Received: 24 May 2021 | Revised: 15 November 2021 | Accepted: 16 November 2021

DOI: 10.1002/sim.9275

RESEARCH ARTICLE

Statistics in Medicine WILEY

Minimum sample size calculations for external validation of a clinical prediction model with a time-to-event outcome

Richard D. Riley[1] | Gary S. Collins[2,3] | Joie Ensor[1] | Lucinda Archer[1] | Sarah Booth[4] | Sarwar I. Mozumder[4] | Mark J. Rutherford[4] | Maarten van Smeden[5] | Paul C. Lambert[4,6] | Kym I. E. Snell[1]

[1]Centre for Prognosis Research, School of Medicine, Keele University, Keele, UK

[2]Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK

[3]NIHR Oxford Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, Oxford, UK

[4]Biostatistics Research Group, Department of Health Sciences, George Davies Centre, University of Leicester, Leicester, UK

[5]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, University of Utrecht, Utrecht, The Netherlands

[6]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

Correspondence
Richard D. Riley, Centre for Prognosis Research, School of Medicine, Keele University, Keele, Staffordshire ST5 5BG, UK.
Email: r.riley@keele.ac.uk

Funding information
Cancer Research UK, Grant/Award Numbers: C14183/A29739, C41379/A27583, C49297/A27294; NIHR Biomedical Research Centre, Oxford; Research Trainees Coordinating Centre, Grant/Award Number: NIHR300100

Abstract

Previous articles in *Statistics in Medicine* describe how to calculate the sample size required for external validation of prediction models with continuous and binary outcomes. The minimum sample size criteria aim to ensure precise estimation of key measures of a model's predictive performance, including measures of calibration, discrimination, and net benefit. Here, we extend the sample size guidance to prediction models with a time-to-event (survival) outcome, to cover external validation in datasets containing censoring. A simulation-based framework is proposed, which calculates the sample size required to target a particular confidence interval width for the calibration slope measuring the agreement between predicted risks (from the model) and observed risks (derived using pseudo-observations to account for censoring) on the log cumulative hazard scale. Precise estimation of calibration curves, discrimination, and net-benefit can also be checked in this framework. The process requires assumptions about the validation population in terms of the (i) distribution of the model's linear predictor and (ii) event and censoring distributions. Existing information can inform this; in particular, the linear predictor distribution can be approximated using the *C-index* or Royston's D statistic from the model development article, together with the overall event risk. We demonstrate how the approach can be used to calculate the sample size required to validate a prediction model for recurrent venous thromboembolism. Ideally the sample size should ensure precise calibration across the entire range of predicted risks, but must at least ensure adequate precision in regions important for clinical decision-making. Stata and R code are provided.

KEYWORDS

calibration, external validation, prediction model, sample size, time-to-event & survival data

1 | INTRODUCTION

Clinical prediction models aim to inform the diagnosis and prognosis of individuals. They utilize multiple predictors in combination to predict an individual's outcome value (for continuous outcomes) or outcome event risk (for binary or time-to-event outcomes).[1-3] Before such models are considered for use in clinical practice, it is important to evaluate their predictive accuracy in new data. This usually requires an external validation study (and typically more than one), where a prediction model is applied to new individuals that were not part of the model development dataset (and may even be from a different population), so that predicted outcomes can be compared with observed outcomes.

We recently published two articles in *Statistics in Medicine* that describe how to calculate the minimum sample size for external validation of a clinical prediction model with either a continuous outcome,[4] or with a binary outcome.[5] The sample size criteria is based on ensuring precise estimation of key measures of a model's predictive performance in external validation, including calibration and—for binary outcomes—discrimination and net benefit. Calibration refers to the agreement between observed and predicted values, and can be minimally summarized by the validation sample by the calibration-in-the-large (ie, mean observed compared to mean predicted value for continuous outcomes, or mean observed compared to mean predicted event risk for binary outcomes) and calibration slope (ie, the slope of a linear or logistic regression model that includes the model's linear predictor as the only covariate). Discrimination refers to how a model's linear predictor (ie, predicted risks of the outcome event) separates between those with and without the outcome event, and is usually quantified by the *C-index*. Net benefit is a measure of clinical utility, and provides a weighted measure of the potential benefits to harms of using a model at a defined clinical risk threshold.[6]

Many prediction models are developed using time-to-event (survival) models (such as QRISK3[7] and QCOVID[8]), and require external validation in datasets that will naturally containing censoring. As previous publications focused on continuous and binary outcomes,[4,5,9,10] we now address the issue of how to calculate the minimum sample size required for external validation of prediction models with a time-to-event outcome. We focus predominantly on the sample size required to precisely estimate the calibration slope as our previous articles,[4,5,9] and those of others,[11,12] have shown in applied examples and simulation studies that the calibration slope is usually the measure that requires the largest sample size to estimate precisely. However, we also show that other measures such as discrimination and net benefit can be examined in the proposed framework. The remaining article outline is as follows. In Section 2 we explain how calibration can be examined at a particular time point using pseudo-observations. In Section 3 we propose a simulation-based framework for identifying the sample size that targets a precise estimate of calibration (in terms of calibration slope and calibration curves) at a particular time point, which requires the user to specify the anticipated overall event risk at that time, the distribution of the prediction model's linear predictor and the censoring and event distribution. Section 4 provides an illustrative example aiming to ensure precise calibration across the entire spectrum of predicted risks. Section 5 then considers sample size in the context of potential clinical decision making, where regions of predicted risk may be more important to focus on. Section 6 extends to other issues including unknown linear predictor distributions, multiple time points and competing risks, and Section 7 concludes with discussion.

2 | ESTIMATING CALIBRATION AT A PARTICULAR TIME POINT

We assume that the developed prediction model enables the calculation of predicted risks over time for each individual in the validation dataset. For example, a regression-based time-to-event prediction model would provide all the regression coefficients and—crucially—the baseline survival (or baseline event) probability at the time points of interest for prediction. Then, for an individual (i) their predicted risk ($\widehat{F}_i(t)$) of having the event by time t can be calculated. For example, a model developed using a proportional hazards regression will lead to risk predictions of the form,

$$\begin{aligned}\widehat{F}_i(t) &= 1 - \widehat{S}_i(t) \\ &= 1 - \widehat{S}_0(t)^{\exp\left(\widehat{\beta}_1 X_{1i} + \widehat{\beta}_2 X_{2i} + \widehat{\beta}_3 X_{3i} + \cdots + \widehat{\beta}_k X_{ki}\right)} \\ &= 1 - \widehat{S}_0(t)^{\exp(\mathrm{LP}_i)} \end{aligned} \tag{1}$$

where $\widehat{S}_i(t)$ is the survival probability by time t, and $\widehat{S}_0(t)$ is the baseline survival probability by time t (where "baseline" refers to individuals whose LP_i is zero). The model's linear predictor (LP_i) is the $\widehat{\beta}_1 X_{1i} + \widehat{\beta}_2 X_{2i} + \widehat{\beta}_3 X_{3i} + \cdots + \widehat{\beta}_k X_{ki}$

component of the model, and is sometimes referred to as the prognostic index.[13] As mentioned, we assume that $\widehat{S}_0(t)$ (or similarly $\widehat{F}_0(t)$) has been reported by the model developers for each time point of interest so that absolute risk predictions can be calculated. If only the linear predictor (eg, from a Cox model) is provided, it is not possible to make risk predictions and so the reported model is not fit for purpose.[13]

For validation, we assume that there is a key time point for which accurate risk predictions are required. For example, QRISK3 focuses primarily at predicting risk of cardiovascular disease by 10 years and subsequent clinical decision making is based on these risks. Hence, a calibration assessment in the validation study should examine whether the predicted risks agree with the observed risks by this time point. Although calibration could be checked across all time points simultaneously (eg, to measure the calibration slope averaged across the whole follow-up period),[14] this is not especially informative to potential users of the model such as clinicians and patients, as for decision and communication purposes the event risk by a particular time (eg, 10 years) is more important.

