Inverse Probability Weighting Adjustment of the Logistic Regression Calibration-in-the-Large

MA Barrowman A Pate GP Martin CJM Sammut-Powell M Sperrin

Contents

Last updated: 23 May

Abstract

Introduction

Methods

Results

Discussion

1 Introduction

Clinical prediction models (CPMs) are statistical models/algorithms that aim to predict the presence (diagnostic) or furture occurence (prognostic) of an event of interest, conditional on a set of predictor variables. Before they be implemented in practice, CPMs must be robustly validated. They need to be validated before they are used and a fundamental test of their validity is calibration: the agreement between observed and predicted outcomes. This requires that among individuals with $p\%$ risk of an event, $p\%$ of those have the event across the full risk range (Steyerberg 2008). The simplest assessment of calibration is the calibration-in-the-large, which tests for agreement in mean calibration (the weakest form of calibration) (Calster et al. 2016). With continuous or binary outcomes, such a test is straight-forward: it can be translated to a test for a zero intercept in a regression model with an appropriately transformed linear predictor as an offset, and no other predictors. More complicated measurements of calibration can also be assessed to descibe how calibration changes across the risk range, such as calibration slope (see Appendix **??**). Calibration alone is not enough to fully assess a model's performance however and so we also need measures of discrimination (how well models discern between different patients), e.g the c-statistic and overall accuracy, e.g. the Brier Score.

In the case of time to event models, however, estimation of calibration is complicated in three ways. First, calibration can be computed at multiple time-points and one must decide which time-points to evaluate, and how to integrate over these time-points. The choice and combination of time-points determines what we mean by calibration; this is problem-specific and not the focus of this paper. Calibration can also be integrated over time using the martingale residuals (Crowson, Atkinson, and Therneau 2016); however we focus on the case where calibration at a specific time point is of interest - e.g. as is common in clinical decision support. Second, there exists no explicit intercept in the model because of the non-parametric baseline hazard function (Royston and Altman 2013). The lack of intercept can be overcome provided sufficient information concerning the baseline survival curve is available (although this is rarely the case as seen in QRISK (**???**), ASCVD (Goff et al. 2014) and ASSIGN (de la Iglesia et al. 2011). Once this is established, estimated survival probabilities are available.

Third, censoring needs to be handled in an appropriate way. This is commonly overcome by using Kaplan-Meier estimates (Royston and Altman 2013; Hippisley-Cox et al. 2007), but the censoring assumptions required for the Kaplan-Meier estimate are stronger than those required for the Cox model: the former requiring unconditional independence (random censoring), the latter requiring independence conditional on covariates only. This is a problem because when miscalibration is found using this approach, it is not clear whether this is genuine miscalibration or a consequence of the different censoring assumptions. Royston (Royston 2014, 2015) has proposed the comparison of KM curves within risk groups, which alleviates the strength of the independence assumption required for the censoring handling to be comparable between the Cox model and the KM curves (since the KM curves now only assume independent censoring within risk group). In these papers a fractional polynomial approach to estimating the baseline survival function (and thus being able to share it efficiently) is also provided. However, this does not allow calculations of the overall calibration of the model, which is of primary interest here.

QRISK used the overall KM approach in the 2007 paper (Hippisley-Cox et al. 2007) with good results (6.34% predicted vs 6.25% observed in women and 8.86% predicted vs 8.88% observed in men), but worse results in

the QRISK3 update (Hippisley-Cox, Coupland, and Brindle 2017) (4.7% predicted v 5.8% observed in women and 6.4% predicted vs 7.5% observed in men). This may be because, as follow-up extends, the dependence of censoring on the covariates increases (QRISK had 12 years follow-up, QRISK3 had 18) and an important change between the update was the lower age limit moved from 35 to 25, as well as the implementation of QRISK in clinical practice [**I remember discussing this with Alex & Matt a while ago as to whether the use of QRISK had a feedback loop when updated after it's own implementation. Did this go any further?**].

Royston (Royston 2014) also presented an alternative approach for calibration at external validation. He uses the approach of pseudo-observations, as described by Perme and Anderson (Perme and Andersen 2008) to overcome the censoring issue and produce observed probabilities at individual level; however, this assumes that censoring is independent of covariates.

A solution to this problem is to apply a weighting to uncensored patients based on their probability of being censored according to a model that accounts for covariates. The Inverse Probability of Censoring Weighting (IPCW) relaxes the assumption that patients who were censored are identical to those that remain at risk and replaces it with the assumption that they are exchangeable conditional on the measured covariates. The weighting inflates the patients who were similar to the censored population to account for those patients who are no longer available at a given time.

