






# 1 calibmsm: An R package for calibration plots of the transition 2 probabilities in a multistate model

3 Alexander Pate, Centre for Health Informatics, University of Manchester   
Matthew Sperrin, Centre for Health Informatics, University of Manchester   
Richard D. Riley, Institute of Applied Health Research, University of Birmingham   
Ben Van Calster, Leiden University Medical Center   
Glen P. Martin, Centre for Health Informatics, University of Manchester 

4 June 12, 2024

## Abstract

5 **Background and objective:** Multistate models, which allow the prediction of complex  
6 multistate survival processes such as multimorbidity, or recovery, relapse and death following  
7 treatment for cancer, are being used for clinical prediction. It is paramount to evaluate the  
8 calibration (as well as other metrics) of a risk prediction model before implementation of the  
9 model. While there are a number of software applications available for developing multistate  
10 models, currently no software exists to aid in assessing the calibration of a multistate model,  
11 and as a result evaluation of model performance is uncommon. **calibmsm** has been developed  
12 to fill this gap.

13 **Methods:** Assessing the calibration of predicted transition probabilities between any two  
14 states is made possible through three approaches. The first two utilise calibration techniques  
15 for binary and multinomial logistic regression models in combination with inverse probability  
16 of censoring weights, whereas the third utilises psuedo-values. All methods are implemented in  
17 conjunction with landmarking to allow calibration assessment of predictions made at any time  
18 beyond the start of follow up. This study focuses on calibration curves, but the methodological  
19 framework also allows estimation of calibration slopes and intercepts.

20 **Results:** This article provides a comprehensive example on how to use **calibmsm** to assess the  
21 calibration of a multistate model developed to predict recovery, adverse events, relapse and  
22 survival in patients with blood cancer after a transplantation. The calibration plots indicate  
23 that predictions of relapse made at the time of transplant are poorly calibrated, however  
24 predictions of death are well calibrated. The calibration of all predictions made at 100 days  
25 post transplant appear to be poor, although a larger validation sample is required to make  
26 stronger conclusions.

27 **Conclusions:** **calibmsm** is an R package which allows users to assess the calibration of  
28 predicted transition probabilities from a multistate model. Evaluation of model performance  
29 is a key step in the pathway to model implementation, yet evaluation of the performance of  
30 predictions from multistate models is not common. We hope availability of software will help

31 model developers evaluate the calibration of models being developed.

32 **Keywords:** Clinical prediction models, calibration, model evaluation, multistate, multi-state,  
33 R

## 1. Introduction

34 Risk prediction models enable the prediction of clinical events in either diagnostic or prog-  
35 nostic settings (van Smeden *et al.* 2021) and are used widely to inform clinical practice. A  
36 multistate model (Putter *et al.* 2007) may be used when there are multiple outcomes of in-  
37 terest, or when a single outcome of interest may be reached via intermediate states. For  
38 example, prediction of death after local recurrence or distant metastasis in patients with  
39 breast cancer following surgery (Putter *et al.* 2006); prediction of death following progression  
40 of chronic kidney disease (Lintu *et al.* 2022); prediction of non-AIDS events and death in  
41 individuals living with HIV (Masia *et al.* 2017). Using a multistate model for prediction is  
42 important when the development of an intermediate condition occurring post index date may  
43 have an impact on the risk of future outcomes of interest. Risk prediction models developed  
44 for use in clinical practice should be evaluated in a relevant cohort, or preferably multiple  
45 settings/cohorts, prior to implementation (Steyerberg and Harrell Jr 2016). If the intended  
46 use of this model is known, targeted validation in a specific setting may be preferred (Sperrin  
47 *et al.* 2022). A key part of the validation process is assessment of the calibration of the model  
48 (Van Calster *et al.* 2019). Calibration assesses whether the predicted risks match the observed  
49 event rates in the cohort of interest. Ideally calibration curves should be produced, which  
50 estimate observed event rates as a function the predicted risks over the entire distribution of  
51 predicted risk. This corresponds to a moderate assessment of calibration (Van Calster *et al.*  
52 2016). Methodology on this topic is well developed for binary outcomes (Van Calster *et al.*  
53 2016), survival outcomes (Crowson *et al.* 2016; Austin *et al.* 2020) and survival outcomes in  
54 the presence of competing risks (Gerds *et al.* 2014; Austin *et al.* 2022), however less so for  
55 multistate models, where there is often interest in prediction of more than one outcome state,  
56 and in predictions made at landmark times.

57 The R (R Core Team 2023) package **mstate** (de Wreede *et al.* 2011) provides a compre-  
58 hensive set of tools to develop a multistate model and estimate patient-specific predictions  
59 for a continuously observed multistate survival process. **mstate** focuses on non-parametric  
60 and semi-parametric multistate models where the cause-specific hazards have been fitted us-  
61 ing cox-proportional hazards models. The **flexsurv** package (Jackson 2016) builds on the  
62 functionality of **mstate**, allowing users to fit parametric multistate models (still using the  
63 cause-specific hazards approach), as well as an approach that uses mixture models. Both  
64 **mstate** and **flexsurvreg** allow fitting of clock-forward (Markov) and clock reset (Semi-Markov)  
65 models. The **SemiMarkov** package (Król and Saint-Pierre 2015) contains functions specif-  
66 ically for fitting semi-Markov models. The **msm** package (Jackson 2011) focuses on fit-  
67 ting multistate models to continuous time processes that are observed at arbitrary times  
68 (panel data). The **flexmsm** package provides a general estimation framework for multistate  
69 Markov processes, with flexible specification of the transition intensities. Transition intensi-  
70 ties can be specified through Generalised Additive Models, and allows models with forward  
71 and backward transitions to be fitted. The Lexis functions from the **Epi** package provide  
72 a way to represent and manipulate data from multistate models, and provides an inter-  
73 face to the **mstate**. For a full list of packages available for fitting multistate models, see

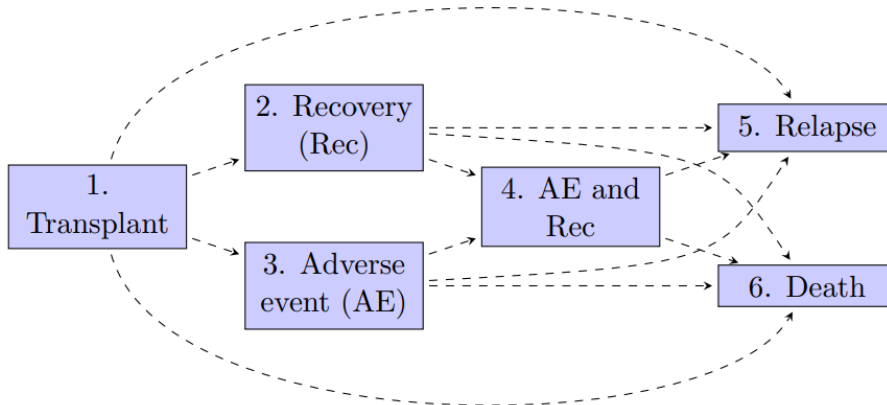


Figure 1: A six-state model for leukemia patients after bone marrow transplantation. Figure taken from (de Wreede *et al.* 2011).

74 <https://cran.r-project.org/web/views/Survival.html>.

75 Despite a wide range of packages for developing multistate models, currently no software  
 76 exists to aid researchers in assessing the calibration of a multistate model that has been  
 77 developed for the purposes of individual risk prediction. The R package **calibmsm** has been  
 78 developed to enable researchers to estimate calibration curves and scatter plots using three  
 79 approaches outlined in Pate *et al.* (2024), which focused on assessing the calibration of the  
 80 transition probabilities out of the starting state. The work in this paper extends the framework  
 81 to assess the calibration of transition probabilities out of any state  $j$  at any time  $s$  using  
 82 landmarking (van Houwelingen 2007; Dafni 2011), provides more details on estimation of the  
 83 inverse-probability of censoring weights (where relevant), and demonstrates the process for  
 84 estimating confidence intervals. **calibmsm** is available from the Comprehensive R Archive  
 85 Network at <https://CRAN.R-project.org/package=calibmsm>.