2.1 | Situations with no censoring before the time horizon for prediction

If there is no censoring by the time point of interest, then the outcome event status (usually presence or absence of a certain health outcome) is truly known for all individuals in the validation dataset. Then, the calibration slope at time point t can be estimated by fitting a calibration model in the form of the following logistic regression model,[5]

$$\operatorname{logit}(p_i) = \alpha + \beta \operatorname{logit}\left(\widehat{F}_i(t)\right) \tag{2}$$

where p_i is the underlying risk of the event occurring by time t for individual i in the validation dataset, β is the calibration slope, and $\widehat{F}_i(t)$ is the predicted risk of the event by time t as derived from the existing prediction model.

2.2 | Situations with censoring before the time horizon for prediction

If there is censoring before the time point of interest in the validation dataset, then the true outcome event status is unknown for the censored individuals. This makes it problematic to directly examine the calibration of predicted risks with observed risks at the time point of interest. A common approach is to create risk groups (eg, 10 groups defined by tenths of predicted risk), and to plot the average predicted risk against the observed (1—Kaplan-Meier) risk for each group. However, this is unsatisfactory, as the number of risk groups and the thresholds used to define them are arbitrary, and what we desire is a plot that examines calibration across the entire range of predicted risks (0-1).

To address this, the use of pseudo-observations (or pseudo-values) has been proposed,[15-18] which are derived using the jackknife (ie, leave-one-out) estimator, and give pseudo-observed event probabilities for each individual accounting for non-informative right censoring. We denote these pseudo-observations by $\widetilde{F}_i(t)$ and provided that censoring is independent of covariates, they yield unbiased estimates of the true $F_i(t)$.[16] In brief, the pseudo-observation for individual i is calculated in the validation dataset as,

$$\widetilde{F}_i(t) = nF_{KM}(t) - \left[(n-1)F_{KM(-i)}(t)\right],$$

where $F_{KM}(t) = 1 - S_{KM}(t)$ is the Kaplan-Meier (or another nonparametric) estimate of the cumulative incidence at time t using all n individuals in the validation dataset, and $F_{KM(-i)}(t)$ is the cumulative incidence estimate recalculated on the $n-1$ individuals after removing individual i. Note that the $\widetilde{F}_i(t)$ are not constrained to fall within the range 0 to 1, but the key is that $\mathrm{E}\left(\widetilde{F}_i(t)\right) = F_i(t)$.[19] This allows the generation of calibration plots at time point t (avoiding risk grouping) and the calculation of the corresponding calibration slope and calibration curves, as shown by Royston.[17] In particular, a generalized linear model can be fitted with $\widetilde{F}_i(t)$ as the outcome response, and the expected value $\left(\mathrm{E}\left(\widetilde{F}_i(t)\right)\right)$ modelled using a particular link function. Royston suggests using a complementary log-log link function,[17] such that.

$$\ln\left(-\ln\left(1-\mathrm{E}\left(\widetilde{F}_i(t)\right)\right)\right) = \alpha + \beta\ \ln\left(-\ln\left(1-\widehat{F}_i(t)\right)\right) \tag{3}$$

where β is the calibration slope, and $\widehat{F}_i(t)$ is the predicted risk calculated from the existing prediction model for individual i at time t. This is equivalent to modelling the log of the cumulative hazard function ($H(t)$), as $-\ln\left(1-\widehat{F}_i(t)\right)=\widehat{H}(t)$, and this scale is often used in survival modelling. Other link functions (eg, logit) could be chosen to model $\mathrm{E}\left(\widetilde{F}_i(t)\right)$), and we return to this issue in Section 5.

Model fitting is discussed by Andersen and Perme,[16] who suggest the use of generalized estimating equations followed by a sandwich estimator for the variance of parameter estimates (which is needed to account for the correlation of the pseudo-observations themselves). This is implemented in various software modules, including *stcoxcal* for Stata,[17] specifically developed to allow researchers to fit Equation (3) and estimate the calibration slope of time-to-event prediction models in new data. More generally, pseudo-observations can be derived directly (eg, using *stpci* in Stata,[20] or using the *pseudo* or *prodlim* packages in R[21]) and then generalized linear models fitted (eg, using *glm* in Stata or *geese* in R), with robust standard errors calculated.

Equation (3) examines a linear calibration, but the actual relationship between observed and predicted risks may be non-linear. Hence, it is also important to estimate a flexible calibration curve with a 95% confidence interval,[12,22] for example by fitting a spline or fractional polynomial function and using robust standard errors (or bootstrapping) to derive the confidence interval. Alternatively, the *stcoxcal* Stata package by Royston (and *calPlot* in the *pec* package in R*)* uses a nearest neighbor smoothed running line to produce the calibration curve and its confidence interval,[17,23] and we use this approach in our article.

3 | CALCULATING THE SAMPLE SIZE TO ESTIMATE THE CALIBRATION SLOPE PRECISELY

We now consider how to calculate the sample size for a cohort study aiming to precisely estimate the calibration slope, assuming that the validation dataset will contain observations censored before the time point of interest for prediction. Our approach requires the user to provide some information and to make some assumptions, as in any sample size calculation.

3.1 | Specify the anticipated linear predictor distribution

Fundamentally, the model's anticipated linear predictor (LP_i) distribution in the validation population must be specified. For example, if the validation population is anticipated to be similar to that used for model development, then it makes sense to assume the linear predictor distribution for validation will be similar in shape to that observed for development. Sometimes this distribution is presented in the model development article as a histogram (either for all individuals, or for each of the events and non-events groups separately), or perhaps the developers report it was approximately normally distributed and give the mean and SD. If information about the linear predictor distribution is unavailable, then the model development team could be asked directly to provide details. We illustrate this in our example later.

If Royston's D statistic is presented,[24,25] then the SD (σ) of the linear predictor distribution can be estimated by,

$$\widehat{\sigma}=\mathrm{D}/(\sqrt{8/\pi}) \tag{4}$$

assuming the linear predictor is normally distributed, which may be a strong assumption. The mean of this normal distribution can be simply set to zero, as without loss of generality it can be assumed centered by its original mean. The only impact of this is on the magnitude of the baseline hazard for simulating survival times (see below).

Jinks et al[26] propose the following equation, based on empirical evidence, for approximating Royston's D when only Harrell's *C-index* is reported[2,27]:

$$\mathrm{D}=5.50(C-0.5)+10.26(C-0.5)^3 \tag{5}$$

Thus, if only Harrell's *C-index* is reported, we can use Equation (5) to approximate Royston's D statistic, and then apply Equation (4) to estimate the SD of the linear predictor distribution assuming it is normally distributed. When using the *C-index* or Royston's D statistic reported from a model development study, ideally their values should have been adjusted for optimism due to any overfitting. For example, this could be the *C-index* after optimism-adjustment based on results from Harrell's bootstrapping approach[2]; after a penalized regression approach has been used; or based on the *C-index* estimated in any independent validation (test) datasets (providing the data used for independent validation is reflective of the target population in whom the model is to be further validated). Note that there are many other types of discrimination measures for survival data, such as Uno's C which (unlike Harrell's *C-index*) does not ignore censored observations,[28] and further work is needed for how to use them to approximate Royston's D and the distribution of the linear predictor.

