Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *doi: 10.18637/jss.v000.i00*

calibmsm: An R package for calibration plots of the transition probabilities in 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 Centre

Glen P. Martin
University of Manchester

Abstract

Multistate models, which allow users to answer a wide range of clinical questions, are becoming a more commonly used tool 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 in practice. Currently no software exists to aid in assessing the calibration of a multistate model. **calibmsm** has been developed to simplify this process for practicing 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.

This article details the methodology and provides a comprehensive example on how to assess the calibration of a model developed to predict recovery, adverse events, relapse and survival in patients with blood cancer after a transplantation. This is an illustrative example to showcase how to use the package and motivate a discussion around which of the calibration methods is most appropriate.

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 (**?**) and are used widely to inform clinical practice. A multistate model (**?**) may

be used when there are multiple outcomes of interest, or when a single outcome of interest may be reached via intermediates states. 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. Multistate models are increasingly being developed given the many medical scenarios where modelling patient movement between clinical 'states' is of intrest (**????**). All risk prediction models developed for use in clinical practice should be validated in a relevant cohort prior to implementation (**??**). A key part of the validation process is assessment of the calibration of the model (**?**). Ideally calibration curves should be produced which allow evaluation of the calibration over the entire distribution of predicted risk, corresponding to moderate assessment of calibration (**?**).

The R package **mstate** (**?**) 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. **calibmsm** has been developed to enable researchers to estimate calibration curves and scatter plots using three approaches outlined by Pate et al **XXXX REF project 6 (may have to put on Arxiv)**, 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 (**??**), 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 **REF XXXX**.

? used data from the European Society for Blood and Marrow Transplantation (**?**) to showcase how to develop a multistate model for clinical prediction of outcomes after bone morrow transplantation in leukemia patients (Figure **??**). 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 setting for our example and steps for data preparation and are described in section 3. In section 4 we show how to estimate calibration curves and scatter plots using **calibmsm**. Section 5 contains a discussion and summary.

2. Methodology

2.1. Setup

Let $X(t) \in \{1, ..., K\}$ 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 $s < t$. The aim is to estimate observed event probabilities:

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

In the absence of censoring, this can be done using cross sectional calibration techniques in

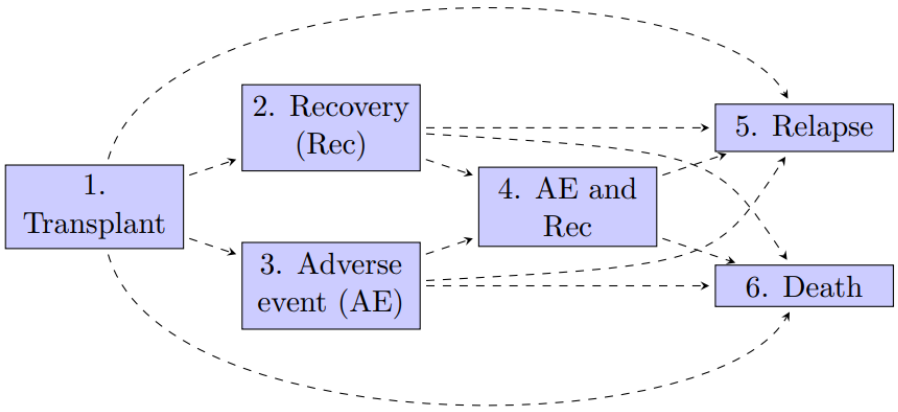

Figure 1: A six-state model for leukemia patients after bone marrow transplantation. Figure taken from (**?**).

a landmark (**??**) cohort of individuals who are in state j at time s (i.e. methods to assess the calibration of models predicting binary or polytomous 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. BLR-IPCW calibration curves

The first approach produces calibration curves using a framework for binary logistic regression models with inverse probability of censoring weights (BLR-IPCW). 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 uncensored at time t, weighted using inverse probability of censoring weights (section **??**). We suggest using a loess smoother:

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

or a logistic regression model with restricted cubic splines:

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

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 $\{\hat{p}_{j,k}(s,t), \hat{o}_{j,k}(s,t)\}$. **This method assumes that in the reweighted population, each outcome $I_k(t)$ is independent from the censoring mechanism**.

2.3. MLR-IPCW calibration 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 an polytomous indicator variable taking values $I_X(t) \in \{1, ..., K\}$ such that $I_X(t) = k$ if an individual is in state k at time t. The nominal recalibration framework of **??** is then applied in the landmark cohort of individuals uncensored at time t, weighted using inverse probability of censoring weights (section **??**). First calculate the log-ratios of the predicted transition probabilities:

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

Then fit the following multinomial logistic regression model:

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

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 (**?**). 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 (**?**) form of calibration than that evaluated by BLR-IPCW, and will also result in observed event probabilities which sum to 1.**This method assumes that in the reweighted population, the outcome $I_X(t)$ is independent from the censoring mechanism**.