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

## 2. Methods and Theory

### 96 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:

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

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

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

## 106 2.2. Binary logistic regression with inverse probability of censoring weights 107 (BLR-IPCW) calibration curves

108 The first approach produces calibration curves using a framework for binary logistic regression  
109 models in conjunction with inverse probability of censoring weights to account for informative  
110 censoring. Let  $I_k(t)$  be an indicator for whether an individual is in state  $k$  at time  $t$ .  $I_k(t)$   
111 is then modeled using a flexible approach with  $\hat{p}_{j,k}(s, t)$  as the sole predictor. This model is  
112 fit in the landmark cohort (in state  $j$  at time  $s$ ) of individuals who are also still uncensored  
113 at time  $t$ . This cohort is weighted using inverse probability of censoring weights (see section  
114 2.4). We suggest using a loess smoother (Austin and Steyerberg 2014):

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

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

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

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

## 121 2.3. Multinomial logistic regression with inverse probability of censoring 122 weights (MLR-IPCW) calibration scatter plots

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

128  $t$ , weighted using inverse probability of censoring weights (section 2.4). First calculate the  
 129 log-ratios of the predicted transition probabilities:

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

130 Then fit the following multinomial logistic regression model:

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

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

#### 143 2.4. Estimation of the inverse probability of censoring weights

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

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

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

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

157 In **calibmsm**, unless otherwise specified,  $\hat{P}[t_{cens} > t | t > s, X(s) = j, \mathbf{Z}]$  is estimated using a  
 158 cox proportional hazards model where all predictors  $\mathbf{Z}$  are assumed to have a linear effect on

159 the log-hazard. This is highly restrictive, users can therefore also input their own vector of  
 160 weights, which is strongly recommended. Given the BLR-IPCW and MLR-IPCW approaches  
 161 are both reliant on correct estimation of the weights, we encourage users to take the time to  
 162 carefully estimate the inverse probability of censoring weights using a well specified model.  
 163 The limitations of using the **calibmsm** internal functions for estimating the weights in this  
 164 clinical example (section 3) are discussed in more detail later, and explored in [vignette-](#)  
 165 [Evaluation-of-estimation-of-IPCWs](#).

166 Stabilised weights can be estimated by multiplying by the weights  $w_j(s, t)$  by the mean prob-  
 167 ability of being uncensored:

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

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

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

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

172 This can be estimated using the same approach as for  $w_j(s, t)$ , except there is no requirement  
 173 to be in state  $j$  when landmarking at time  $s$ . If the censoring mechanism is non-informative  
 174 after conditioning on  $\mathbf{Z}$ , then  $w(s, t) = w_j(s, t)$ , and any consistent estimator for  $w(s, t)$  will  
 175 be a consistent estimator of  $w_j(s, t)$ . The advantage is that  $\hat{w}(s, t)$  is calculated by developing  
 176 a model in the cohort of individuals uncensored at time  $s$ , which is a larger cohort than those  
 177 uncensored and in state  $j$  at time  $s$ . Therefore  $\hat{w}(s, t)$  will be a more precise estimator than  
 178  $\hat{w}_j(s, t)$ . On the contrary, if the assumption of non-informative censoring after conditioning on  
 179  $\mathbf{Z}$  does not hold, there is a risk of bias in estimation of the weights. We therefore recommend  
 180 using the estimator  $w_j(s, t)$  unless sample size (number of individuals in state  $j$  at time  $s$ )  
 181 is low, which may be assessed using sample size formula for prediction models with time-to-  
 182 event outcomes (Riley *et al.* 2019). If the sample size is deemed insufficient, one may consider  
 183 using  $w(s, t)$ , but the risk of bias associated with this estimator must be carefully considered.

184 Finally, we state the importance of using inverse probability of censoring weights, even if the  
 185 censoring mechanism is believed to be completely non-informative (i.e. happens at random).  
 186 All multistate models must have an absorbing state, entry into which prevents censoring from  
 187 happening. This induces a dependence between the outcome and the censoring mechanism  
 188 which must be adjusted for using inverse probability of censoring weights. This issue was  
 189 highlighted in the supplementary material of previous work (Pate *et al.* 2024)

## 190 2.5. Pseudo-value calibration plots

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

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

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

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

## 215 2.6. Estimation of confidence intervals

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

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

229 This will produce a number of bootstrapped calibration curves, all plotted over the same  
 230 vectors of predicted transition probabilities. Taking the  $\frac{\alpha}{2}$  and  $(1 - \frac{\alpha}{2})$  percentiles of the  
 231 observed event probabilities for each predicted transition probability gives the required  $1 - \alpha$   
 232 confidence interval around the estimated calibration curve. To estimate confidence intervals

233 for the pseudo-value calibration curves using bootstrapping, the same procedure is applied  
234 except the third step is replaced with 'calculate the pseudo-values within the landmarked  
235 bootstrapped dataset'. This will be highly computationally demanding as the pseudo-values  
236 must be estimated in every bootstrap dataset.

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

249 [Description of package functions and interface

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

258 The validation cohort must be provided to `calib_msm` in two different formats. The `data.raw`  
259 argument requires a `data.frame` (one observation per individual) and is used to fit the calibra-  
260 tion models. For methods BLR-IPCW and MLR-IPCW, `data.raw` should contain variables  
261 `dtcens` (censoring time) and `dtcens.s` (censoring indicator, `dtcens.s = 1` if the individual  
262 is censored at time `dtcens`, `dtcens.s = 0` otherwise), plus any baseline predictors  $\mathbf{Z}$  used  
263 to estimate the weights. For the pseudo-value approach, this dataset should contain any  
264 baseline predictors  $\mathbf{Z}$  which variables will be grouped by before calculating the pseudo-values.  
265 The `data.ms` argument requires a dataset of class `msdata`, which is used to implement the  
266 landmarking and estimate the Aalen-Johansen estimator for the pseudo-value approach. A  
267 dataset of this class can be produced using the package `mstate` ([de Wreede et al. 2011](#)). Both  
268 `data.ms` and `data.raw` should contain corresponding patient ID variables `id`. The predicted  
269 transition probabilities out of state  $j$  at time  $s$  must then be specified through the `tp.pred`  
270 argument, which must contain a column for each transition  $k$ , even if the transition from  $j$   
271 to  $k$  has zero probability. The rows in `tp.pred` must be ordered in the same way as those in  
272 `data.raw`. The datasets described in section 3.1 meet these criteria.

273 The methods in `calibmsm` require continuously observed data, however are agnostic to the  
274 type of multistate model used to estimate the transition probabilities. This includes Markov,  
275 Semi-Markov or non-Markov models, and non-parametric, semi-parametric or parametric  
276 models. A dataset of class `msdata` from `mstate` is required as input, however this is only  
277 required to apply landmarking, and determine the occupied state for each individual at time



278 *t*. The estimated transition probabilities, supplied through `tp.pred` can be estimated using  
 279 any statistical software.

### 3. Clinical example and typical program run

#### 280 3.1. Clinical setting and data preparation

281 We utilise data from the European Society for Blood and Marrow Transplantation ([EBMT](#)  
 282 [2023](#)), containing multistate survival data after a transplant for patients with blood cancer.  
 283 The start of follow up is the day of the transplant and the initial state is alive and in remission.  
 284 There are three intermediate events (2: recovery, 3: adverse event, or 4: recovery + adverse  
 285 event), and two absorbing states (5: relapse and 6: death). This data is available from the  
 286 `mstate` package ([de Wreede et al. 2011](#)). We assume the user of `calibmsm` has experience with  
 287 handling the type of data used to develop a multistate model as outlined by [de Wreede et al.](#)  
 288 ([2011](#)).