If no information is available for informing the distribution of the linear predictor, and also when the validation population is expected to be very different to that used for development, then a pilot study could be used to observe the distribution more closely.[5]

3.2 | Specify the anticipated distribution of survival and censoring times

The assumed distribution of survival times must also be specified for the validation population. To make the approach practical, here we will assume survival times follow an exponential survival distribution, with baseline rate parameter (λ) chosen to ensure $F(t)$ (and thus $S(t)$) in the validation population are as anticipated for the time point of interest for prediction. Crucially, λ needs to be chosen conditional on the effect of the linear predictor. For example, we suggest using trial and error (or a simple one parameter non-linear optimization strategy) to identify the value of λ that gives the desired $F(t)$ at the time point of interest in a large simulated dataset of survival times from an assumed exponential distribution with baseline rate parameter λ and a single covariate for LP_i whose corresponding parameter value is 1 (ie, assuming risks from the existing prediction model will have a perfect calibration slope across all time points).

The exponential distribution assumes a constant hazard over time and so will, in most situations, be an overly simple approximation of the truth. More complex survival functions could be assumed if more information is available (eg, if the model development data were available), although we anticipate the exponential distribution will be a pragmatic compromise for most users. The impact of different survival distributions is considered in Section 5.

The censoring distribution must also be specified. For example, an exponential distribution could be assumed with censoring rate matching that reported for the development dataset or in similar studies in the same field, to ensure the censoring proportion matches that anticipated by the time point of interest for prediction. Our example illustrates this later. However, it is also important for the censoring distribution to reflect how and when participants will enter the cohort study. Most censoring is administrative (eg, end of study follow-up), and so a study that recruits everyone at the study onset is likely to have a different censoring pattern than a study that recruits individuals gradually over time or at specific intervals. We examine the impact of censoring distributions in Section 5.

3.3 | Specify the target SE for the calibration slope

Finally, the target SE for the calibration slope must be specified. This choice is subjective, and potentially context specific as what constitutes "precise" depends on the magnitude and range of risk predictions arising from the model, which are defined by the outcome event rate and the linear predictor distribution. It may also depend on the potential clinical implementation of the model; in particular, it may be more important to have precise calibration in some ranges of the calibration plot (eg, predicted risks between 0.05 and 0.20) where clinical decision thresholds are likely to be made, compared to other ranges (eg, predicted risks between 0.5 and 1.0) where miscalibration is less of a concern. As in our previous articles,[4,5,9] we recommend starting with a target SE of 0.051, which corresponds to a narrow confidence interval for the calibration slope of 0.9 to 1.1. The aim is to identify the sample size that is expected to (ie, will on average) give this target SE (and thus target confidence interval width), and we describe how to do this below.

3.4 | Use a simulation-based approach to estimate the required sample size

After the set-up phase, further steps are required using a simulation-based framework,[29,30] which proceeds iteratively until the sample size is identified that meets the target SE value on average. The steps are outlined in Figure 1. The process starts by assuming the model will be well calibrated in the validation dataset, which is a pragmatic place to start. This could be changed, for example if the calibration slope is anticipated to be less than 1 upon validation due to suspected overfitting

- Step 1: Set-up process:
  - Specify the time point of interest for checking calibration performance.
  - Specify the model's anticipated linear predictor (LP_i) distribution in the validation population. See main text for guidance on how to choose this, for example based on reported histograms, or the D statistic and *C-index.*
  - Specify the anticipated overall $F(t)$ (or $S(t)$) in the validation population at the time point of interest.
  - Specify the distribution of survival times in the validation population, conditional on the effect of LP_i. In the absence of other evidence, we suggest assuming an exponential distribution for simplicity, but other distributions should be specified if distributional information exists
  - Specify the effect of LP_i. We suggest assuming the existing model will have good calibration, such that the log hazard ratio for the effect of LP_i is 1 (ie, calibration slope is 1). Ensure the chosen parameter(s) of the distribution (eg, baseline rate parameter for exponentially distributed survival times) correspond to the anticipated $F(t)$ (or $S(t)$) at the time point of interest is met, conditional on assuming the calibration slope is 1.
  - Specify the assumed distribution of censoring times in the validation population, and maximum follow-up time. We suggest assuming an exponential distribution in the absence of other information.
  - Specify the target value for the SE, of the calibration slope (eg, 0.051).
- Step 2: Specify a starting sample size and generate a dataset containing the same number of individuals as this sample size.
- Step 3: For each individual i in the dataset, simulate values of LP_i from the assumed linear predictor distribution.
- Step 4: For the time point of interest (prediction time horizon), generate values of $\hat{F}_i(t)$ for each individual by applying the existing prediction model equation. For example, the existing model will typically have an equation of the form $\hat{F}_i(t) = 1 - \hat{S}_0(t)^{\exp(\mathrm{LP}_i)}$ where $\hat{S}_0(t)$ is the model's reported baseline survival probability, and LP_i is the value of the linear predictor for individual i, generated from step 3.
- Step 5: Randomly generate survival times for each individual according to the assumed distribution from step 1 and conditional on their LP_i value. For example, using the *survsim* package in Stata,[29] or *simsurv* in R.[30] For each individual, set their outcome status to be 1 (ie, event) and their follow-up time to be their survival time.
- Step 6: Randomly generate a censoring time for each individual under the censoring distribution assumed in step 1. Also specify the maximum follow-up time for all individuals in the validation study. For those individuals whose survival time (from step 5) is later than their generated censoring time or the maximum follow-up time, change their event status to 0 (ie, no event) and change their follow-up time to their censoring time or the maximum follow-up time (whichever is earlier).
- Step 7: For the chosen time point of interest for prediction, generate pseudo-observations and fit Equation (3) to estimate the calibration slope and its SE, for example by using the *stcoxcal* package in Stata. Store the results (and those for any other measures of interest, eg, net benefit).
- Step 8: Repeat steps 2 to 7 many times (eg, 1000), and each time store the obtained estimates and SEs of the calibration slope.
- Step 9: Summarize the mean calibration slope (simply to check it equals 1 as assumed in step 1) and the mean SE. If the mean SE equals the targeted value specified in step 1, then the sample size in step 2 is the required sample size. Otherwise, repeat steps 2 to 9 with a different chosen sample size.

Once the required sample size is identified, plot flexible calibration curves (and confidence intervals) for the simulated datasets, to ascertain whether their spread (and confidence interval width) appears acceptable. Particular attention might be given to regions of risk that are key to clinical decision making.

FIGURE 1 Suggested process to identify the sample size required to precisely estimate the calibration slope (and subsequent calibration curves) at a key time point when externally validating the performance of a prediction model with a time-to-event outcome

during model development; however, this usually leads to lower sample sizes and therefore assuming a slope of 1 is more conservative.[4,5,9] The process focuses on precise estimation of calibration slope, but precision of other performance measures should also be checked, such as the calibration-in-the-large, the *C-index*, and net benefit. If the sample size is adequate for precisely estimating the calibration slope, then usually it will be adequate for these other measures too.[5]

Once this sample size has been identified, we recommend producing calibration plots for 10-20 of the already simulated datasets of this sample size, including a flexible calibration curve with a 95% confidence interval for each,[12,22] as described in Section 2. The calibration curve and its confidence interval on these plots can then be inspected, to ascertain whether the precision of the curve appears generally acceptable, or whether a smaller or larger sample size is actually needed. Similarly, the empirical distribution of calibration curves could be displayed by presenting the entire set of simulated calibration curves on the same graph (without confidence intervals). This helps to reveal the potential range of calibration curves that might be observed in practice, and if variability is too high then a larger sample size is needed. Our applied example will illustrate these ideas.