Gerds & Schumacher (Gerds and Schumacher 2006) have thoroughly investigated the requirements and advantages of applying an IPCW to a performance measure for modelling using the Brier score as an example and demonstrating the efficacy of its use, which was augmented by Spitoni et al (Spitoni, Lammens, and Putter 2018) who demonstrated that any proper scoring rule can be improved by the use of the IPCW. This work has been extended by Han et al (Han, Zhang, and Shao 2017) and Liu et al (Liu, Jin, and Graziano 2016) who demonstrated one can also apply IPCW to the c-statistic (a measure of discrimination).

In this paper we present an approach to assessing the calibration intercept (calibration-in-the-large) and calibration slope in time-to-event models based on estimating the censoring distribution, and reweighting observations by the inverse of the censoring probability. We first show, theoretically, how this method can be used and evidence that the metrics for calibration are amenable to its use. We then compare simulation results from using this weighted estimate to an unweighted estimate within various commonly used methods of calibration assessment.

2 Methods

2.1 Theory

[**Lots of Theory work on the probabilities. May need to drop this if we're unable to do it between us.**]

2.2 Aims

The aim of this simulation study is to investigate the bias induced by applying different methods of assessing model calibration to data that is susceptible to censoring and to compare it to the bias when this data has been adjusted by the Inverse Probability of Censoring Weighting (IPCW).

2.3 Data Generating Method

We simulated populations of patients with survival and censoring times, and took the observed event time as the minimum of these two values along with an event indicator of whether this was the survival or censoring time (Burton et al. 2006). Each population was simulated with three parameters: β, γ and η, which defined

the proportional hazards coefficients for the survival and censoring distributions and the baseline hazard function, respectively.

Patients were generated with a single covariate $Z \sim N(0,1)$ from which, we then generated a survival time, T and a censoring time, C. Survival times were simulated with a baseline hazard $\lambda_0(t) = t^\eta$ (i.e. Weibull), and a proportional hazard of $e^{\beta Z}$. This allows the simulation of a constant baseline hazard ($\eta = 0$) as well as an increasing ($\eta = {}^1/_2$) and decreasing ($\eta = -{}^1/_2$) hazard function Censoring times were simulated with a constant baseline hazard, $\lambda_{C,0}(t) = 1$ and a proportional hazard of $e^{\gamma Z}$. This combines to give a simulated survival function, S as

$$S(t|Z = z) = \exp\left(-\frac{e^{\beta Z} t^{\eta+1}}{\eta+1}\right)$$

and a simulated censoring function, S_c as

$$S_c(t|Z = z) = \exp\left(-e^{\gamma Z} t\right)$$

Once the survival and censoring times were generated, the event time, $X = \min(T, C)$, and the event indicator, $\delta = I(T = X)$, were generated. In practice, only Z, X and δ would be observed.

During each simulation, we varied the parameters to take all the values,$\gamma = \{-2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2\}$, $\beta = \{-2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2\}$ and $\eta = \{-{}^1/_2, 0, {}^1/_2\}$. For each combination of parameters, we generated $N = 100$ populations of $n = 10,000$ patients (a high number of patients was chosen to improve precision of our estimates)

2.4 Prediction Models

[**New section, taken from previous snippets, highlighting/strikethroughs will show the new changes**]

For each population, we used three distinct prediction models for survival. F_P was chosen to exactly model the Data Generating Mechanism (DGM) to emulate a perfectly specified model:

$$F_P(t|Z = z) = 1 - \exp\left(-\frac{e^{\beta Z} t^{\eta+1}}{\eta+1}\right)$$

From this, we also derived a prediction model that would systematically over-estimate the prediction model, F_O, and one which would systematically under-estimate the prediction, F_U. These are defined as:

$$\begin{aligned} F_U(t|Z = z) = &\quad \text{logit}^{-1}\left(\text{logit}\left(F_P(t|z) - 0.2\right)\right) \\ F_O(t|Z = z) = &\quad \text{logit}^{-1}\left(\text{logit}\left(F_P(t|z) + 0.2\right)\right) \end{aligned}$$

These prediction models were used to generate an estimate of the Expected probability that a given patient, with covariate z, will have an event at the given time.