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

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

308 The predicted transitions probabilities from each state  $j$  at times  $s = 0$  and  $s = 100$  are  
309 contained in stacked datasets `tps0` and `tps100` respectively. A leave-one-out approach was  
310 used when estimating these transition probabilities. This means each individual was removed  
311 from the development dataset when fitting the multistate model to estimate their transition  
312 probabilities. This approach allows validation to be assessed in the same dataset that the  
313 model was developed with minimal levels of in-sample optimism. Note that for `tps100` the  
314 predicted probabilities for some states  $k$  are equal to 0. This is because no individuals in  
315 state  $j = 1$  at time  $s = 100$  transition into states 3 or 4. This may be due to the definition  
316 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
```

317 **3.2. Calibration plots for the transition probabilities out of**  
318 **state  $j = 1$  at time  $s = 0$**

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

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

325 We first evaluate calibration using the BLR-IPCW approach by specifying `calib.type =`  
 326 `"blr"`. We choose to estimate the calibration curves using restricted cubic splines, although  
 327 the use of loess smoothers would be equally valid. When using restricted cubic splines the  
 328 number of knots must always be specified by the user, and 3 knots are chosen here given the  
 329 reasonably small size of the dataset. Calibration curves could be estimated using  
 330 the internal estimation procedure and the predictor variables `year`, `agec1`, `proph` and `match`.  
 331 The `w.landmark.type` argument assigns whether weights are estimated using all individuals  
 332 uncensored at time  $s$ , or only those uncensored and in state  $j$  at time  $s$ , as discussed in section  
 333 2.4. The maximum weight (`w.max = 10`) and stabilisation of weights (`w.stabilised = TRUE`)  
 334 are left as default. Weights can also be manually specified using the `weights` argument. We  
 335 request 95% confidence intervals for the calibration curves calculated through bootstrapping  
 336 with 200 bootstrap replicates.

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

337 The first element of `dat.calib.blr` (named `plotdata`) contains 6 data frames. One for the  
 338 calibration curves of the transition probabilities into each of the six states,  $k \in \{1, 2, 3, 4, 5, 6\}$ .  
 339 Each data frame contains five columns, `id`: the identifier of each individual; `pred`: the  
 340 predicted transition probabilities; `obs`: the observed event probabilities; `obs.lower` and  
 341 `obs.upper`: the confidence interval for the observed event probabilities. The second ele-  
 342 ment (named `metadata`) is a metadata argument containing information about the data and  
 343 chosen calibration analysis. The plot data and metadata can be viewed using the `print` and  
 344 `metadata` commands respectively. However, it is recommended to get acquainted with the  
 345 underlying object structure, as accessing the plot data will be useful if wanting to customise  
 346 plots or apply bootstrapping manually.

```
R> print(dat.calib.blr)
```

```
$state1
  id      pred      obs obs.lower obs.upper
2  2 0.11401890 0.1095897 0.09090921 0.1287940
4  4 0.13838778 0.1036308 0.08508630 0.1275751
5  5 0.12332255 0.1051035 0.08904819 0.1234944
7  7 0.09737975 0.1236322 0.08986378 0.1606467
```

```
10 10 0.11371889 0.1097779 0.09073130 0.1290061
```

```
$state2
```

	id	pred	obs	obs.lower	obs.upper
2	2	0.2316569	0.1698031	0.1163660	0.2158913
4	4	0.1836189	0.1855591	0.1550775	0.2178034
5	5	0.1609740	0.1759804	0.1446957	0.2095528
7	7	0.2121470	0.1785688	0.1413952	0.2083255
10	10	0.2315632	0.1698443	0.1164773	0.2158726

```
$state3
```

	id	pred	obs	obs.lower	obs.upper
2	2	0.08442692	0.12485834	0.09431447	0.1596038
4	4	0.07579429	0.11666056	0.08980336	0.1517412
5	5	0.05508100	0.09189341	0.05299086	0.1371077
7	7	0.06154308	0.10011560	0.06577209	0.1392661
10	10	0.08440940	0.12484341	0.09431072	0.1595563

```
$state4
```

	id	pred	obs	obs.lower	obs.upper
2	2	0.2328398	0.2427580	0.2011478	0.2843494
4	4	0.2179331	0.2243106	0.1889370	0.2563107
5	5	0.1828176	0.1851051	0.1547167	0.2138236
7	7	0.2206335	0.2275985	0.1906828	0.2592605
10	10	0.2326989	0.2425807	0.2010257	0.2840853

```
$state5
```

	id	pred	obs	obs.lower	obs.upper
2	2	0.1481977	0.1909795	0.1631746	0.2165364
4	4	0.1538475	0.1654523	0.1488839	0.1834957
5	5	0.1425950	0.2215190	0.1760482	0.2650808
7	7	0.1441960	0.2123460	0.1718196	0.2505304
10	10	0.1488068	0.1879278	0.1611251	0.2130000

```
$state6
```

	id	pred	obs	obs.lower	obs.upper
2	2	0.1888598	0.2069354	0.1837972	0.2328139
4	4	0.2304185	0.2542212	0.2274923	0.2820832
5	5	0.3352099	0.3163102	0.2867808	0.3521109
7	7	0.2641006	0.2800368	0.2576373	0.3044745
10	10	0.1888028	0.2068586	0.1837227	0.2327092

```
R> metadata(dat.calib.blr)
```

```
$valid.transitions
```

```
[1] 1 2 3 4 5 6
```

```
$assessed.transitions
```

```
[1] 1 2 3 4 5 6
```

```
$CI
```

```
[1] 95
```

```
$CI.type
```

```
[1] "bootstrap"
```

```
$CI.R.boot
```

```
[1] 200
```

```
$j
```

```
[1] 1
```

```
$s
```

```
[1] 0
```

```
$t
```

```
[1] 1826
```

```
$calib.type
```

```
[1] "blr"
```

```
$curve.type
```

```
[1] "rcs"
```

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

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

359 Next we use the pseudo-value approach to assess calibration by specifying `calib.type =`  
360 `"pv"`. Instead of specifying how the weights are estimated, we now specify variables to define  
361 groups within which pseudo-values will be calculated (see section 2.5). The goal is to induce  
362 uninformative censoring within the chosen subgroups. We chose to calculate pseudo-values in  
363 individuals with the same year of transplant (`pv.group.vars = c("year")`), and then split