4 | APPLIED EXAMPLE

Ensor et al developed a prognostic time-to-event model to predict the risk of a recurrent venous thromboembolism (VTE) following cessation of therapy for a first VTE.[31] The model included predictors of age, gender, site of first clot, D-dimer level, and the lag time from cessation of therapy until measurement of D-dimer (often around 30 days). These predictors corresponded to six parameters in the model, which was developed using the flexible parametric survival modeling framework of Royston and Parmar.[32,33]

A key time point for prediction was 3 years, and the final model equation for calculating predicted risks was given as,

$$\widehat{F}_i(3) = 1 - \widehat{S}_0(3)^{\exp(\mathrm{LP}_i)}$$

where $\widehat{S}_0(3) = 0.9983$, and $\mathrm{LP}_i = [(-0.0105 \times \mathrm{Age}) + (0.545 \times \mathrm{Male}) + (1.735 \times \mathrm{Site{:}\ Proximal\ DVT}) + (1.756 \times \mathrm{Site{:}\ PE}) + (0.701 \times \ln[\mathrm{D\ dimer}]) + (-0.291 \times \ln[\mathrm{lag\ time}])]$.

At the design stage of a validation study, we must now determine the minimum sample size needed to precisely estimate the calibration slope of this model when used to predict risk at 3 years, in order to precisely examine calibration across the whole range of predicted risks. Assuming the validation population will be similar to the development population, we used the process described in Figure 1 to calculate the required sample size. Stata code to implement the approach for this example is provided in the Supplementary Material (and R code available at https://www.github.com/gscollins1973/). The process and findings are now summarized.

4.1 | Step 1: Set-up process

- **Specify the time point of interest for checking calibration performance**: The key time point of interest for validation was 3 years.
- **Specify the model's anticipated linear predictor ($\mathbf{LP}_i$) distribution in the validation population**: The distribution of LP_i values was not described in the model development article, however upon request the authors provided a histogram. This is displayed in Figure 2, and it could be closely approximated by assuming LP_i follows a skew normal distribution with a mean of 4.60, a variance of 0.65, a skewness parameter of −0.5, and a kurtosis parameter of 5.
- **Specify the assumed $\boldsymbol{F(t)}$ (or $\boldsymbol{S(t)}$) in the validation population at the time point of interest**: The model development article reports a Kaplan-Meier survival curve in the development sample, and by 3 years about 17% of the population had a VTE recurrence, and thus $F(3) = 0.17$ and $S(3) = 0.83$. Hence, we assumed that this would be the same in the validation population.
- **Specify the assumed distribution of survival times in the validation population, conditional on the effect of $\mathbf{LP}_i$**: We assumed the existing model is well calibrated, and identified (in a large simulated dataset) that an exponential distribution with baseline rate parameter of about 0.00050 ensures $F(3)$ is 0.17 when the log hazard ratio for the effect of LP_i is 1 (ie, calibration slope is 1). Hence, survival times were drawn from this distribution.
- **Specify the assumed distribution of censoring times in the validation population, and maximum follow-up time**: The model developers told us that the censoring rate was high, with about 72% of individuals censored before 3 years. To mirror this—and assuming a constant rate of censoring—we drew censoring times from an exponential distribution with rate parameter of 0.426 (as this gives a probability of being censored by 3 years of about 0.72), and then classed the individual as censored if their generated censoring time was less than their survival time. We also assumed the maximum follow-up time would be 3 years and so all individuals with a survival and censoring times generated to be after 3 years were censored immediately after 3 years.

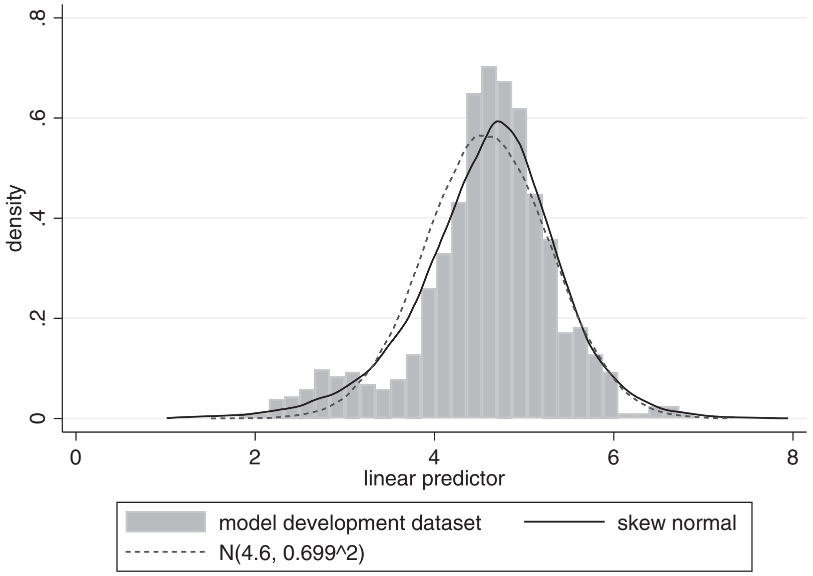

FIGURE 2 Comparison of the linear predictor distribution from the model development dataset (as supplied by the model developers) and an assumed skew normal distribution (mean 4.60, variance 0.65, skewness −0.5, kurtosis 5) and an assumed N(4.60, 0.699^2) distribution

- **Specify the target value for the SE of the calibration slope**: We initially chose a SE of 0.051, to target a confidence interval width of 0.2 (ie, 0.9-1.1 assuming the slope is 1) for the calibration slope as measured on the complementary log-log scale of the calibration model of Equation (3). We emphasize this SE is a starting point, and may need to be refined subsequent to observing calibration plots and curves in simulated data, as described below.

4.2 | Steps 2 to 9: Simulation-based approach to identify the required sample size

Under the assumptions from step 1, the iterative process from steps 2 to 9 identified that about 14250 individuals (with about 1430 observed events by 3 years) are required to estimate the calibration slope with a target SE of 0.051. To better appreciate this level of precision, we plotted the calibration curve and its 95% confidence interval for the simulated datasets of this sample size. Figure 3A provides a representative example, which shows precise calibration across the whole spectrum of predicted risks, especially in the range of risks between 0 and 0.5. This is also reflected by the spread of calibration curves from all the simulated datasets with a sample size of 14250, as shown in Figure 4A for a random sample of 500 of the curves. There is little variation in the range of 0 to 0.5, but it gradually increases as the predicted risk moves closer to 1, because the number of participants in the upper range is far fewer than at the lower range (as the overall risk is 0.17). Therefore, the sample size of 14250 participants is the minimum required to target a precise calibration assessment across the entire range of predicted risks. When reducing the target SE to 0.012 (ie, a target confidence interval width of 0.4 for the calibration slope), there is more uncertainty in estimated curves (Figure 3B) and more variation in observed curves across simulations (Figure 4B).

The sample size required to achieve precise calibration slope estimates is usually larger than for other performance measures, such as calibration-in-the-large and discrimination. This should still be checked. For example, in steps 7 to 9 we also calculated the mean confidence interval width for the *C-index*, which was 0.028 when the sample size was 14250 individuals, and so very precise around the assumed *C-index* of 0.69.