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. 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 $t_{abs} < t$. 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`, $\hat{P}[t_{cens} > t | t > s, X(s) = j, \mathbf{Z}]$ is estimated using a cox proportional hazards model where all predictors $\mathbf{Z}$ are assumed to have a linear effect on the log-hazard. However users can also input their own vector of weights. 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 weights themselves using non-linear models of $\mathbf{Z}$ and interaction terms when appropriate.

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 the predictors $\mathbf{Z}$, 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 $\mathbf{Z}$, there is a risk of bias.

2.5. Pseudo-value calibration plots

The third approach produces calibration curves using pseudo-values. Pseudo-values can be used to estimate quantities of interest in censored survival data and multistate survival data. For certain estimators $\hat{\theta}$ (where θ must take the form of an expectation), 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 transition probabilities in a multistate model, $P[X(t) = k | X(s) = j]$, is the Landmark Aalen-Johansen estimator (**?**). Note that this is equivalent to landmarking the cohort and estimating the Aalen-Johansen estimator (**?**), which is what's done in this package. The resulting pseudo-values are vectors with K elements, one for each possible transition. These pseudo-values can replace the outcome $I_k(t)$ in equations (**??**) and (**??**) in order to estimate $o_{j,k}(s,t)$.

However, the pseudo-values are based on the same assumptions as the underlying estimator $\hat{\theta}$. The Landmark Aalen-Johansen estimator requires non-informative 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 of state k. Both options are

implemented in this package. whether the pseudo-values are calculated within subgroups or not, they are used as the outcome in models (**??**) and (**??**) 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 (**??**) 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. 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'. For the pseudo-value approach, 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 be calculated using parametric estimates of the standard error when making predictions of the observed event probabilities (i.e. when making predictions from model (**??**) or (**??**)). To obtain a parametric confidence interval when using the BLR-IPCW approach, a robust sandwich-type estimator should be used to estimate the standard error (**?**). In **calibmsm**, this has been implemented for calibration curves estimated using restricted cubic splines. However, a sandwich estimator was shown to be biased when estimating the standard error of a treatment effect when using inverse probability of treatment weighting in survival analysis (**?**). The performance of robust sandwich-type estimators when using inverse probability of censoring weights in a binary logistic regression is still relatively unknown **I'm struggling to find references on this**. Furthermore, the size of the confidence interval will be underestimated as uncertainty in estimation of the weights is not considered, which is more of an issue at small sample sizes.

Both the parametric and bootstrap methods have been built into **calibmsm**. However for the reasons outlined above, we recommend using parametric confidence intervals for the pseudo-value calibration curves and bootstrapping for the BLR-IPCW calibration curves.

3. Clinical setting and data preperation

We utilise data from the European Society for Blood and Marrow Transplantation (**?**), 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 (**?**).

Four datasets are provided to enable assessment of a multistate model fitted to these data. 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 `msebmtcal`, which is the `ebmt` dataset converted into `msdata` format using the process outlined in **?**. 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("msebmtcal")
R> head(msebmtcal)

  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
```

In the work of **?**, the focus is on predicting transition probabilities made at times $s = 0$ and $s = 100$ 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 **?**, following the theory of **?**. 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 estimator 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. The code for deriving all these datasets is provided in the source code for **calibmsm**.

```
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
```

4. Assessing calibration using calibration curves and scatter plots

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 **??**. The second function, `plot.calib_blr`, `plot.calib_pv` or `plot.calib_mlr`, produces the plots using **ggplot2**. Separating these processes allows users the flexibility of producing their own Figures once the calibration data has been estimated.

The functions for estimating the calibration curves have three arguments where the data provided must be in a specific format. The `data.mstate` argument requires an object of class `msdata`, and is used to implement the landmarking. A dataset of this class must be produced using the package **mstate** (**?**). 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. Both `data.mstate` and `data.raw` should contain a patient ID variable `id`. The `tp.pred` argument must contain a column for each transition k, even if the transition from j to k has zero probability. Each row in `tp.pred` should correspond to the equivalent row in `data.raw`. The datasets described in section **??** meet these criteria.

4.1. Plots 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 first evaluate calibration using the BLR-IPCW approach, implemented through the function `calib_blr`.

The predicted transition probabilities from state $j = 1$ at time $s = 0$ (`tp.pred`) are extracted from the object `tps0`. We choose to estimate the calibration curves using restricted cubic splines, and 3 knots are chosen given the reasonably small size of the dataset. Weights are estimated using the internal estimation procedure and the predictor variables `year`, `agecl`, `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 **??**. The maximum weight (`w.max = 10`) and stabilisation of weights (`stabilised = FALSE`) 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.eval = t.eval,
+                  tp.pred = tps0 |>
+                    dplyr::filter(j == 1) |>
+                    dplyr::select(any_of(paste("pstate", 1:6, sep = ""))),
+                  curve.type = "rcs",
+                  rcs.nk = 3,
+                  w.covs = c("year", "agecl", "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. These data frames have less rows than `ebmtcal` because calibration can only be assessed in individuals uncensored at time `t.eval` **replace t.eval with t**. The second element (named `metadata`) is a metadata argument containing a vector of the possible transitions from state j (all states cannot necessarily be reached from state j), the size of the confidence interval (currently `false`), and the type of calibration curve (`rcs` or `loess`).