```
R> plot.blr
```

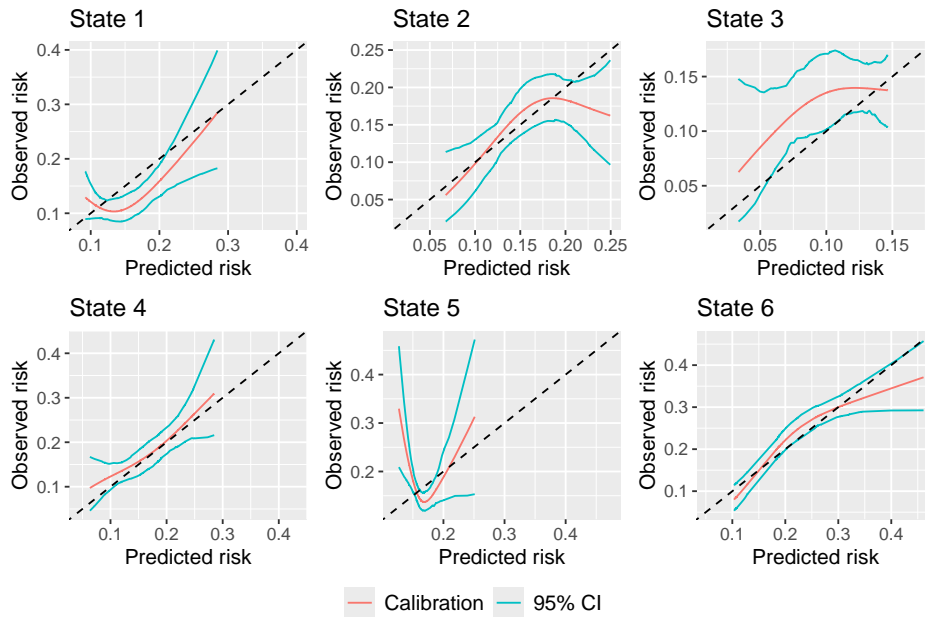


Figure 2: BLR-IPCW calibration curves out of state  $j = 1$  at time  $s = 0$ .

364 individuals into a further three groups defined by their predicted risk (`pv.n.pctls = 3`). The  
 365 number of percentiles should be increased in bigger validation datasets, although guidance  
 366 on specific numbers is currently lacking. Year of transplant was identified as a subgrouping  
 367 variable because a later transplant resulted in a shorter possible follow up, an earlier admin-  
 368 istrative censoring time, and it was therefore highly predictive of being censored. Your data  
 369 should be explored to identify appropriate variables for subgrouping (see [vignette-Evaluation-](#)  
 370 [of-estimation-of-IPCWs](#)). A parametric confidence interval is estimated as recommended in  
 371 section 2.6.

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

372 Calibration curves were then generated using `plot`. The pseudo-value calibration curves

```
R> plot.pv
```

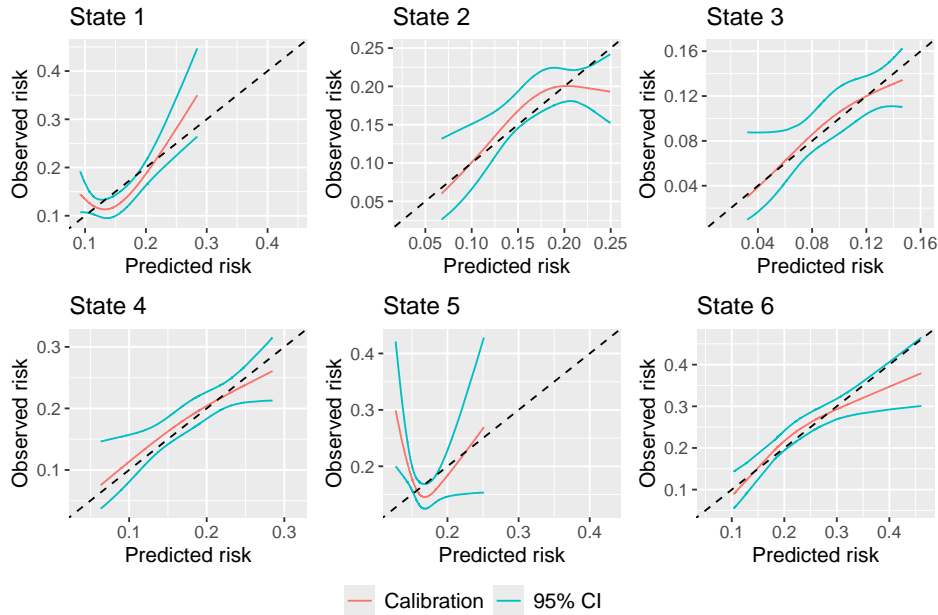


Figure 3: Pseudo-value calibration curves out of state  $j = 1$  at time  $s = 0$ .

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

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

385 Next we use the MLR-IPCW to evaluate calibration which produces a calibration scatter plot  
 386 by specifying `calib.type = "mlr"`. The inputs for calculating the weights are the same as  
 387 for the BLR-IPCW approach, but a confidence interval is no longer requested which is not  
 388 possible for the MLR-IPCW approach.

```
R> dat.calib.mlz <-  

  +   calib_msm(data.ms = msebmtcal,  

  +             data.raw = ebmtcal,  

  +             j=1,
```



```
R> plot.mlr
```

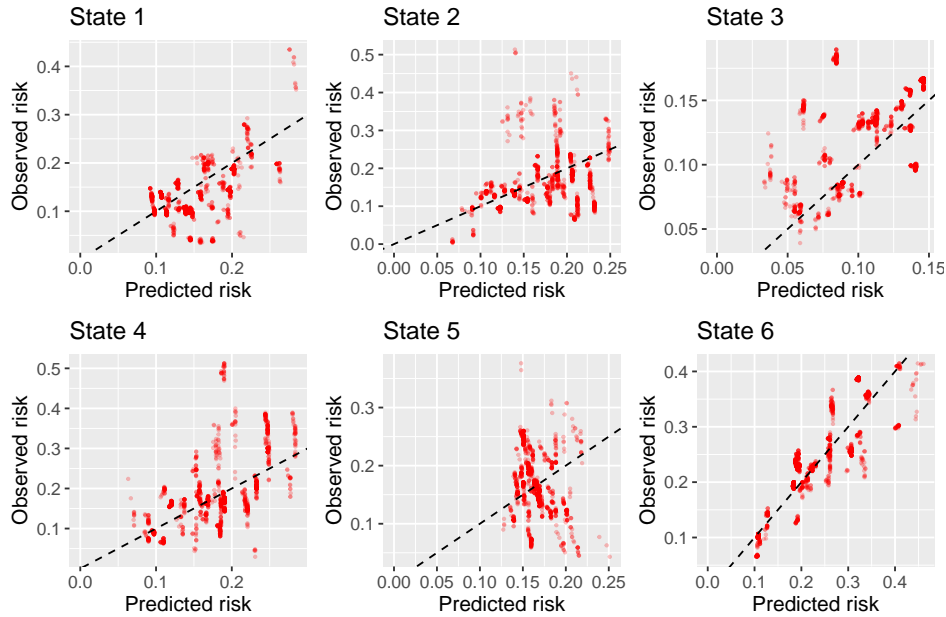


Figure 4: MLR-IPCW calibration scatter plots out of state  $j = 1$  at time  $s = 0$ .

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

389 The MLR-IPCW calibration scatter plots, produced using `plot` are contained in Figure 4.  
 390 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  $\neq k$ .  
 391 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  
 392 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 re-  
 393 mains 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  
 394 the predicted transition probabilities of the other states. This insight can be gained because  
 395 MLR-IPCW is a stronger (Van Calster *et al.* 2016) form of calibration assessment than the  
 396 BLR-IPCW and pseudo-value approaches.  
 397  
 398  
 399  
 400

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

401 **3.3. Calibration plots for the transition probabilities out of**  
 402 **states  $j = 1$  and 3 at time  $s = 100$**

403 In the work of de Wreede *et al.* (2011) focus then shifts to comparing transition probabilities  
 404 when  $s = 100$  depending on whether an individual has had an adverse event (state 3) or  
 405 remains in state 1 (post transplant). Our focus therefore now shifts to assessing the calibration  
 406 of these transition probabilities. This is done through landmarking as described in section 2.  
 407 We start by extracting the predicted transition probabilities from state  $j = 1$  and 3 at time  
 408  $s = 100$  from the object `tps100`. These are the transition probabilities we aim to assess the  
 409 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 = "")))
```

410 The process for estimating the calibration curves remains the same, changing the inputted  
 411 values `j` and `s`, and specifying the appropriate predicted transition probabilities to the ar-  
 412 gument `tp.pred`. We start by producing the calibration plots for  $j = 1$  and  $s = 100$  using  
 413 the BLR-IPCW (Figure 5) and pseudo-value (Figure 6) methods. Given the small number of  
 414 data points in this analysis induced by landmarking, we do not produce calibration scatter  
 415 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_msm(data.ms = msebmtcal,
+             data.raw = ebmtcal,
+             j=1,
+             s=100,
+             t = t.eval,
+             tp.pred = tp.pred.j1s100,
+             calib.type = "blr",
+             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_msm(data.ms = msebmtcal,
+             data.raw = ebmtcal,
+             j=1,
+             s=100,
+             t = t.eval,
+             tp.pred = tp.pred.j1s100,
+             calib.type = "pv",
+             curve.type = "rcs",
+             rcs.nk = 3,
```