5 | CONSIDERATION OF CLINICAL UTILITY AND DECISION-MAKING

The sample size approach described in Sections 3 and 4 aims for a precise estimate of the calibration curve across the entire range of predictions, and the target confidence interval width of 0.9 to 1.1 for the calibration slope is a helpful way to quantify this. However, sample sizes will often need to be very large to achieve this, as seen in the VTE example

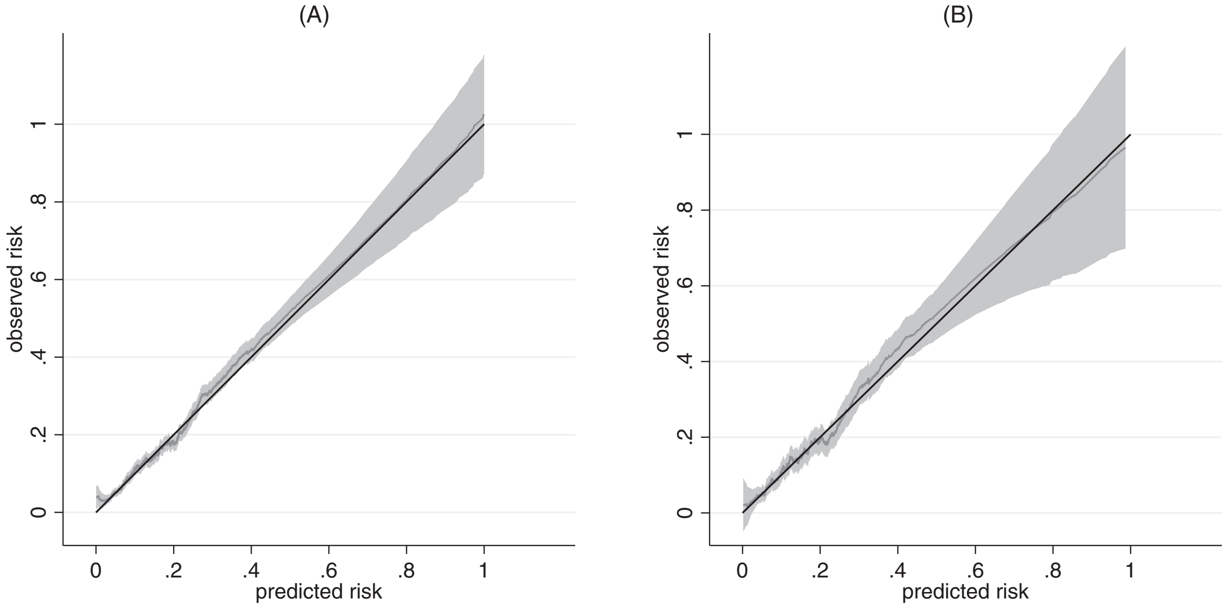

FIGURE 3 Example calibration curves when validating the VTE prediction model in a simulated sample of (A) 14250 individuals, which targets a SE of 0.051 for the calibration slope (confidence interval width of 0.2), and (B) 3600 individuals, which targets a SE of 0.102 for the calibration slope (confidence interval width of 0.4). Black line indicates perfect calibration; grey line is the estimated calibration curve; shaded region is 95% confidence interval

where about 14250 individuals and 1430 events are required. Such numbers may not be possible in practice (eg, from a new cohort study), unless an existing routinely collected (electronic) database with all the relevant information available could be identified. However, depending on the intended role of the model, ensuring precise calibration estimates across the entire range (0-1) may be too stringent. In particular, we may opt for greater precision in regions of predicted risk relevant to clinical decision making and accept lower precision in other regions where miscalibration is less important. For example, for VTE recurrence, predicted risks between about 0.03 and 0.20 have been suggested to warrant clinical action, such as remaining on anticoagulation therapy. Hence, slight to moderate miscalibration in ranges of highest risk (0.5-1) is potentially acceptable in this context, as it is unlikely to change decisions that are made based on thresholds defined by low risks (0-0.2, say).

In such situations, lower sample sizes might be deemed acceptable as long as precision remained quite high in the regions linked to clinical decision making. For example, the target confidence interval width for the calibration slope might be lowered to 0.3 or 0.4, instead of a width of 0.2 as used previously. We investigated this in the VTE example by lowering the required target SE for the calibration slope to 0.102, to aim for a confidence interval width of 0.4. Repeating steps 1 to 9, we found that 3600 individuals (about 365 events) are required to meet this target, and so a substantial reduction of more than 10000 individuals and 1000 events compared to when aiming for a confidence interval width of 0.2.

Examining the calibration plot from a random selection of the simulated datasets of size 3600, the confidence interval for the calibration curve is still reasonably narrow in the range where decision making thresholds may lie (0-0.2). An example is given in Figure 3B. However, the uncertainty is noticeably bigger than before, and the variation in observed calibration curves is also much larger, with some curves drifting down toward observed risks of 0.5 when predicted risks are close to 1. Hence, it would be inadvisable to lower the sample size any further, and so we deemed the 3600 individuals as the *minimum* sample size required. The word "minimum" is important—there is a strong argument that risk thresholds vary across individuals and settings, and so focusing only on a narrow range of thresholds (eg, 0 to 0.2) is subjective and an incomplete picture, compared to ensuring precise calibration across the entire spectrum of risks from 0 to 1.

Alongside consideration of calibration, it is also helpful to check whether the net benefit of the model is precisely estimated in the region of clinically relevant risk thresholds. The net benefit measures the overall consequences of using a prediction model for clinical decisions, which requires the weighing of the benefits (eg, improved patient outcomes)

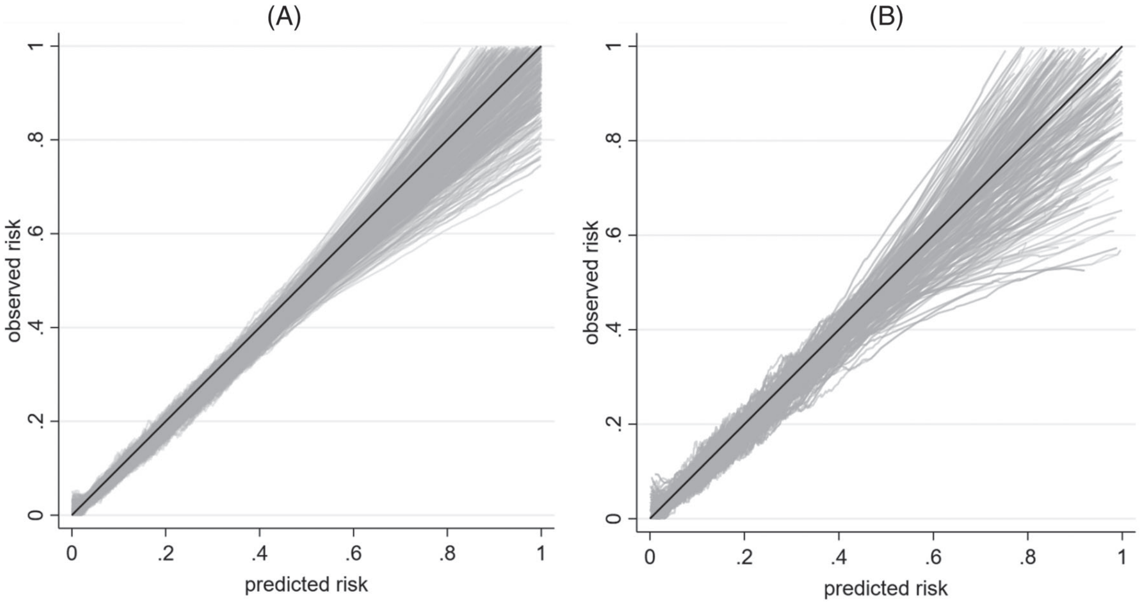

FIGURE 4 Spread of 500 estimated calibration curves when validating the VTE prediction model in a simulated sample of (A) 14250 individuals, which targets a SE of 0.051 for the calibration slope (confidence interval width of 0.2), and (B) 3600 individuals, which targets a SE of 0.102 for the calibration slope (confidence interval width of 0.4). Black line indicates perfect calibration; grey lines show the estimated calibration curves