```
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] 0.0877 0.0796 0.0836 0.0877 0.0879 ...
  ..$ obs.upper: num [1:1778] 0.128 0.126 0.123 0.16 0.128 ...
 $ 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] 0.12 0.157 0.146 0.144 0.12 ...
  ..$ obs.upper: num [1:1778] 0.232 0.22 0.209 0.211 0.232 ...
 $ 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] 0.1008 0.0911 0.0459 0.0584 0.1008 ...
  ..$ obs.upper: num [1:1778] 0.156 0.149 0.145 0.146 0.156 ...
 $ 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] 0.204 0.195 0.16 0.196 0.204 ...
  ..$ obs.upper: num [1:1778] 0.281 0.257 0.223 0.261 0.281 ...
 $ 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 0.149 0.175 0.172 0.163 ...
  ..$ obs.upper: num [1:1778] 0.215 0.185 0.267 0.25 0.211 ...
 $ 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] 0.183 0.225 0.275 0.252 0.183 ...
  ..$ obs.upper: num [1:1778] 0.229 0.283 0.361 0.31 0.229 ...
```

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

```
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.eval              : num 1826
```

Calibration curves can then be generated using `plot.calib_blr`. This S3 generic was written for objects of class `calib_blr` and produces the calibration plots using **ggplot2**. The calibration curves (Figure **??**) indicate the level of calibration is different for the transition probabilities into each of the different states. The calibration into states 4 and 6 is good, as both contain the line of prefect calibration across the entire range of predicted risk. 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, and large portions of the confidence intervals do not contain the line of perfect calibration. Importantly the calibration of the transition probabilities into state 5 (Relapse), a key clinical outcome in this clinical setting, is extremely poor.

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 **??**). We choose to calculate pseudo-values within individuals with the same year of transplant (`group.vars`), and then split individuals into a further three groups defined by their predicted risk (`n.pctls`). The number of percentiles should be increased in bigger validation datasets. Year of transplant

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

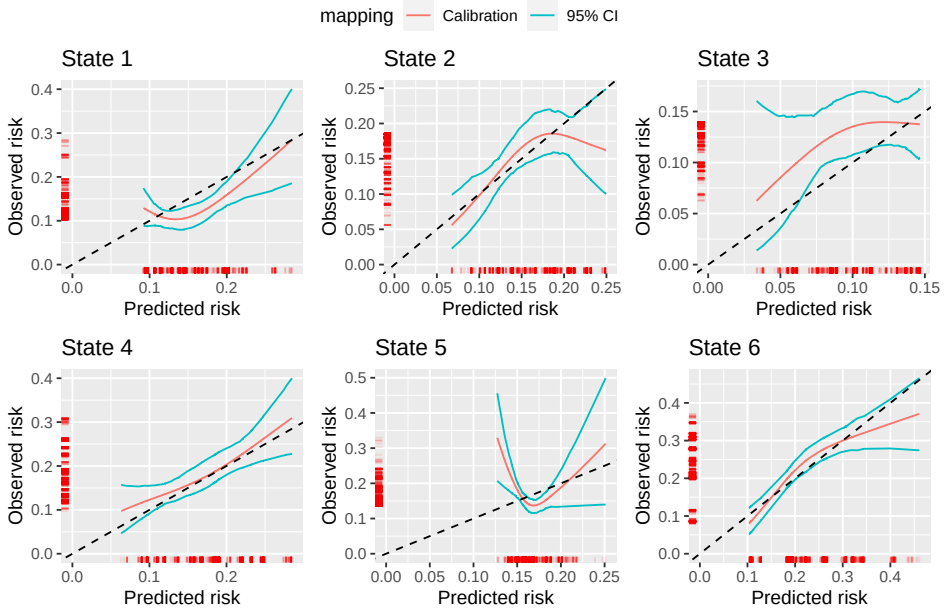

Figure 2: BLR-IPCW calibration curves out of state j = 1 at time s = 0.

was identified as a subgrouping variable through exploration of the dataset (supplementary material XXXX). A later transplant resulted in a shorter possible follow up, an earlier administrative censoring time, and it was therefore highly predictive of the probability of being censored. A parametric confidence interval is estimated as recommended in section **??**.