```

+           pv.group.vars = c("year"),
+           CI = 95,
+           CI.type = "parametric")

```

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

```

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

```

```

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

```

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

```

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

```

R> plot.blr.j1.s100

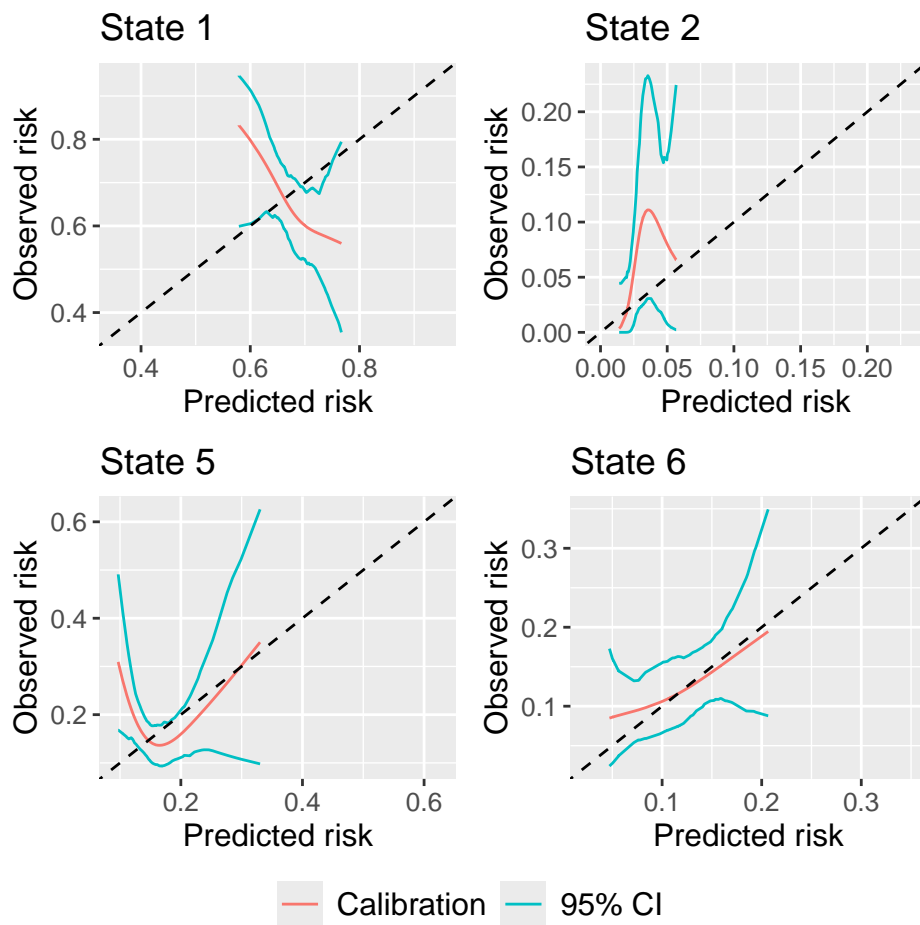


Figure 5: BLR-IPCW calibration curves out of state  $j = 1$  at time  $s = 100$ .

R> plot.pv.j1.s100

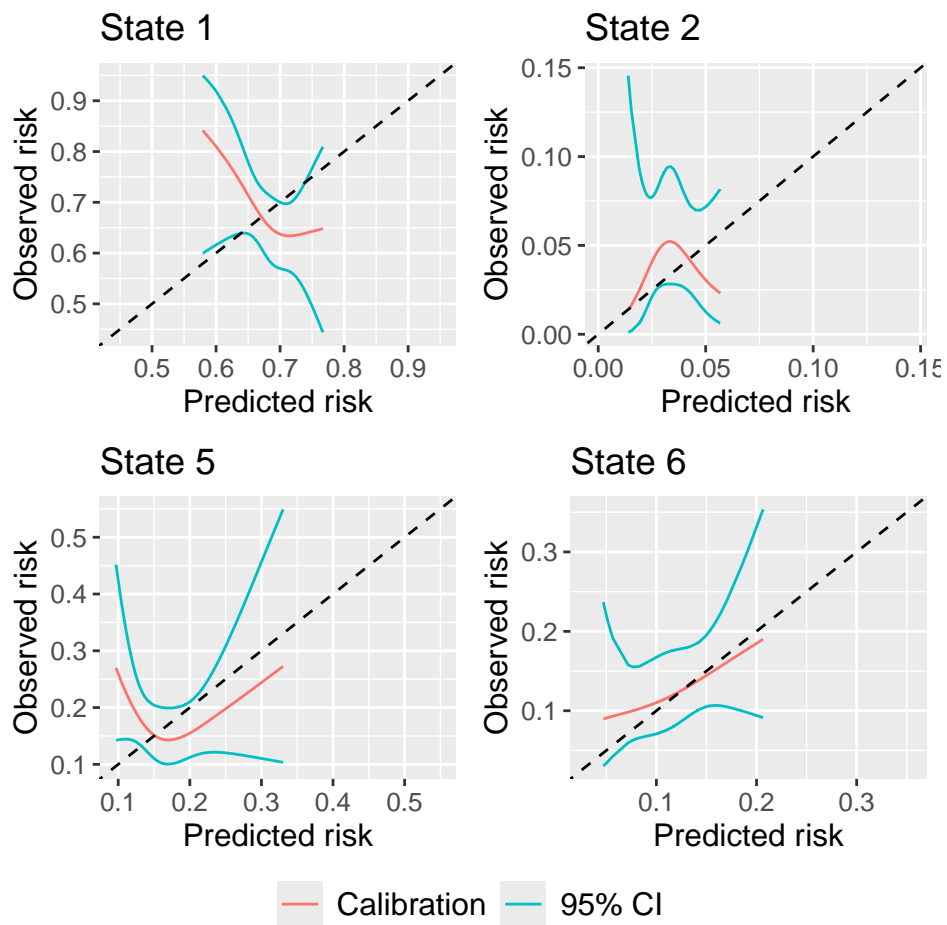


Figure 6: Pseudo-value calibration curves out of state  $j = 1$  at time  $s = 100$ .

```

+           s=100,
+           t = t.eval,
+           tp.pred = tp.pred.j3s100,
+           calib.type = "pv",
+           curve.type = "rcs",
+           rcs.nk = 3,
+           pv.group.vars = c("year"),
+           CI = 95,
+           CI.type = "parametric")

```

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