2.5 The IPCW

In order to apply the IPCW, we need to calculate a censoring prediction model. For our purposes, we will again use a perfectly specified censoring distribution, G, to be derived directly from the DGM:

$$G(t|Z = z) = 1 - \exp\left(-e^{\gamma Z} t\right)$$

This is used to calculate an IPCW for all non-censored patients at the last time they were observed (t for patients who have not had an event, and X_i for patients who have had the event), This is defined as:

$$\omega(t|z) = \frac{1}{1 - G(\min(t, X_i)|z)}$$

2.6 Calibration Measurements

The prediction models were assessed at 100 time points, evenly distributed between the 25th and 75th percentile of observed event times, X. At each of these time points, we compare Observed outcomes (O) with the Expected outcomes (E) of the prediction models based on four choices of methodology (Royston 2014, 2015; Riley et al. 2019; Andersen and Pohar Perme 2010) to produce measures for the calibration-in-the-large

- Kaplan-Meier (KM) - A Kaplan-Meier estimate of survival is estimated from the data and the value of the KM curve at the current time is taken to be the average Observed number of events within the population and this is compared with the average Expected value.
- Logistic Unweighted (LU) - Logistic regression is performed on the non-censored population to predict the binary Observed value using the logit(Expected) value as an offset and the Intercept of the regression is the estimate of calibration-in-the-large.
- Logistic Weighted (LW) - As above, but the logistic regression is performed using the IPCW as a weighting for each non-censored patient.
- Pseudo-Observations (PO) - The contribution of each patient (including censored patients) to the overall Observed value is calculated by removing them from the population and aggregating the difference. Regression is performed with the complimentary log-log function as a link function and the log cumulative hazard as an offset with the Intercept representing the estimate of calibration-in-the-large.

Some of these methods produce unusual results for the regressions. Firstly, the weights within the LW method cause the "number of events" being processed (i.e the sum of the weighted events) to be non-integer. This is a minor issue and can be dealt with by most software packages (Wildscop 2013). Secondly, the PO method produces outcomes that are outside of the (0,1) range (Perme and Andersen 2008) required for the complimentary log-log function. To combat this, we re-scale the values produced to be with this range and perform the regression as normal.

2.7 Estimands

For each set of parameters and methodology, our estimand at time, t, measured in simulation $i = 1, ..., N$ is $\theta_i(t)$, the set of estimates of the calibration-in-the-large for the F_P, F_U and F_O models in order. Therefore our underlying truth for all time points is

$$\theta = (0, 0.2, -0.2)$$

From this, we can also define our upper and lower bound for a 95% confidence interval as the vectors $\theta_{i,L}(t)$ and $\theta_{i,U}(t)$.

2.8 Performance Measures

The measures we will take as performance measures as the Bias, the Empirical Standard Error and the Coverage at time, t, along with relevant standard errors and confidence intervals as per current recommendations (Morris, White, and Crowther 2019). These measures can be seen in table 1. For these estimates at each time point, Method and Model, the top and bottom 5% of all simulation estimates will be omitted, leaving $N = 90$ to avoid biasing the results from singly large random effects. The bias provides a measure of how close our estimate is to the true value as per our data generating mechanisms. The coverage will demonstrate how often our confidence intervals surrounding our estimate actually include this true value. The Empirical Standard Error will show us how precise our estimates are.

Table 1: Performance Measures to be taken at each time point

Performance Measure	Estimation	SE
Bias	$\hat{\theta}(t) = \frac{1}{N}\sum_{i=1}^{N}\theta_i(t) - \theta$	$\hat{\theta}_{SE}(t) = \sqrt{\frac{1}{N(N-1)}\sum_{i=1}^{N}\left(\theta_i(t) - \hat{\theta}(t)\right)^2}$
EmpSE	$\hat{E}(t) = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(\theta_i(t) - \hat{\theta}(t)\right)^2}$	$\hat{E}_{SE}(t) = \frac{\hat{E}(t)}{\sqrt{2(N-1)}}$
Coverage	$\hat{C}(t) = \frac{1}{N}\sum_{i=1}^{N} I\left(\theta_{i,L}(t) \leq \theta \leq \theta_{i,U}(t)\right)$	$\hat{C}_{SE}(t) = \frac{\hat{C}(t)\left(1-\hat{C}(t)\right)}{N}$

2.9 Software

All analysis was done in `R 3.6.3` (Team, n.d.) using the various `tidyverse` packages (Wickham 2017), Kaplan-Meier estimates were found using the `survival` package (Therneau 2020), Pseudo-Observations were evaluated with the `pseudo` package (Perme, Gerster, and Rodrigues 2017), and the results app was developed using `shiny`(Chang et al. 2020). The code used for this simulation study is available on Github and the results can be seen in a shiny app

3 Results

[**Results shown here are new and improved from the previous version. No highlighting is shown**] Figure 1 shows the results when censoring is independent of covariates ($\gamma = 0$). The LW method