```
R> dat.calib.pv <-
+    calib_pv(data.mstate = msebmtcal,
+                  data.raw = ebmtcal,
+                  j=1,
+                  s=0,
+                  t.eval = t.eval,
+                  tp.pred = tps0 |>
+                   dplyr::filter(j == 1) |>
+                   dplyr::select(any_of(paste("pstate", 1:6, sep = ""))),
+                  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.calib_pv`. The pseudo-value calibration curves (Figure **??**) are largely similar to the BLR-IPCW calibration curves (Figure **??**). 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

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

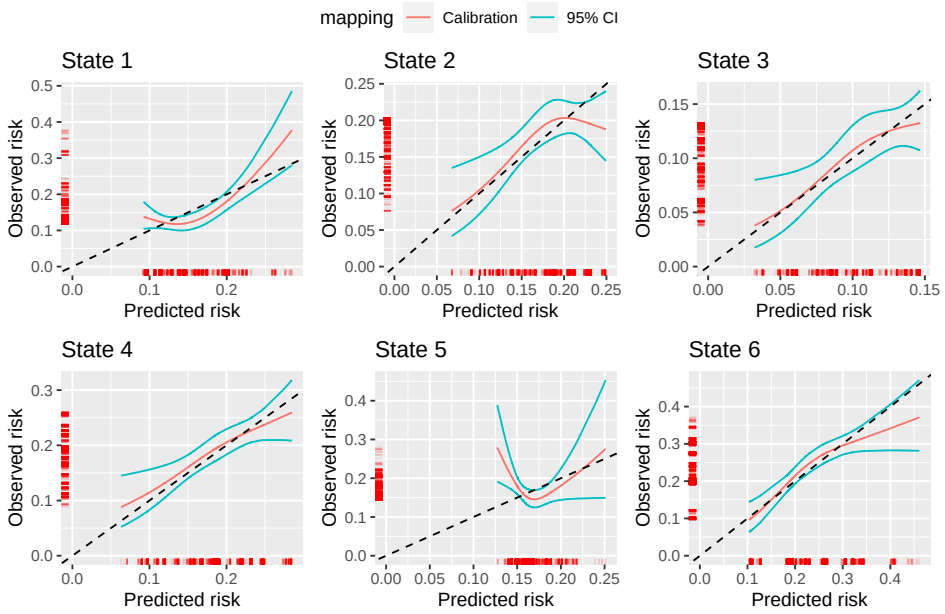

Figure 3: Pseudo-value calibration curves out of state j = 1 at time s = 0.

the pseudo-value calibration plot indicates the transition probabilities are well calibrated, but the BLR-IPCW calibration plot indicates the transition probabilities under predict. If developing this model in a clinical setting, further exploration of this difference would be required. In particular, how the variable year of transplant acts on the censoring mechanism, and for which method the assumptions about independence of the censoring mechanism (section **??**) are most likely to hold.

I could delve into this a bit more deeply. I think the pseudo-value method is probably more appropriate, as I reckon the model for estimating the weights in `calib_blr` is probably misspecified as the proportional hazard assumption unlikely to hold within levels of year of transplant. This then leads to a conversation around why the BLR-IPCW plot and pseudo-value plot are the same for all other states. I think possibly because state 3 has some odd properties, in that patients may not move into state 3 after 100 days. I think this discussion is valuable, as it highlights the need to correct specify weights/review the assumptions of each method, but also wary of over complicating this paper.