against the harms (eg, worse patient outcomes, additional costs).[6,34,35] It requires the researchers to choose a risk threshold, such that if an individual's predicted risk of the outcome event is above that risk there will be a clinical action (eg, further treatment, referral to specialist, etc.). Often a range of thresholds are important to consider, such as 0.05 to 0.20 in the VTE application. Net benefit examination can be incorporated within steps 7 to 9 of the simulation framework. For example, in the VTE example using a risk threshold of 0.05, Figure 5A shows a histogram of the observed net benefit values from 200 simulated datasets of size 3600, and the empirical SE of these was 0.0084. This is small relative to the anticipated net benefit of about 0.125 at this threshold, and would allow clear evidence demonstrating a difference from "treat none" (net benefit of zero) strategy, and allow a detailed comparison to a "treat all" strategy. Net benefit would also be precisely estimated at other thresholds up to 0.2, and so the 3600 sample size is adequate in this range to examine clinical utility measured by net benefit. However, if higher risk thresholds were of interest, then the sample size may be insufficient; for instance, the distribution of net benefit estimates at a risk threshold of 0.5 overlaps zero (Figure 5B). This emphasizes why, ideally, we want to obtain precise calibration estimates across the entire risk spectrum, and that focusing on a narrower range must be appropriately justified by clinical experts and patient groups.

6 | EXTENSIONS

6.1 | What if the linear predictor distribution was not available?

In our applied example we used a skew normal distribution to approximate the observed linear predictor distribution in the histogram (Figure 2). But what if the histogram depicting the linear predictor distribution had not been provided? In that situation, we could have used the model's Royston D statistic, but this was also not reported. However, a Harrell's *C-index* of 0.69 was reported, and so we used Equation (5) to approximate the corresponding D statistic:

$$\mathrm{D} = 5.50(C - 0.5) + 10.26(C - 0.5)^3 = 5.50(0.69 - 0.5) + 10.26(0.69 - 0.5)^3 = 1.115$$

Then, an estimate of the SD of the linear predictor distribution (assuming it is normally distributed) is $\hat{\sigma} = D/(\sqrt{8/\pi}) = 1.115/(\sqrt{8/\pi}) = 0.699$. Assuming (without loss of generality) that the linear predictor is centered at its mean, we can then assume $LP_i \sim N\left(0, 0.699^2\right)$, and subsequently identify the corresponding baseline hazard that ensures

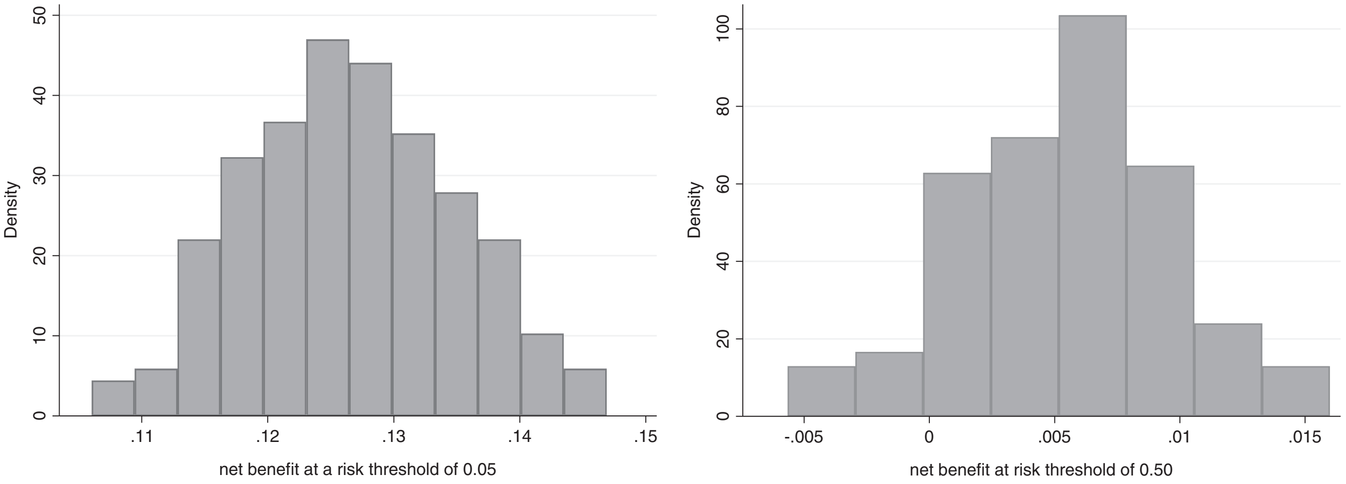

FIGURE 5 Histogram of the observed net benefit at (A) a risk threshold of 0.05 (left figure) and (B) a risk threshold of 0.50 (right figure), from 200 simulated datasets of size 3600

that that $S(3)$ is correct. Recall it was reported that $F(3) = 0.17$. By simulating a large dataset, and through trial and error, $F(3) = 0.17$ corresponds to drawing survival times from a conditional exponential distribution with a baseline rate of about 0.052 and a linear predictor effect of 1.

Therefore, we repeated the simulation process under these assumptions, and targeted a SE of 0.102 (confidence interval width of 0.4), in keeping with Section 5. This identified that a sample size of about 3400 individuals (366 events) is required to achieve this. This is very similar to the sample size of about 3600 individuals (365 events) identified when using the skew normal distribution of Figure 2 for the linear predictor. This is not surprising, because the normal distribution and the skew normal distribution are very similar (as shown in Figure 2 when centering the mean of LP_i at 4.6 for both distributions). Nevertheless, this extension serves to demonstrate that in situations where assuming a normal distribution for the linear predictor is reasonable, and a histogram or summary statistics for the linear predictor are unavailable, then the proposed sample size approach is still feasible even when only the model's *C-index* or Royston's D statistic can be obtained from the model development study.

6.2 | Impact of link function

The calibration model of Equation (3) uses a complementary log-log link function, and hence the target precision of the calibration slope is defined on this scale. The question might be asked whether using a different link function could lead to more precise calibration plots and curves. However, the calibration plots and curves shown in Figures 3 and 4 are produced by a non-parametric running line on the actual risk scale (best fit of "observed" risks based on pseudo-observations vs predicted risks from the existing model), and so are not affected by the link function used to estimate the calibration slope. This is why we emphasized the importance of checking calibration plots and calibration curves for the simulated datasets, to ensure they look acceptable once the sample size for the target SE is identified. In other words, the target SE for the calibration slope is just a starting point, and so the link function is somewhat arbitrary. Furthermore, in our VTE example the mean SE was similar regardless of whether a logit or complementary log-log link was used. For simulated datasets of size 14250, the mean SE of the calibration slope was about 0.051 when using complementary log-log link, and about 0.054 when using a logit link. For simulated datasets of size 3600, the mean SE of the calibration slope was about 0.102 when using a complementary log-log link, and about 0.109 when using a logit link.

6.3 | Impact of changing assumptions

In our example, the simulation-based approach assumed that the survival distribution was exponential, and thus the event rate was constant over time. This is a pragmatic choice of distribution without other information, but more complex survival distributions should be considered if evidence supports them. For example, it might be known that events are more likely to occur in the first year, and then afterwards become rare. We examined this in the VTE example by replacing an exponential distribution with a Weibull distribution (with parameters $\lambda = 0.00119$, $\gamma = 0.2$), with a high event rate in the first year and lower rate thereafter, but still ensuring $F(3) = 0.17$ conditional on the effect of the linear predictor effect being 1. Repeating the simulation exercise whilst still assuming a sample size of 3600 participants, we found that the Weibull distribution leads to a mean SE of about 0.076, substantially lower than the mean SE of 0.102 when assuming an exponential distribution. Hence a lower sample size would be needed if this Weibull distribution was more appropriate. Conversely, assuming events happen mostly in year 3 (Weibull distribution with parameters $\lambda = 0.0000062$ and $\gamma = 5$), a much larger mean SE of about 0.13 is obtained. Hence, the mean SE (and thus the required sample size) is sensitive to the chosen survival distribution form.