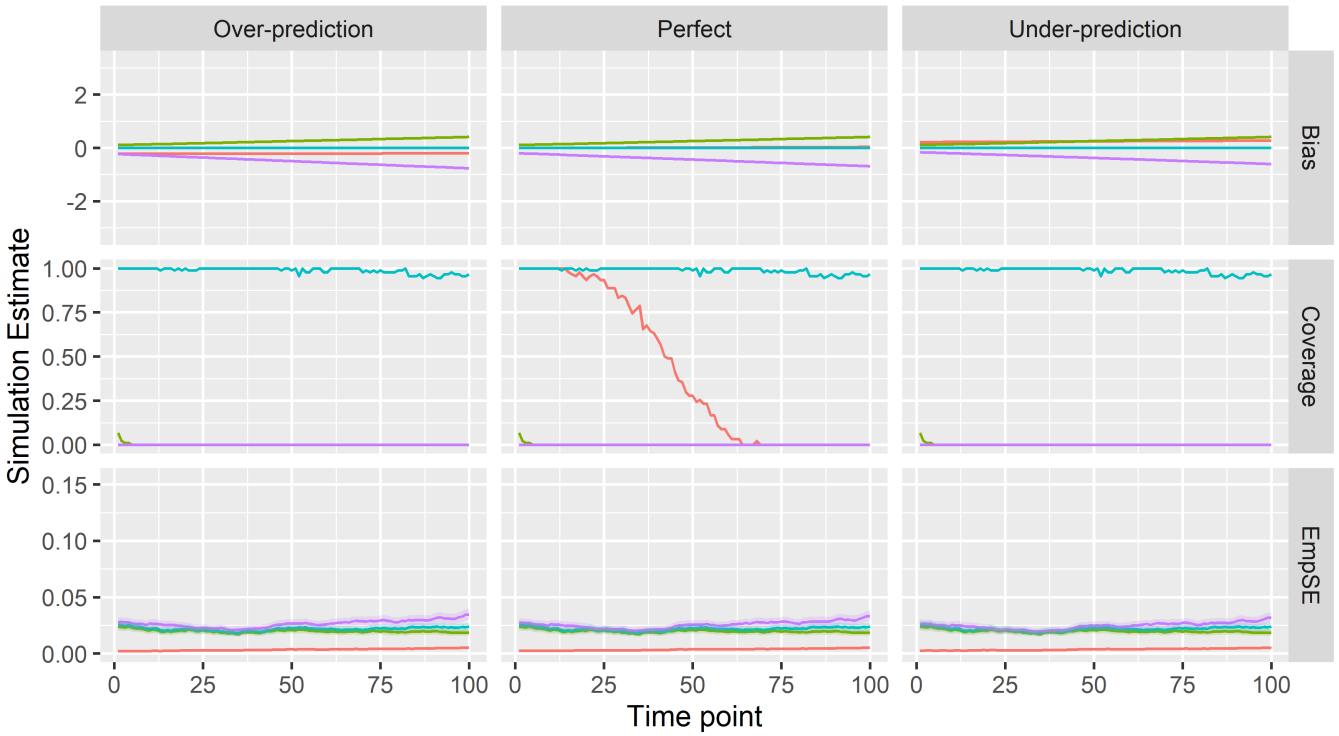


Figure 1: Bias, Coverage and Empirical Standard Error for the Over-estimating, Perfect and Under-Estimating models across all four methods when $\beta = 1$, $\gamma = 0$ and $\eta = {}^{1}/_{2}$. Confidence Intervals are included in the plot, but are tight around the estimate

provides strong coverage across the entire timeframe and miminal bias. The absolute bias for PO and LU

increases over time with PO under-reporting the correct value and LO over-reporting. KM bias remains constant across the timeframe, but for the imperfect models, is constantly under- or over-reported. LU and PO also provide minimal coverage at all time points, whereas KM covers perfect in the early stages of the Perfect Model with coverage dropping off as time progresses. Empirical Standard Error is clse to 0 for all models. Figure 2 shows the results when censoring and the event-of-interest have the same individual