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,
+                    s=0,
+                    t.eval = 1826,
```

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

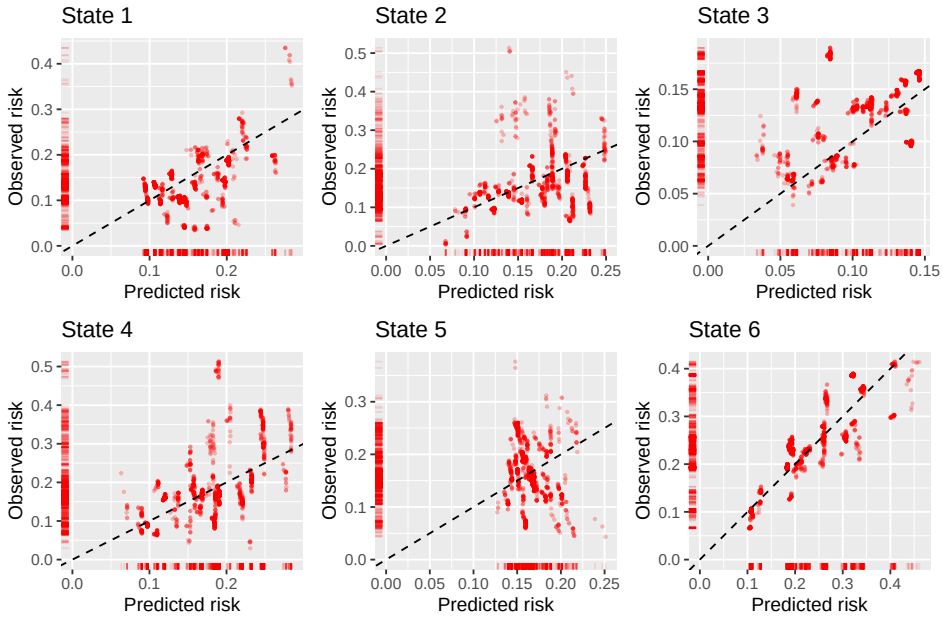

Figure 4: MLR-IPCW calibration curves out of state j = 1 at time s = 0.

```
+                 tp.pred = tps0 |>
+                   dplyr::filter(j == 1) |>
+                   dplyr::select(any_of(paste("pstate", 1:6, sep = ""))),
+                 w.covs = c("year", "agecl", "proph", "match"))
```

The MLR-IPCW calibration plots, produced using `plot.calib_mlr` are contained in Figure **??**. 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 similar calibration according to the BLR-IPCW plots. 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 highly dependent on the predicted transition probabilities of the other states. This insight can be gained because MLR-IPCW is a stronger (**?**) form of calibration assessment than the BLR-IPCW and pseudo-value approaches. As a result, this type of calibration assessment requires a bigger sample size as the confidence intervals around the observed event probabilities will be bigger than for BLR-IPCW. Despite this, it is not clear how to present confidence intervals for all data points simultaneously.

4.2. Plots out of state $j = 1$ and 3 at time $s = 100$

In the work of **?** 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 using the same approaches in combination with landmarking, as described in section **??**. In **calibmsm**, the process remains the same, changing the inputted values `j` and `s`, and providing the appropriate predicted transition probabilities into the argument `tp.pred`. We start by producing the calibration plots for $j = 1$ and $s = 100$ using the BLR-IPCW (Figure **??**) and pseudo-value (Figure **??**) methods. Given the small number of data points in this analysis induced by landmarking, we do not produce calibration scatter plots using MLR-IPCW.

```
R> dat.calib.blr.j1.s100 <-
+    calib_blr(data.mstate = msebmtcal,
+                   data.raw = ebmtcal,
+                   j=1,
+                   s=100,
+                   t.eval = t.eval,
+                   tp.pred = tps100 |>
+                     dplyr::filter(j == 1) |>
+                     dplyr::select(any_of(paste("pstate", 1:6, sep = ""))),
+                   curve.type = "rcs",
+                   rcs.nk = 3,
+                   w.covs = c("year", "agecl", "proph", "match"),
+                   CI = 95,
+                   CI.R.boot = 200)
R> dat.calib.pv.j1.s100 <-
+    calib_pv(data.mstate = msebmtcal,
+                  data.raw = ebmtcal,
+                  j=1,
+                  s=100,
+                  t.eval = t.eval,
+                  tp.pred = tps100 |>
+                   dplyr::filter(j == 1) |>
+                   dplyr::select(any_of(paste("pstate", 1:6, sep = ""))),
+                  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. 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 probability. We follow this up with the pseudo-value calibration plots (Figure **??**) which leads to similar conclusions, as again only stat e$k = 6$ has a monotonically increasing calibration curve. The size of the confidence intervals are much larger than in the calibration plots for state $j = 1$ at time $s = 0$ (Figures **??** and **??**) for both the BLR-IPCW and pseudo-value approaches. For states $k = 2$ and $k = 5$, this could mean the poor calibration is a result of sampling variation as opposed to a misspecified model. A larger validation dataset would be required to get to