```

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

```

```

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

```

## 4. Discussion

446 Multistate models are a unique tool for prediction, handling both competing risks and the  
447 occurrence of intermediate health states in the same model. Development of multistate models  
448 for prediction is becoming more common, yet validation of such models is still very uncommon.  
449 A major barrier to implementation of statistical techniques is often the availability of software  
450 (Pullenayegum *et al.* 2016). **calibmsm** has been developed to aid in the implementation of  
451 techniques to assess the calibration of the transition probabilities from a multistate model.  
452 This paper has extended previously proposed methods for assessing the calibration of the  
453 transition probabilities out of the initial state (Pate *et al.* 2024), to the transition probabilities  
454 out of any state  $j$  at any time  $s$ . While package development has focused on multistate models,  
455 **calibmsm** could, in theory, be used to assess the calibration of predicted risks from a range

R> plot.blr.j3.s100

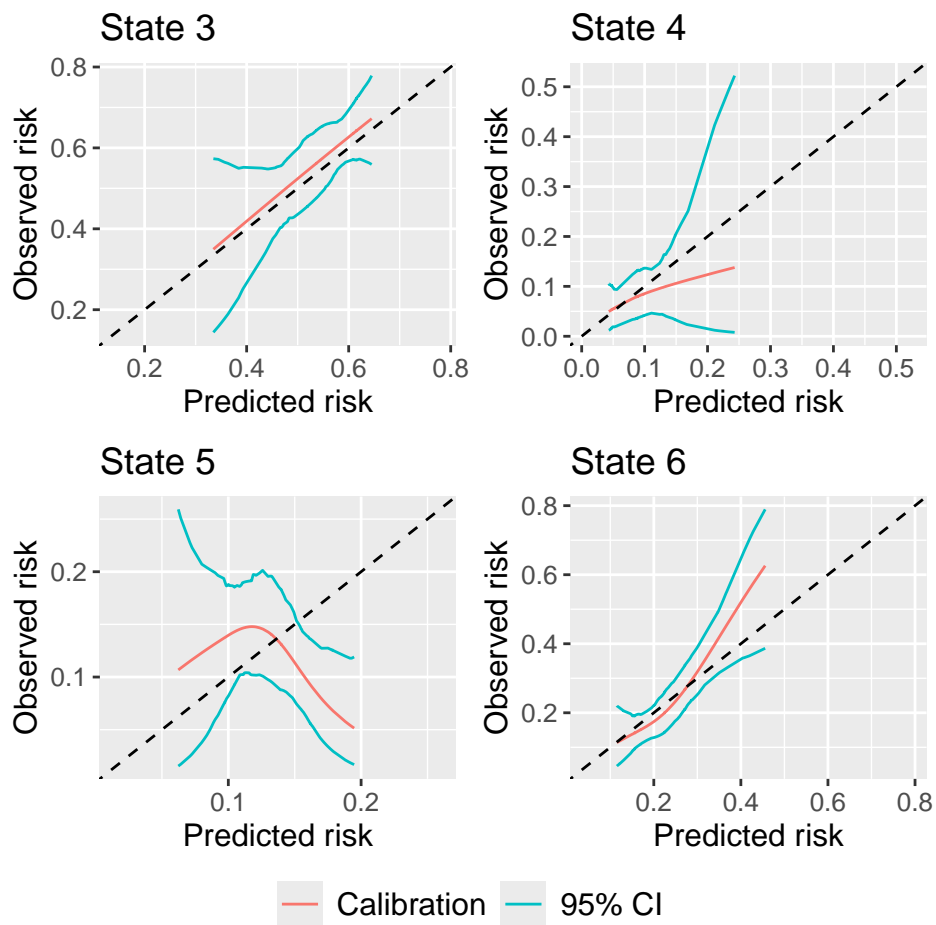


Figure 7: BLR-IPCW calibration curves out of state  $j = 3$  at time  $s = 100$ .

R> plot.pv.j3.s100

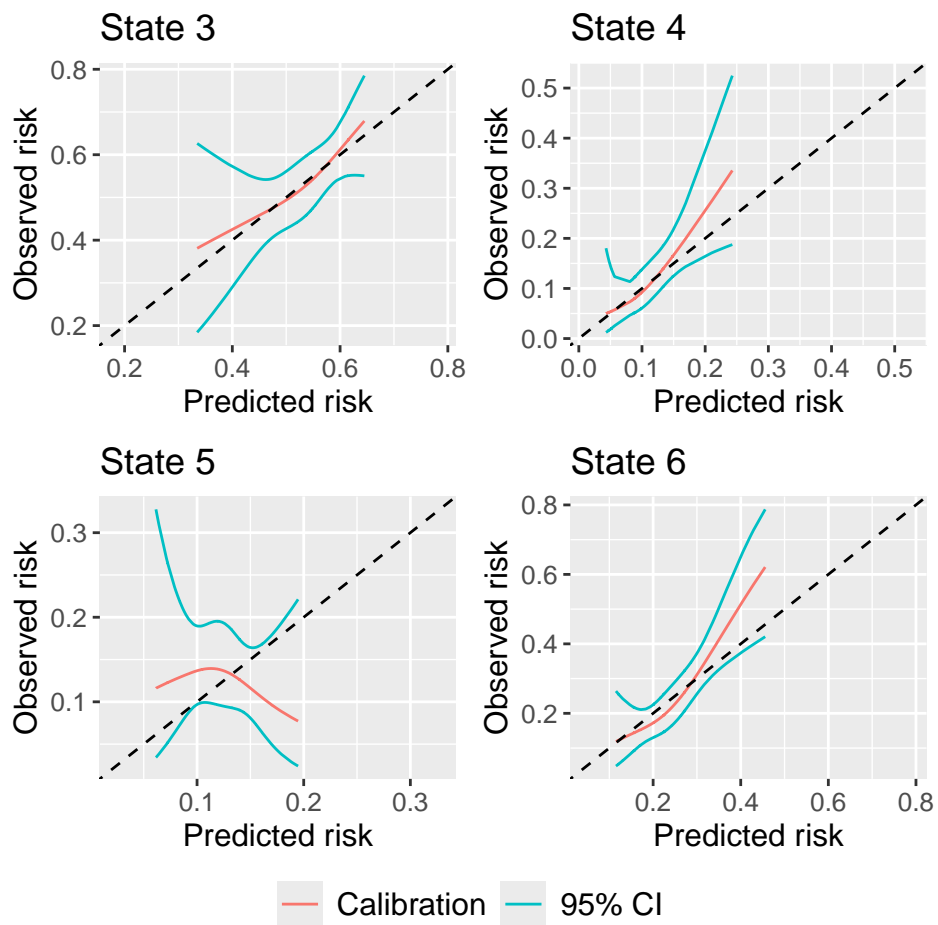


Figure 8: Pseudo-value calibration curves out of state  $j = 3$  at time  $s = 100$ .



456 of other models, including: any model which utilises information post baseline to update  
457 predictions (Bull *et al.* 2020), dynamic models (van Houwelingen 2007; Grand *et al.* 2018),  
458 competing risks models (Putter *et al.* 2006) and standard single outcome survival models,  
459 where predictions can be made at any landmark time.

460 All three methods (BLR-IPCW, MLR-IPCW and pseudo-value) have been shown to give  
461 an unbiased assessment of calibration under non-informative censoring mechanisms, and a  
462 predominately unbiased assessment of calibration under strongly informative censoring (Pate  
463 *et al.* 2024). This paper found broadly similar evaluation of calibration when using the  
464 BLR-IPCW and pseudo-value methods, however there were discrepancies in the evaluation  
465 of calibration of the transition probabilities into state  $k = 3$ . In situations like this, we  
466 recommend testing the assumptions of each method as was done in [vignette-Evaluation-](#)  
467 [of-estimation-of-IPCWs](#). While we concluded that the BLR-IPCW was likely to be biased  
468 in this particular example, this is not a general finding. Further research evaluating each  
469 methods performance in a wider range of simulation scenarios, and by a different research  
470 group (Boulesteix *et al.* 2013), would be highly valuable (Heinze *et al.* 2022).

471 Given it is possible to use **calibmsm** to validate a standard competing risks model (Austin  
472 *et al.* 2022; Gerds *et al.* 2014), we carried out a sensitivity analysis to compare the approaches  
473 described in this paper with the 'graphical calibration curves' of Austin *et al.* (2022), [vignette-](#)  
474 [Comparison-with-graphical-calibration-curves-in-competing-risks-setting](#). BLR-IPCW, pseudo-  
475 values, and graphical calibration curves (MLR-IPCW excluded for not producing a calibration  
476 curve) all resulted in similar calibration curves. This is with the exception of BLR-IPCW  
477 for state  $k = 3$ , which has been previously discussed. The three methods take completely  
478 different approaches to assessing the calibration of a competing risks model. Therefore finding  
479 agreement between these assessments of calibration can provide reassurance that the calibra-  
480 tion plots are correct, and is an exercise that could be repeated in practice. Despite this,  
481 the relative performance of each method in a wider range of competing risks scenarios re-  
482 mains unknown. A comparison of these methods in a simulation when the assumptions of  
483 each method do and do not hold, and under a range of sample sizes and multistate model  
484 structures, would be therefore valuable (Heinze *et al.* 2022).

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

497 Estimation of the weights is clearly of high importance for the BLR-IPCW and MLR-IPCW  
498 approaches. If the model to do so is misspecified, this could lead to incorrect evaluation of  
499 the calibration. It is possible this is what is causing the difference between the BLR-IPCW  
500 and pseudo-value approaches for the calibration of transition probabilities from state  $j = 1$  at  
501 time  $s = 0$  into state  $k = 3$ , as was explored in [vignette-Evaluation-of-estimation-of-IPCWs](#).

502 This package is focused on creation of calibration curves, but is not a dedicated package  
503 for estimating inverse probability of censoring weights. We encourage users to create a well  
504 specified model for the weights (see [Hernan and Robins \(2020\)](#)) if using the BLR-IPCW  
505 or MLR-IPCW approaches. Custom functions for estimating the weights can be specified  
506 through the `w.function` in `calib_msm`. Alternatively, weights can be estimated externally and  
507 then specified through the `weights` argument. In this latter case, the internal bootstrapping  
508 procedure will not work, as the weights need to be re-estimated in each bootstrap dataset.  
509 We have provided a more detailed vignette about how to estimate calibration curves and  
510 confidence intervals using bootstrapping when defining your own function to estimate the  
511 weights ([vignette-BLR-IPCW-manual-bootstrap](#)).