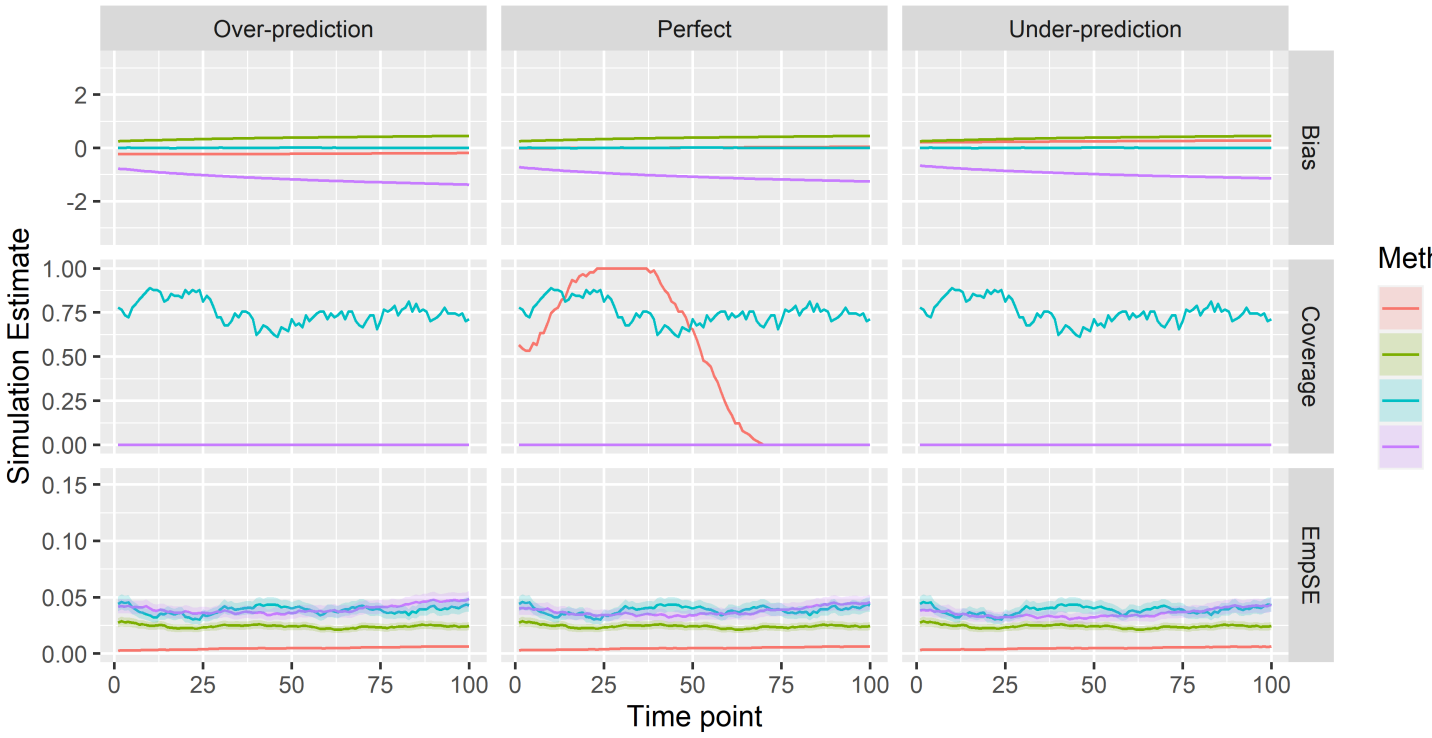


Figure 2: Bias, Coverage and Empirical Standard Error for the Over-estimating, Perfect and Under-Estimating models across all four methods when $\beta = 1$, $\gamma = 1$ and $\eta = {}^{1}/_{2}$. Confidence Intervals are included in the plot, but are tight around the estimate

effects ($\beta = \gamma = 1$). The LW method provides strong coverage across the entire timeframe and miminal bias, although this coverage is reduced compared to the previous set of results shown (approximately 75% throughout). Once again, the absolute bias for PO and LU increases over time, however the under-reporting for PO is much more strongly pronounced. KM bias behaves similarly but for coverage, it starts off at around 50% coverage reaches a peak of full coverage approximately 25% of the way through the timeframe. Figure 3 shows the results when censoring and the event-of-interest have opposite individual effects ($\beta = 1, \gamma = -1$). The bias results are similar to those when censoring is independent. A difference here is that coverage begins greater than zero for the KM, LU and PO methods, but quickly drops to 0 before the 25% time point. For LW, the coverage appears to reduce to around 80% by the end of the time point.

4 Discussion

Weighting = Good.

Not Weighting = Bad.

limitation: Maybe the "True" θ for the under and over predictions were wrong and that would explain the low Coverage.