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

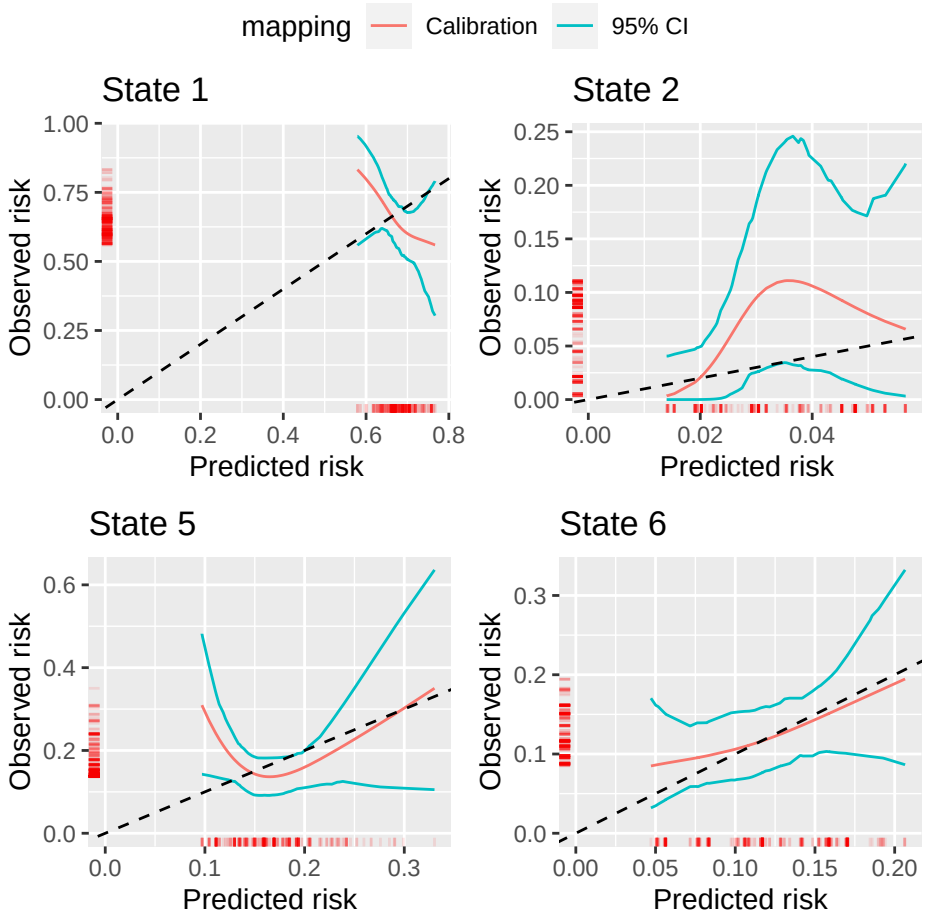

Figure 5: BLR-IPCW calibration curves out of state j = 1 at time s = 100.

the bottom of this. For state $k = 1$, the calibration is extremely poor and is unlikely driven by sampling variation.

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

```
R> dat.calib.blr.j3.s100 <-
+    calib_blr(data.mstate = msebmtcal,
+                  data.raw = ebmtcal,
+                  j=3,
+                  s=100,
+                  t.eval = t.eval,
+                  tp.pred = tps100 |>
+                    dplyr::filter(j == 3) |>
+                    dplyr::select(any_of(paste("pstate", 1:6, sep = ""))),
+                  curve.type = "rcs",
+                  rcs.nk = 3,
```