512 In summary, `calibmsm` provides tools to assess the calibration of the transition probabilities  
513 of a multistate model or competing risks model using three approaches (BLR-IPCW, MLR-  
514 IPCW and pseudo-values). Further comparison of these approaches in targeted simulations  
515 to establish their performance under different censoring mechanisms and assumptions would  
516 be valuable. Future work will aim to develop methodology for other model evaluation metrics  
517 and incorporate these into `calibmsm`.

## Computational details

518 The results in this paper were obtained using R 4.4.0 with the `dplyr` 1.1.4, `tidyr` 1.3.1, `gg-`  
519 `plot2` 3.5.1, `ggpubr` 0.6.0, `Hmisc` 5.1.3, `rms` 6.8.1, `VGAM` 1.1.11, `boot` 1.3.30, `survival` 3.5.8,  
520 `stats` 4.4.0, `magrittr` 2.0.3. R itself and all packages used are available from the Comprehensive  
521 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

- 522 Andersen PK, Pohar Perme M (2010). “Pseudo-observations in survival analysis.” *Sta-*  
523 *tistical Methods in Medical Research*, **19**(1), 71–99. ISSN 09622802. [doi:10.1177/](https://doi.org/10.1177/0962280209105020)  
524 [0962280209105020](https://doi.org/10.1177/0962280209105020).
- 525 Andersen PK, Wandall ENS, Pohar Perme M (2022). “Inference for transition proba-
- 526 bilities in non-Markov multi-state models.” *Lifetime Data Analysis*, **28**(4), 585–604.  
527 ISSN 15729249. [doi:10.1007/s10985-022-09560-w](https://doi.org/10.1007/s10985-022-09560-w). URL [https://doi.org/10.1007/](https://doi.org/10.1007/s10985-022-09560-w)  
528 [s10985-022-09560-w](https://doi.org/10.1007/s10985-022-09560-w).
- 529 Austin PC, Harrell FE, van Klaveren D (2020). “Graphical calibration curves and the inte-
- 530 grated calibration index (ICI) for survival models.” *Statistics in Medicine*, **39**(21), 2714–  
531 2742. ISSN 10970258. [doi:10.1002/sim.8570](https://doi.org/10.1002/sim.8570).

- 532 Austin PC, Putter H, Giardiello D, van Klaveren D (2022). “Graphical calibration curves  
533 and the integrated calibration index (ICI) for competing risk models.” *Diagnostic and*  
534 *Prognostic Research*, **6**(1). doi:10.1186/s41512-021-00114-6.
- 535 Austin PC, Steyerberg EW (2014). “Graphical assessment of internal and external calibration  
536 of logistic regression models by using loess smoothers.” *Statistics in Medicine*, **33**(3), 517–  
537 535. ISSN 02776715. doi:10.1002/sim.5941.
- 538 Boulesteix AL, Lauer S, Eugster MJ (2013). “A Plea for Neutral Comparison Studies in  
539 Computational Sciences.” *PLoS ONE*, **8**(4). ISSN 19326203. doi:10.1371/journal.  
540 pone.0061562.
- 541 Bull LM, Lunt M, Martin GP, Hyrich K, Sergeant JC (2020). “Harnessing repeated mea-  
542 surements of predictor variables for clinical risk prediction: a review of existing methods.”  
543 *Diagnostic and Prognostic Research*, **4**(1). doi:10.1186/s41512-020-00078-z.
- 544 Crowson CS, Atkinson EJ, Therneau TM, Lawson AB, Lee D, MacNab Y (2016). “Assessing  
545 calibration of prognostic risk scores.” *Statistical Methods in Medical Research*, **25**(4), 1692–  
546 1706. ISSN 14770334. doi:10.1177/0962280213497434.
- 547 Dafni U (2011). “Landmark analysis at the 25-year landmark point.” *Circulation: Cardiovas-*  
548 *cular Quality and Outcomes*, **4**(3), 363–371. ISSN 19417713. doi:10.1161/CIRCOUTCOMES.  
549 110.957951.
- 550 de Wreede LC, Fiocco M, Putter H (2011). “mstate: An R Package for the Analysis of  
551 Competing Risks and Multi-State Models.” *Journal of Statistical Software*, **38**(7). URL  
552 <https://cran.r-project.org/package=mstate>.
- 553 EBMT (2023). “Data from the European Society for Blood and Marrow Transplantation.”  
554 URL <https://search.r-project.org/CRAN/refmans/mstate/html/EBMT-data.html>.
- 555 Gerds TA, Andersen PK, Kattan MW (2014). “Calibration plots for risk prediction models in  
556 the presence of competing risks.” *Statistics in Medicine*, **33**(18), 3191–3203. ISSN 10970258.  
557 doi:10.1002/sim.6152.
- 558 Grand MK, de Witte TJM, Putter H (2018). “Dynamic prediction of cumulative incidence  
559 functions by direct binomial regression.” *Biometrical Journal*, **60**(4), 737–747. doi:10.  
560 1002/bimj.201700194.
- 561 Harrell FE (2015). *Regression Modeling Strategies*. Springer s edition. Springer, Cham.
- 562 Heinze G, Boulesteix AL, Kammer M, Morris TP, White IR (2022). “Phases of methodological  
563 research in biostatistics - building the evidence base for new methods.” *Biometrical Journal*,  
564 **Early View**. ISSN 15214036. doi:10.1002/bimj.202200222. 2209.13358, URL <http://arxiv.org/abs/2209.13358>.  
565
- 566 Hernan M, Robins J (2020). “12.2 Estimating IP weights via modeling.” In *Causal Inference:*  
567 *What If*, chapter 12.2. Chapman Hall/CRC, Boca Raton.
- 568 Jackson CH (2011). “Multi-State Models for Panel Data: The msm Package for R.” *Journal*  
569 *of Statistical Software*, **38**(8), 128–129. ISSN 19395108. doi:10.1002/wics.10. URL  
570 <https://cran.r-project.org/package=msm>.

- 571 Jackson CH (2016). “Flexsurv: A platform for parametric survival modeling in R.” *Journal*  
572 *of Statistical Software*, **70**(8). ISSN 15487660. doi:10.18637/jss.v070.i08.
- 573 Król A, Saint-Pierre P (2015). “Semimarkov: An R package for parametric estimation  
574 in multi-state semi-markov models.” *Journal of Statistical Software*, **66**(6), 1–16. ISSN  
575 15487660. doi:10.18637/jss.v066.i06.
- 576 Lintu MK, Shreyas KM, Kamath A (2022). “A multi-state model for kidney disease  
577 progression.” *Clinical Epidemiology and Global Health*, **13**(December 2021), 100946.  
578 ISSN 22133984. doi:10.1016/j.cegh.2021.100946. URL [https://doi.org/10.1016/  