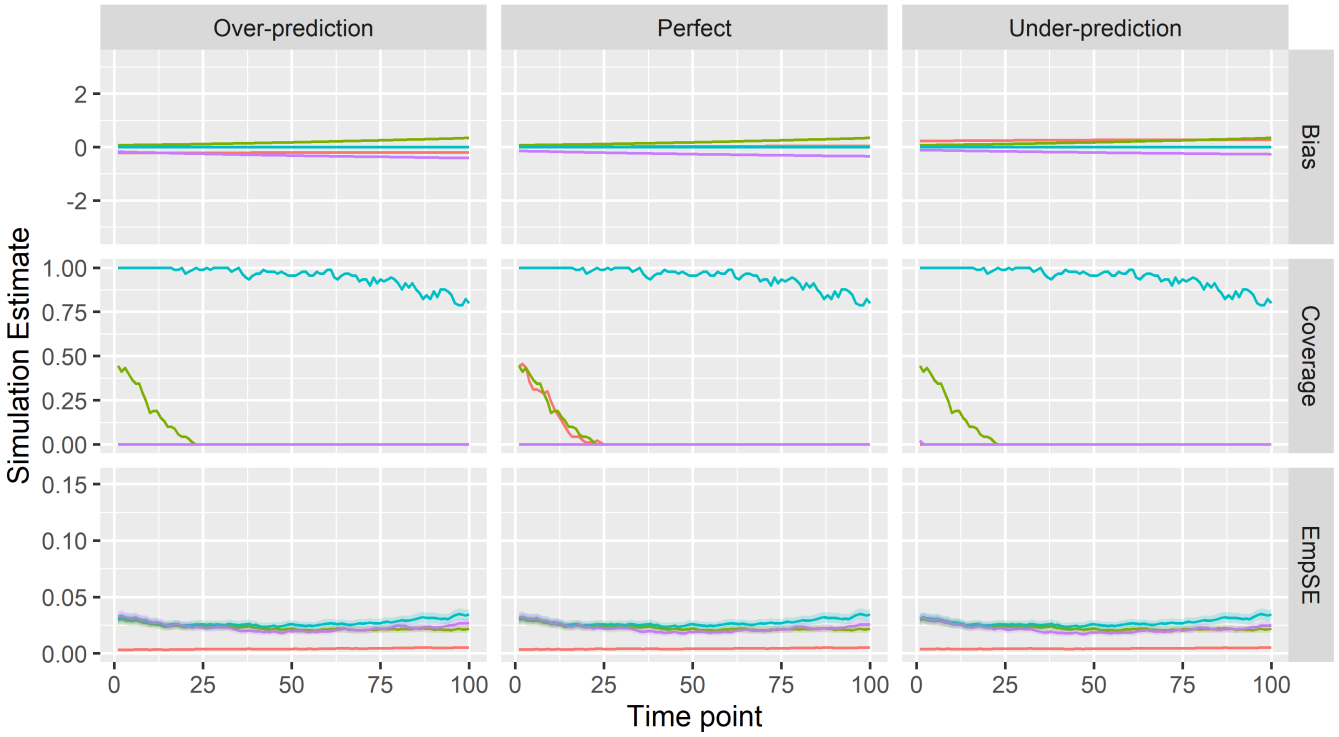


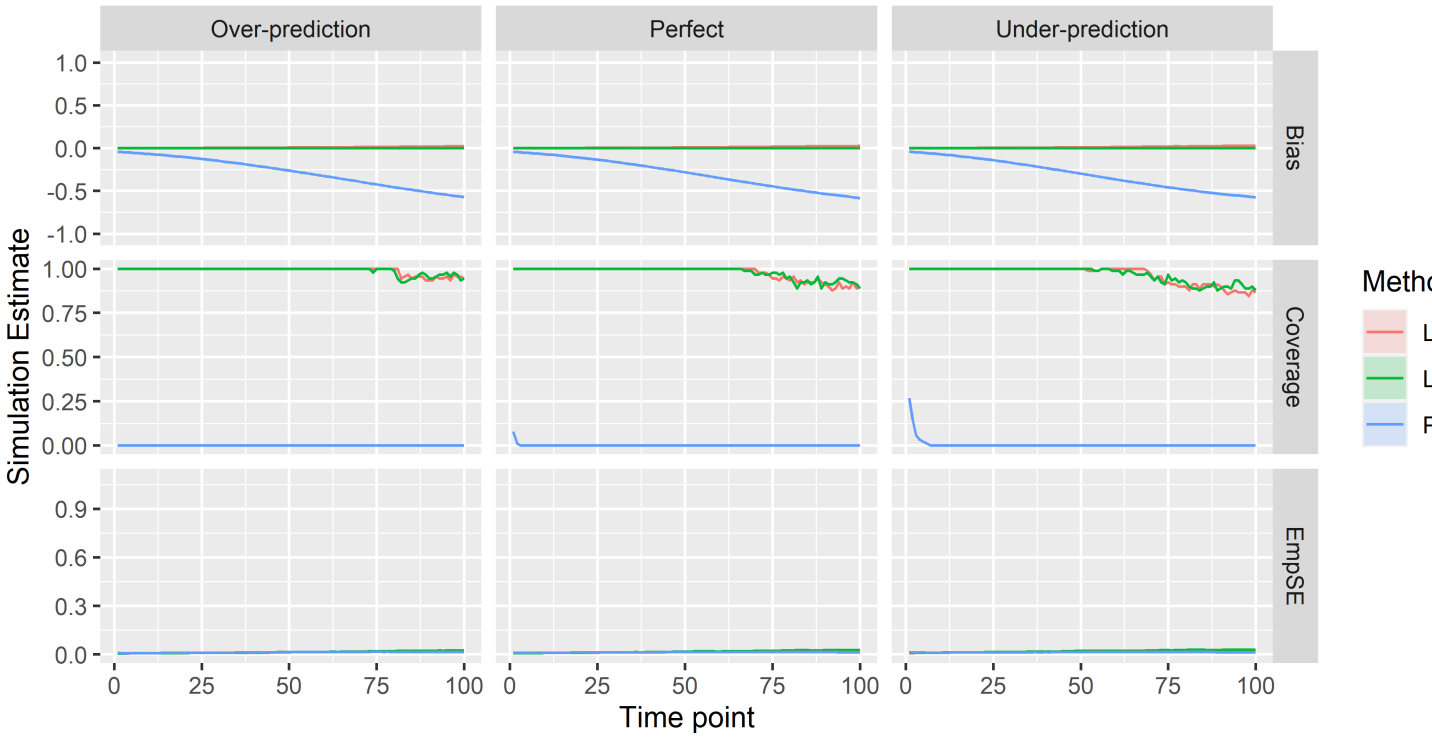
Figure 3: Bias, Coverage and Empirical Standard Error for the Over-estimating, Perfect and Under-Estimating models across all four methods when $\beta = 1$, $\gamma = -1$ and $\eta = {}^{1}/_{2}$. Confidence Intervals are included in the plot, but are tight around the estimate

A Supplementary Material

A.1 Calibration Slope

The main purpose of this paper was to assess the evaluation of calibration-in-the-large at different time points in a time-to-event clinical prediction model. Along with calibration-in-the-large, various methods of calibration can also produce measures of calibration slope. Calibration slope provides an insight into how well the model predicts outcomes across the range of predictions. In an ideal model, the calibration slope would be 1. The Logistic Weighted, Logistic Unweighted and Pseudo-Observation methods described above can provide estimates of the calibration slope. For each of these methods, we first estimate the calibration-in-the-large as above, using a predictor as an offset, then we use this estimate as an offset to predict the calibration slope (without an intercept term).

A.1.1 Results



Results currently show bias/coverage/EmpsE away from 0, rather than 1. Needs fixing. Oops.

A.1.2 Discussion

Brief discussion, much briefer than the main points.

References