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

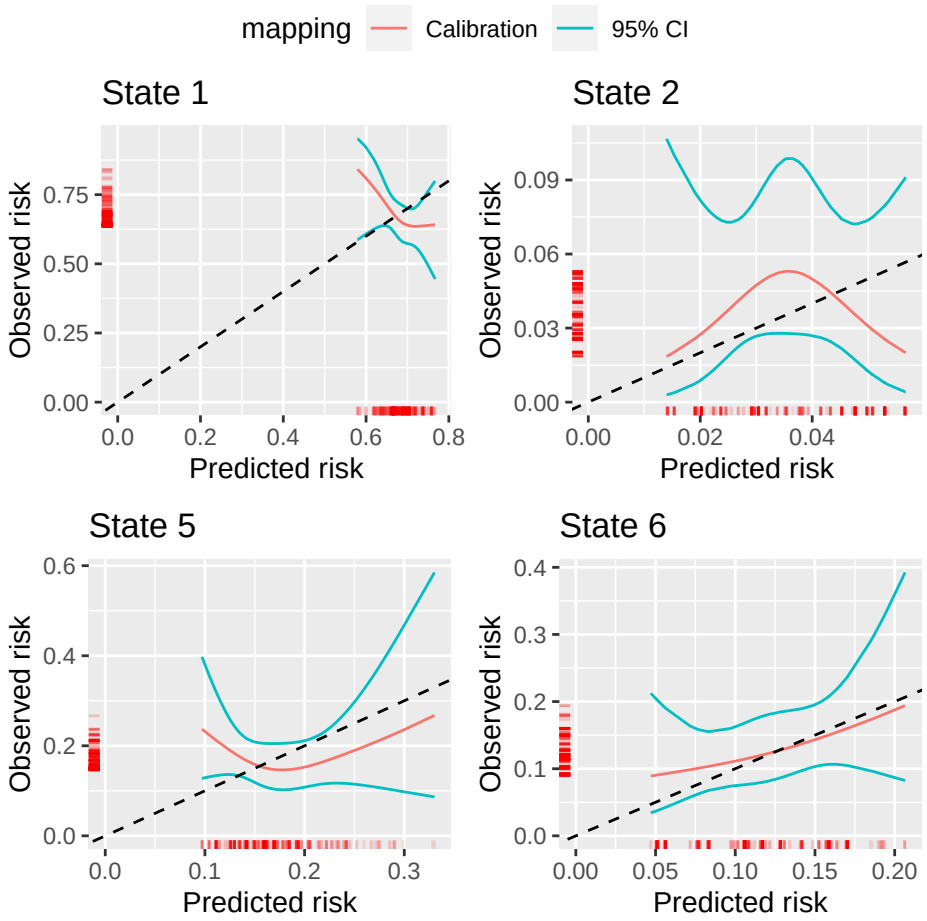

Figure 6: Pseudo-value calibration curves out of state j = 1 at time s = 100.

```
+                    w.covs = c("year", "agecl", "proph", "match"),
+                    CI = 95,
+                    CI.R.boot = 200)
R> dat.calib.pv.j3.s100 <-
+    calib_pv(data.mstate = msebmtcal,
+                   data.raw = ebmtcal,
+                   j=3,
+                   s=100,
+                   t.eval = t.eval,
+                   tp.pred = tps100 |>
+                    dplyr::filter(j == 3) |>
+                    dplyr::select(any_of(paste("pstate", 1:6, sep = ""))),
+                   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). 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 the confidence interval contains the entire line of perfect calibration. 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 evidence about the assessment of calibration, and that the assumptions on which each method is based are satisfied.

References

Aalen OO, Johansen S (1978). "An Empirical Transition Matrix for Non-Homogeneous Markov Chains Based on Censored Observations." *Scandinavian Journal of Statistics*, **5**(3), 141–150.

Austin PC (2016). "Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis." *Statistics in Medicine*, **35**(30), 5642–5655. ISSN 10970258. doi:10.1002/sim.7084.