579 j.cegh.2021.100946](https://doi.org/10.1016/j.cegh.2021.100946).
- 580 Masia M, Padilla S, Moreno S, Barber X, Iribarren JA, Romero J, LIST NTFA (2017).  
581 “Prediction of long-term outcomes of HIV- infected patients developing non-AIDS events  
582 using a multistate approach.” *PLoS ONE*, **112**, 1–16.
- 583 Pate A, Sperrin M, Riley RD, Peek N, Van Staa T, Sergeant JC, Mamas MA, Lip GYH,  
584 Flaherty MO, Barrowman M, Buchan I, Martin GP (2024). “Calibration plots for multistate  
585 risk predictions models.” *Statistics in Medicine*, (April), 1–23. doi:10.1002/sim.10094.  
586 **2308.13394**, URL <https://pubmed.ncbi.nlm.nih.gov/38720592/>.
- 587 Pullenayegum EM, Platt RW, Barwick M, Feldman BM, Offringa M, Thabane L (2016).  
588 “Knowledge translation in biostatistics: A survey of current practices, preferences, and  
589 barriers to the dissemination and uptake of new statistical methods.” *Statistics in Medicine*,  
590 **35**(6), 805–818. ISSN 10970258. doi:10.1002/sim.6633.
- 591 Putter H, Fiocco M, Geskus RB (2007). “Tutorial in biostatistics: Competing risks and  
592 multi-state models.” *Statistics in medicine*, **26**(11), 2389–2430. doi:[https://doi.org/  
593 10.1002/sim.2712](https://doi.org/10.1002/sim.2712). URL <https://doi.org/10.1002/sim.2712>.
- 594 Putter H, Spitoni C (2018). “Non-parametric estimation of transition probabilities in non-  
595 Markov multi-state models: The landmark Aalen–Johansen estimator.” *Statistical Methods*  
596 *in Medical Research*, **27**(7), 2081–2092. ISSN 14770334. doi:10.1177/0962280216674497.
- 597 Putter H, Van Hage JD, De Bock GH, Elgalta R, Van De Velde CJ (2006). “Estimation and  
598 prediction in a multi-state model for breast cancer.” *Biometrical Journal*, **48**(3), 366–380.  
599 ISSN 03233847. doi:10.1002/bimj.200510218.
- 600 R Core Team (2023). “R: A Language and Environment for Statistical Computing.” URL  
601 <https://www.r-project.org/>.
- 602 Riley RD, Snell KI, Ensor J, Burke DL, Harrell FE, Moons KG, Collins GS (2019). “Minimum  
603 sample size for developing a multivariable prediction model: PART II - binary and time-to-  
604 event outcomes.” *Statistics in Medicine*, **38**(7), 1276–1296. ISSN 10970258. doi:10.1002/  
605 [sim.7992](https://doi.org/10.1002/sim.7992).
- 606 Sperrin M, Riley RD, Collins GS, Martin GP (2022). “Targeted validation: validating clinical  
607 prediction models in their intended population and setting.” *Diagnostic and Prognostic*  
608 *Research*, **6**(1), 4–9. ISSN 2397-7523. doi:10.1186/s41512-022-00136-8. URL [https://doi.org/  
609 //doi.org/10.1186/s41512-022-00136-8](https://doi.org/10.1186/s41512-022-00136-8).

- 610 Steyerberg EW, Harrell Jr FE (2016). “Prediction models need appropriate internal, internal-  
611 external, and external validation.” *Journal of Clinical Epidemiology*, **69**, 245–247. doi:  
612 [10.1016/j.jclinepi.2015.04.005](https://doi.org/10.1016/j.jclinepi.2015.04.005).
- 613 Van Calster B, McLernon DJ, Van Smeden M, Wynants L, Steyerberg EW, Bossuyt P,  
614 Collins GS, MacAskill P, Moons KG, Vickers AJ (2019). “Calibration: The Achilles  
615 heel of predictive analytics.” *BMC Medicine*, **17**(1), 1–7. ISSN 17417015. doi:  
616 [10.1186/s12916-019-1466-7](https://doi.org/10.1186/s12916-019-1466-7).
- 617 Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW (2016). “A  
618 calibration hierarchy for risk models was defined: From utopia to empirical data.” *Journal  
619 of Clinical Epidemiology*, **74**, 167–176. ISSN 18785921. doi:[10.1016/j.jclinepi.2015.](https://doi.org/10.1016/j.jclinepi.2015.12.005)  
620 [12.005](https://doi.org/10.1016/j.jclinepi.2015.12.005). URL <http://dx.doi.org/10.1016/j.jclinepi.2015.12.005>.
- 621 Van Hoorde K, Van Huffel S, Timmerman D, Bourne T, Van Calster B (2015). “A spline-  
622 based tool to assess and visualize the calibration of multiclass risk predictions.” *Journal of  
623 Biomedical Informatics*, **54**, 283–293. ISSN 15320464. doi:[10.1016/j.jbi.2014.12.016](https://doi.org/10.1016/j.jbi.2014.12.016).  
624 URL <http://dx.doi.org/10.1016/j.jbi.2014.12.016>.
- 625 Van Hoorde K, Vergouwe Y, Timmerman D, Van Huffel S, Steyerberg W, Van Calster B  
626 (2014). “Assessing calibration of multinomial risk prediction models.” *Statistics in Medicine*,  
627 **33**(15), 2585–2596. doi:[10.1002/sim.6114](https://doi.org/10.1002/sim.6114).
- 628 van Houwelingen HC (2007). “Dynamic Prediction by Landmarking in Event History Anal-  
629 ysis.” *Scandinavian Journal of Statistics*, **34**(1), 70–85.
- 630 van Smeden M, Reitsma JB, Riley RD, Collins GS, Moons KG (2021). “Clinical prediction  
631 models: diagnosis versus prognosis.” *Journal of Clinical Epidemiology*, **132**, 142–145. ISSN  
632 18785921. doi:[10.1016/j.jclinepi.2021.01.009](https://doi.org/10.1016/j.jclinepi.2021.01.009). URL [http://dx.doi.org/10.1016/  
633 j.jclinepi.2021.01.009](http://dx.doi.org/10.1016/j.jclinepi.2021.01.009).
- 634 Wickham H (2016). “ggplot2: Elegant Graphics for Data Analysis.” URL [https://ggplot2.](https://ggplot2.tidyverse.org)  
635 [tidyverse.org](https://ggplot2.tidyverse.org).
- 636 Yee TW (2015). *Vector Generalized Linear and Additive Models*. 1 edition. Springer New  
637 York, NY. ISBN 978-1-4939-4198-8. doi:[10.1007/978-1-4939-2818-7](https://doi.org/10.1007/978-1-4939-2818-7). URL [https:  
638 //link.springer.com/book/10.1007/978-1-4939-2818-7](https://link.springer.com/book/10.1007/978-1-4939-2818-7).

639 **Affiliation:**

640 Alexander Pate  
641 Division of Imaging, Informatics and Data Science  
642 Faculty of Biology, Medicine and Health  
643 University of Manchester M139PR, UK  
644 E-mail: [alexander.pate@manchester.ac.uk](mailto:alexander.pate@manchester.ac.uk)

645