Andersen, Per Kragh, and Maja Pohar Perme. 2010. "Pseudo-Observations in Survival Analysis." *Statistical Methods in Medical Research* 19 (1): 71–99. https://doi.org/10.1177/0962280209105020.

Burton, Andrea, Douglas G. Altman, Patrick Royston, and Roger L. Holder. 2006. "The Design of Simulation Studies in Medical Statistics." *Statistics in Medicine* 25 (24): 4279–92. https://doi.org/10.1002/sim.2673.

Calster, Ben Van, Daan Nieboer, Yvonne Vergouwe, Bavo De Cock, Michael J. Pencina, and Ewout W. Steyerberg. 2016. "A Calibration Hierarchy for Risk Models Was Defined: From Utopia to Empirical Data." *Journal of Clinical Epidemiology* 74 (June): 167–76. https://doi.org/10.1016/j.jclinepi.2015.12.005.

Chang, Winston, Joe Cheng, J. J. Allaire, Yihui Xie, Jonathan McPherson, RStudio, jQuery Foundation (jQuery library and jQuery UI library), et al. 2020. "Shiny: Web Application Framework for R."

Crowson, Cynthia S., Elizabeth J. Atkinson, and Terry M. Therneau. 2016. "Assessing Calibration of Prognostic Risk Scores." *Statistical Methods in Medical Research* 25 (4): 1692–1706. https://doi.org/10.1177/0962280213497434.

de la Iglesia, B., J. F. Potter, N. R. Poulter, M. M. Robins, and J. Skinner. 2011. "Performance of the ASSIGN Cardiovascular Disease Risk Score on a UK Cohort of Patients from General Practice." *Heart* 97 (6): 491–99. https://doi.org/10.1136/hrt.2010.203364.

