.1016/j.cegh.2021.100946. URL <https://doi.org/10.1016/j.cegh.2021.100946>.
- Masia M, Padilla S, Moreno S, Barber X, Iribarren JA, Romero J, LIST NTFA (2017). “Prediction of long-term outcomes of HIV- infected patients developing non-AIDS events using a multistate approach.” *PLoS ONE*, **112**, 1–16.
- Pate A, Sperrin M, Riley RD, Peek N, Staa TV, Sergeant C, Mamas MA, Lip GYH, Flaherty MO (2023). “Calibration plots for multistate risk predictions models : an overview and simulation comparing novel approaches.” *ArXiv*. doi:10.48550/arXiv.2308.13394. 2308.13394.

- Pullenayegum EM, Platt RW, Barwick M, Feldman BM, Ofringa M, Thabane L (2016). “Knowledge translation in biostatistics: A survey of current practices, preferences, and barriers to the dissemination and uptake of new statistical methods.” *Statistics in Medicine*, **35**(6), 805–818. ISSN 10970258. doi:10.1002/sim.6633.
- Putter H, Fiocco M, Geskus RB (2007). “Tutorial in biostatistics: Competing risks and multi-state models.” *Statistics in medicine*, **26**(11), 2389–2430. doi:https://doi.org/10.1002/sim.2712. URL https://doi.org/10.1002/sim.2712.
- Putter H, Spitoni C (2018). “Non-parametric estimation of transition probabilities in non-Markov multi-state models: The landmark Aalen–Johansen estimator.” *Statistical Methods in Medical Research*, **27**(7), 2081–2092. ISSN 14770334. doi:10.1177/0962280216674497.
- Putter H, Van Hage JD, De Bock GH, Elgalta R, Van De Velde CJ (2006). “Estimation and prediction in a multi-state model for breast cancer.” *Biometrical Journal*, **48**(3), 366–380. ISSN 03233847. doi:10.1002/bimj.200510218.
- R Core Team (2023). “R: A Language and Environment for Statistical Computing.” URL https://www.r-project.org/.
- Riley RD, Snell KI, Ensor J, Burke DL, Harrell FE, Moons KG, Collins GS (2019). “Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes.” *Statistics in Medicine*, **38**(7), 1276–1296. ISSN 10970258. doi:10.1002/sim.7992.
- Sperrin M, Riley RD, Collins GS, Martin GP (2022). “Targeted validation: validating clinical prediction models in their intended population and setting.” *Diagnostic and Prognostic Research*, **6**(1), 4–9. ISSN 2397-7523. doi:10.1186/s41512-022-00136-8. URL https://doi.org/10.1186/s41512-022-00136-8.
- Steyerberg EW, Harrell Jr FE (2016). “Prediction models need appropriate internal, internal-external, and external validation.” *Journal of Clinical Epidemiology*, **69**, 245–247. doi:10.1016/j.jclinepi.2015.04.005.
- Van Calster B, McLernon DJ, Van Smeden M, Wynants L, Steyerberg EW, Bossuyt P, Collins GS, MacAskill P, Moons KG, Vickers AJ (2019). “Calibration: The Achilles heel of predictive analytics.” *BMC Medicine*, **17**(1), 1–7. ISSN 17417015. doi:10.1186/s12916-019-1466-7.
- Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW (2016). “A calibration hierarchy for risk models was defined: From utopia to empirical data.” *Journal of Clinical Epidemiology*, **74**, 167–176. ISSN 18785921. doi:10.1016/j.jclinepi.2015.12.005. URL http://dx.doi.org/10.1016/j.jclinepi.2015.12.005.
- Van Hoorde K, Van Huffel S, Timmerman D, Bourne T, Van Calster B (2015). “A spline-based tool to assess and visualize the calibration of multiclass risk predictions.” *Journal of Biomedical Informatics*, **54**, 283–293. ISSN 15320464. doi:10.1016/j.jbi.2014.12.016. URL http://dx.doi.org/10.1016/j.jbi.2014.12.016.
- Van Hoorde K, Vergouwe Y, Timmerman D, Van Huffel S, Steyerberg W, Van Calster B (2014). “Assessing calibration of multinomial risk prediction models.” *Statistics in Medicine*, **33**(15), 2585–2596. doi:10.1002/sim.6114.

- van Houwelingen HC (2007). “Dynamic Prediction by Landmarking in Event History Analysis.” *Scandinavian Journal of Statistics*, **34**(1), 70–85.
- van Smeden M, Reitsma JB, Riley RD, Collins GS, Moons KG (2021). “Clinical prediction models: diagnosis versus prognosis.” *Journal of Clinical Epidemiology*, **132**, 142–145. ISSN 18785921. doi:10.1016/j.jclinepi.2021.01.009. URL <http://dx.doi.org/10.1016/j.jclinepi.2021.01.009>.
- Wickham H (2016). “ggplot2: Elegant Graphics for Data Analysis.” URL <https://ggplot2.tidyverse.org>.
- Yee TW (2015). *Vector Generalized Linear and Additive Models*. 1 edition. Springer New York, NY. ISBN 978-1-4939-4198-8. doi:10.1007/978-1-4939-2818-7. URL <https://link.springer.com/book/10.1007/978-1-4939-2818-7>.

Affiliation:

Alexander Pate
Division of Imaging, Informatics and Data Science
Faculty of Biology, Medicine and Health
University of Manchester M139PR, UK
E-mail: alexander.pate@manchester.ac.uk