Another challenge for the user is to accurately specify the potential censoring distribution and rate in the validation dataset. In the VTE example, we assumed 72% of individuals would be censored before 3 years, based on the censoring reported for the model development study. This is high, and the validation study may hope to retain people for longer through better follow-up procedures. If the censoring proportion by 3 years is assumed to be 50%, then a sample size of about 2400 individuals (and about 300 events) is required to obtain a mean SE for the calibration slope of 0.102. This is over 1000 fewer individuals and 60 fewer events than when assuming the censoring proportion would be 72%. Furthermore, the assumed censoring distribution is also important, because it dictates how many person-years will be contributed by those individuals censored before the time point of interest. For example, if all censoring occurs in the third year of follow-up, this is very different to it all occurring in the first six months or at a constant rate over all three years. Hence, if more information is known about the expected censoring (eg, due to administrative reasons), then this should help inform the censoring distribution assumed for simulation.

In summary, the assumed survival and censoring distributions are influential; our basic choices (eg, exponential) will often be the most pragmatic choice, but if further information about the distributions can be obtained it will be more accurate. For example, if a Kaplan-Meier curve (with censoring points) is provided in a publication about the target population, then digital software could be used to read in the corresponding points and help reconstruct survival and censoring times to inform the underlying distributions.[36]

6.4 | Comparison to logistic regression

Pavlou et al suggest that sample size for validation of time-to-event prediction models can follow the sample size required for logistic prediction models "when the proportion of censored observations is high".[10] To examine this in the VTE example, we applied the closed-form sample size approach of Riley et al for validation of a logistic regression prediction model (with 3-year predicted risks generated from VTE model),[5] and this calculates that 1203 individuals (about 203 events) are required for a target SE of 0.102 for the calibration slope. This is substantially smaller than the 3600 individuals (360 events) required when accounting for the assumed high censoring rate of 70%. Hence, we do not recommend using the logistic regression approach unless the censoring rate is very low, as mentioned in Section 3. If censoring is very low, then our simulation-based approach for survival data will give very similar answers to that based on logistic regression; however, the sample size calculation for the logistic approach is much quicker, as it uses closed-form solutions and so does not require simulation.

6.5 | Multiple time points

If more than one time point is of interest (eg, 1, 5, and 10 years), then a separate validation is required for each time point, and so a separate sample size calculation is needed for each. The largest sample size is then the one required for the validation study. As events occur less by earlier time points, it is likely that larger sample sizes are needed for the earlier time points. Multiple time points will also require a more complex survival distribution to be assumed than exponential, to ensure that the population risk is as anticipated at all the time points of interest.

6.6 | Extension to settings with competing risks

In the case that there are other events that may preclude the event of interest, competing risks must be considered.[23] However, there are various complications when adapting the process in Figure 1, some of which are not immediately obvious. First, it should be clear what the chosen estimands are, as the interest may include the risk of one event, or the total risk partitioned into each event-specific risks. The cause-specific cumulative incidence function for the main cause of interest might be the chosen estimand (ie, the risk of a specific event). However, this is no longer likely to span the entire risk range from 0 to 1 since the competing events will affect the number of individuals censored, thus reducing the total number of events for the event of interest. On the other hand, if interested in partitioning total risk to obtain each event-specific cumulative incidence function (event-specific risk by a particular time-point) it is important to consider that sample sizes will vary for each event depending on how often they occur. Therefore, a decision will need to be made on which event to calculate the sample size on, and it may be most natural to consider the event that occurs least frequently.

A further complication on adapting the process in Figure 1 to competing risks is that a choice must be made on how to model predictor effects. One can either model covariate effects directly on the event-specific risk via the subdistribution hazard function, or model effects on each event-specific hazard rate then transform to obtain the event-specific risk. The process for a standard survival analysis can easily be extended when the estimand of interest is in one event-specific risk that is modeled on the subdistribution hazards scale, such as the Fine and Gray model. On the other hand, interest may be in all event-types, which may motivate the need for cause-specific models, which necessitates that some steps may require more thought. For instance, the data generating mechanism in the simulation for competing risks data becomes more complicated as the LP_i for each model may be correlated between different competing events. To circumvent this, individual participant data or details of correlation between each of the LP_i for the different events may be required to simulate the competing risks data more accurately. Assuming an exponential distribution for each cause is also likely to be too simplistic.

Further details on various ways to simulate competing risks data are outlined in Beyersmann et al,[37] and the use of pseudo-observations and calibration curves in competing risk settings is given by Gerds et al[23] The calculation of pseudo-observations must be made using the Aalen-Johansen estimate of the event-specific cumulative incidence function such that,

$$\widetilde{F}_i(t) = nF_{AJ}(t) - \left[(n-1)F_{AJ(-i)}(t)\right],$$

where, $F_{AJ}(t)$, is the Aalen-Johansen estimator.[38]

7 | DISCUSSION

Building on our earlier work for models predicting binary outcomes,[5,9] we have shown how to use a simulation-based approach to identify the minimum sample size required to externally validate a clinical prediction model with a time-to-event outcome. Our approach is focused on precise estimation of the calibration slope and calibration curves, but precision of other measures should also be checked in the simulation process, such as the C-index, D statistic, and net benefit. Obtaining a precise estimate of calibration is important, but what defines "precise" is context specific, and the visual inspection of simulated calibration curves can be a helpful guide, especially when particular ranges of risk are of more interest for clinical decision making.

Our work identifies a sample size that targets a confidence interval of a particular width. This means that, if the data generating assumptions are correct, this width will be achieved on *average* when a dataset of the identified sample size is generated. Of course, in a *single* dataset the confidence interval width may be narrower or wider by chance, even when the assumptions made are correct. A more stringent criterion is to identify a sample size such that the range of standard errors that could be produced are unlikely to be above a particular value. This will require a larger sample size.

The observed confidence interval width will also depend on deviations from the modeling assumptions, for example in terms of the observed calibration slope, survival and censoring distributions, and linear predictor distribution. This issue is the same for any sample size calculation: if the input assumptions are incorrect, then the actual study will deviate from the target precision; for example, if the censoring rate in practice is higher than that assumed, a larger number of events will be required.[26] This could lead to a prospective study recruiting more individuals than necessary, or conversely recruiting fewer than needed. However, this is typically the case in observational research.

Jinks et al propose a closed-form approach based on precise estimation of the D statistic.[26] This is much quicker than our simulation-based approach, but only focuses on discrimination and not calibration. Calibration is a fundamental measure that is often neglected in validation studies,[39,40] and we think it should be the main focus on a sample size calculation for validation studies, as calibration is the aspect that is most affected when evaluated in new data. The precision of the D statistic and other discrimination measures could be summarized as part of steps 7 to 9 of our simulation process, alongside the calibration slope and any other measure of interest. Indeed, the flexibility of the simulation approach is its great advantage, especially as it allows the generation of calibration plots and flexible calibration curves with confidence intervals, to help the user to better visualize the precision on a meaningful scale.