Gerds, Thomas A., and Martin Schumacher. 2006. "Consistent Estimation of the Expected Brier Score in General Survival Models with Right-Censored Event Times." *Biometrical Journal* 48 (6): 1029–40. https://doi.org/10.1002/bimj.200610301.

Goff, David C., Donald M. Lloyd-Jones, Glen Bennett, Sean Coady, Ralph B. D'Agostino, Raymond Gibbons, Philip Greenland, et al. 2014. "2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines." *Circulation* 129 (25 suppl 2): S49–S73. https://doi.org/10.1161/01.cir.0000437741.48606.98.

Han, Xiaoxia, Yilong Zhang, and Yongzhao Shao. 2017. "On Comparing Two Correlated C Indices with Censored Survival Data." *Statistics in Medicine* 36 (25): 4041–9. https://doi.org/10.1002/sim.7414.

Hippisley-Cox, Julia, Carol Coupland, and Peter Brindle. 2017. "Development and Validation of QRISK3 Risk Prediction Algorithms to Estimate Future Risk of Cardiovascular Disease: Prospective Cohort Study." *BMJ* 357 (May). https://doi.org/10.1136/bmj.j2099.

Hippisley-Cox, Julia, Carol Coupland, Yana Vinogradova, John Robson, Margaret May, and Peter Brindle. 2007. "Derivation and Validation of QRISK, a New Cardiovascular Disease Risk Score for the United Kingdom: Prospective Open Cohort Study." *BMJ (Clinical Research Ed.)* 335 (7611): 136. https://doi.org/10.1136/bmj.39261.471806.55.

Liu, Xinhua, Zhezhen Jin, and Joseph H. Graziano. 2016. "Comparing Paired Biomarkers in Predicting Quantitative Health Outcome Subject to Random Censoring." *Statistical Methods in Medical Research* 25 (1): 447–57. https://doi.org/10.1177/0962280212460434.

Morris, Tim P., Ian R. White, and Michael J. Crowther. 2019. "Using Simulation Studies to Evaluate Statistical Methods." *Statistics in Medicine* 38 (11): 2074–2102. https://doi.org/10.1002/sim.8086.

Perme, Maja Pohar, and Per Kragh Andersen. 2008. "Checking Hazard Regression Models Using Pseudo-Observations." *Statistics in Medicine* 27 (25): 5309–28. https://doi.org/10.1002/sim.3401.

Perme, Maja Pohar, Mette Gerster, and Kevin Rodrigues. 2017. "Pseudo: Computes Pseudo-Observations for Modeling."

Riley, Richard D., Danielle van der Windt, Peter Croft, and Karel G. M. Moons. 2019. *Prognosis Research in Healthcare: Concepts, Methods, and Impact*. First. Oxford University Press.

Royston, Patrick. 2014. "Tools for Checking Calibration of a Cox Model in External Validation: Approach Based on Individual Event Probabilities:" *The Stata Journal*, December. https://doi.org/10.1177/1536867X1401400403.

———. 2015. "Tools for Checking Calibration of a Cox Model in External Validation: Prediction of Population-Averaged Survival Curves Based on Risk Groups." *The Stata Journal* 15 (1): 275–91. https://doi.org/10.1177/1536867X1501500116.

Royston, Patrick, and Douglas G. Altman. 2013. "External Validation of a Cox Prognostic Model: Principles and Methods." *BMC Medical Research Methodology* 13 (1): 33. https://doi.org/10.1186/1471-2288-13-33.

Spitoni, Cristian, Violette Lammens, and Hein Putter. 2018. "Prediction Errors for State Occupation and Transition Probabilities in Multi-State Models." *Biometrical Journal. Biometrische Zeitschrift* 60 (1): 34–48. https://doi.org/10.1002/bimj.201600191.

Steyerberg, Ewout W. 2008. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer Science & Business Media.

Team, R Core. n.d. "R: A Language and Environment for Statistical Computing." Vienna, R Foundation for Statistical Computing, Vienna, Austria.

Therneau, Terry. 2020. "A Package for Survival Analysis in R," March, 89.

Wickham, Hadley. 2017. "The Tidy Tools Manifesto." https://cran.r-project.org/web/packages/tidyverse/vignettes/manifest

Wildscop. 2013. "Biostatistics and Epidemiology with R: Weighted Logistic Regression in R, SPSS, Stata." *Biostatistics and Epidemiology with R.*