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.

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

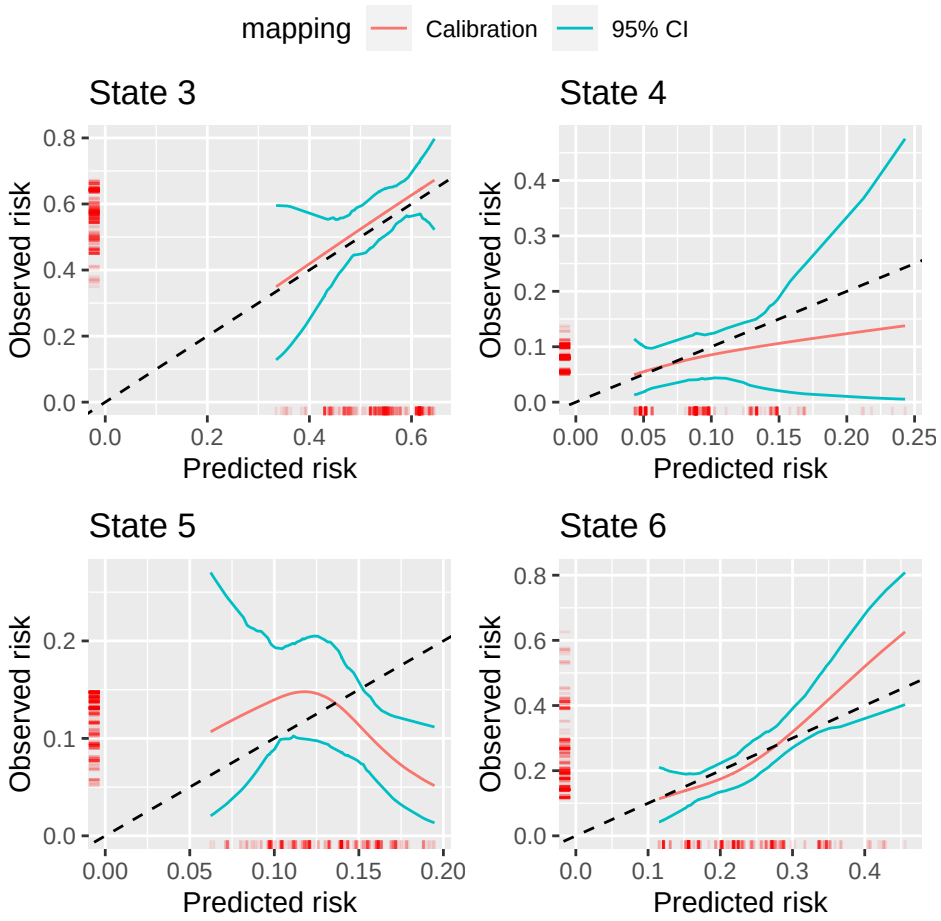

Figure 7: BLR-IPCW calibration curves out of state j = 3 at time s = 100.

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

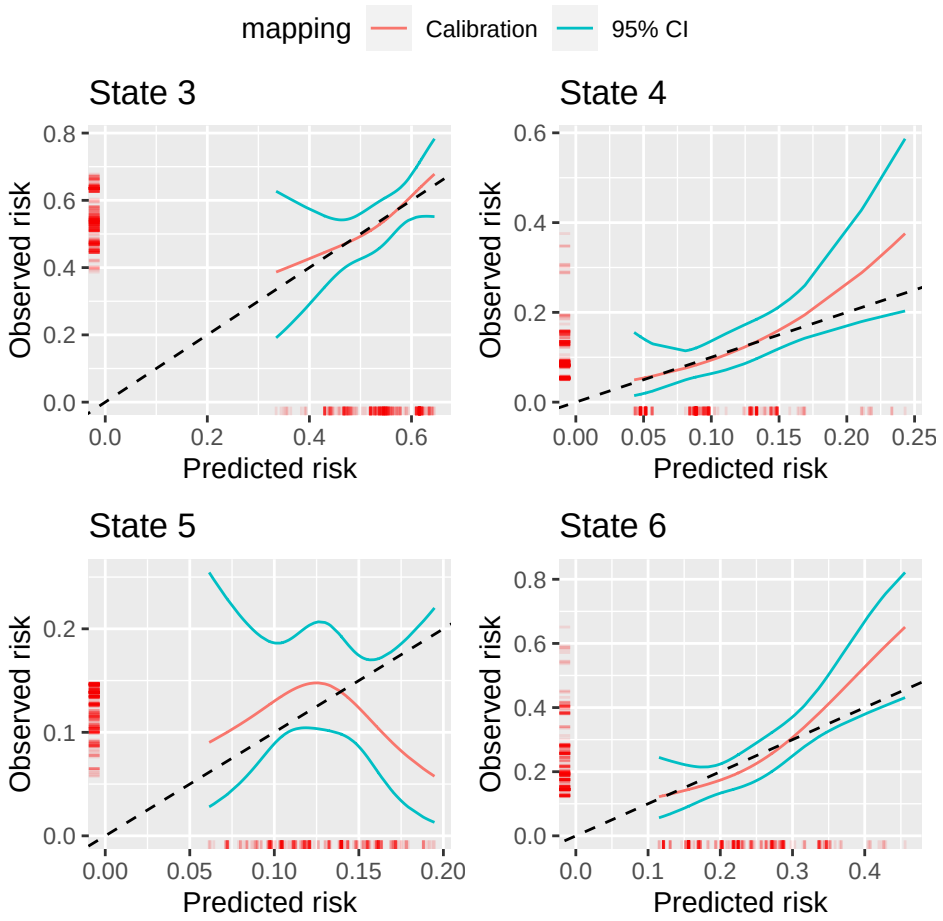

Figure 8: Pseudo-value calibration curves out of state j = 3 at time s = 100.

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).

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.

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.

Le-Rademacher JG, Peterson RA, Therneau TM, Sanford BL, Stone RM, Mandrekar SJ (2018). "Application of multi-state models in cancer clinical trials." *Clinical Trials*, **15**(5), 489–498. ISSN 17407753. doi:10.1177/1740774518789098.

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.

Putter H, Fiocco M, Geskus RB (2007). "Tutorial in biostatistics: Competing risks and multi-state models." *Statistics in medicine*, **26**(11), 2389–2430. doi:10.1002/sim.

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.

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.

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

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
MMMMMM YYYY, Volume VV, Issue II *Submitted:* yyyy-mm-dd
doi:10.18637/jss.v000.i00 *Accepted:* yyyy-mm-dd