Our simulation procedure assumes that censoring is non-informative, which is a pragmatic decision and eases the computation process. However, in practice censoring may be informative (ie, conditional on covariates and thus underlying risk), and so once data are actually collected for the validation, this should be considered. In particular, pseudo-observations assume non-informative censoring, and so to help mitigate against this, we suggest deriving them separately within each of, say, 10 or 20 groups defined by tenths or twentieths of predicted risk. Rather than using pseudo-observations, Austin et al suggest using flexible adaptive hazard regression or a Cox model using restricted cubic splines, which also allows observed risks to be produced (and thus calibration curves) assuming uninformative censoring conditional on the predicted risk and proportional hazards.[22]

In conclusion, we have proposed a simulation-based framework for examining the sample size required to precisely validate a prediction model for a time-to-event outcome. This encourages users to focus on ensuring precise estimates of calibration, ideally across the entire range of predicted risks, but at the very least in regions of key importance for clinical decision-making.

ACKNOWLEDGEMENTS

We thank an anonymous reviewer for their helpful comments and attention to detail that helped us improve our article upon revision. Gary S. Collins was supported by the NIHR Biomedical Research Centre, Oxford, and Cancer Research UK (programme Grant: C49297/A27294). Sarwar I. Mozumder was supported by the National Institute for Health Research (NIHR Advanced Fellowship, Dr. Sarwar I. Mozumder, NIHR300100). Mark J. Rutherford was supported by Cancer Research UK project Grant (C41379/A27583). Sarah Booth was supported by Cancer Research UK project Grant (C14183/A29739).

DATA AVAILABILITY STATEMENT

the work presented involves applying equations using inputted or simulated data, and therefore no actual data is available for sharing. Simulation code is provided in the Supplementary Material.

ORCID

Richard D. Riley https://orcid.org/0000-0001-8699-0735
Gary S. Collins https://orcid.org/0000-0002-2772-2316
Joie Ensor https://orcid.org/0000-0001-7481-0282
Lucinda Archer https://orcid.org/0000-0003-2504-2613
Sarah Booth https://orcid.org/0000-0003-1799-3144
Sarwar I. Mozumder https://orcid.org/0000-0001-9644-7525
Mark J. Rutherford https://orcid.org/0000-0003-1557-6697
Maarten van Smeden https://orcid.org/0000-0002-5529-1541
Paul C. Lambert https://orcid.org/0000-0002-5337-663X
Kym I. E. Snell https://orcid.org/0000-0001-9373-6591

REFERENCES

1. Riley RD, van der Windt D, Croft P, et al., eds. *Prognosis Research in Healthcare: Concepts, Methods and Impact*. Oxford, UK: Oxford University Press; 2019.

2. Harrell FE Jr. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. 2nd ed. New York, NY: Springer; 2015.
3. Steyerberg EW. *Clinical Prediction Models: a Practical Approach to Development, Validation, and Updating*. 2nd ed. New York, NY: Springer; 2019.
4. Archer L, Snell KIE, Ensor J, Hudda MT, Collins GS, Riley RD. Minimum sample size for external validation of a clinical prediction model with a continuous outcome. *Stat Med*. 2021;40(1):133-146.
5. Riley RD, Debray TP, Collins GS, et al. Minimum sample size for external validation of a clinical prediction model with a binary outcome; 2021.
6. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ*. 2016;352:i6.
7. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ*. 2017;357:j2099.
8. Clift AK, Coupland CAC, Keogh RH, et al. Living risk prediction algorithm (QCOVID) for risk of hospital admission and mortality from coronavirus 19 in adults: national derivation and validation cohort study. *BMJ*. 2020;371:m3731.
9. Snell KIE, Archer L, Ensor J, et al. External validation of clinical prediction models: simulation-based sample size calculations were more reliable than rules-of-thumb. *J Clin Epidemiol*. 2021;135:79-89.
10. Pavlou M, Qu C, Omar RZ, et al. Estimation of required sample size for external validation of risk models for binary outcomes. *Stat Methods Med Res*. 2021;9622802211007522:2187-2206.
11. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. *Stat Med*. 2016;35(2):214-226.
12. Van Calster B, Nieboer D, Vergouwe Y, et al. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol*. 2016;74:167-176.
13. Royston P, Altman DG. External validation of a Cox prognostic model: principles and methods. *BMC Med Res Methodol*. 2013;13:33.
14. van Houwelingen HC. Validation, calibration, revision and combination of prognostic survival models. *Stat Med*. 2000;19(24):3401-3415.
15. Parner ET, Andersen PK. Regression analysis of censored data using pseudo-observations. *Stata J*. 2010;10(3):408-422.
16. Andersen PK, Perme MP. Pseudo-observations in survival analysis. *Stat Methods Med Res*. 2010;19(1):71-99.
17. Royston P. Tools for checking calibration of a Cox model in external validation: approach based on individual event probabilities. *Stata J*. 2014;14(4):738-755.
18. Perme MP, Andersen PK. Checking hazard regression models using pseudo-observations. *Stat Med*. 2008;27(25):5309-5328.
19. Graw F, Gerds TA, Schumacher M. On pseudo-values for regression analysis in competing risks models. *Lifetime Data Anal*. 2009;15(2):241-255.
20. Overgaard M, Andersen PK, Parner ET. Regression analysis of censored data using pseudo-observations: an update. *Stata J*. 2015;15(3):809-821.
21. Klein JP, Gerster M, Andersen PK, Tarima S, Perme MP. SAS and R functions to compute pseudo-values for censored data regression. *Comput Methods Programs Biomed*. 2008;89(3):289-300.
22. Austin PC, Harrell FE Jr, van Klaveren D. Graphical calibration curves and the integrated calibration index (ICI) for survival models. *Stat Med*. 2020;39(21):2714-2742.
23. Gerds TA, Andersen PK, Kattan MW. Calibration plots for risk prediction models in the presence of competing risks. *Stat Med*. 2014;33(18):3191-3203.
24. Royston P. Explained variation for survival models. *Stata J*. 2006;6:83-96.
25. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. *Stat Med*. 2004;23(5):723-748.
26. Jinks RC, Royston P, Parmar MK. Discrimination-based sample size calculations for multivariable prognostic models for time-to-event data. *BMC Med Res Methodol*. 2015;15:82.
27. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*. 1996;15(4):361-387.
28. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat Med*. 2011;30(10):1105-1117.
29. Crowther MJ, Lambert PC. Simulating biologically plausible complex survival data. *Stat Med*. 2013;32(23):4118-4134.
30. Brilleman SL, Wolfe R, Moreno-Betancur M, et al. Simulating survival data using the simsurv R Package. *J Stat Softw*. 2021;97(1):27.
31. Ensor J, Riley RD, Jowett S, et al. Prediction of risk of recurrence of venous thromboembolism following treatment for a first unprovoked venous thromboembolism: systematic review, prognostic model and clinical decision rule, and economic evaluation. *Health Technol Assess*. 2016;20(12):1-190.
32. Royston P, Parmar MKB. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Stat Med*. 2002;21:2175-2197.
33. Royston P, Lambert PC. *Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model*. College Station, Texas: CRC Press; 2011.
34. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*. 2006;26(6):565-574.
35. Vickers AJ, Cronin AM, Elkin EB, Gonen M. Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. *BMC Med Inform Decis Mak*. 2008;8:53.
36. Guyot P, Ades AE, Ouwens MJ, et al. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC Med Res Methodol*. 2012;12:9.

37. Beyersmann J, Latouche A, Buchholz A, Schumacher M. Simulating competing risks data in survival analysis. *Stat Med*. 2009;28(6):956-971.
38. Aalen OO, Johansen S. An empirical transition matrix for non-homogeneous Markov chains based on censored observations. *Scand J Stat*. 1978;5(3):141-150.
39. Van Calster B, McLernon DJ, van Smeden M, et al. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17(1):230.
40. Collins GS, de Groot JA, Dutton S, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol*. 2014;14:40.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Riley RD, Collins GS, Ensor J, et al. Minimum sample size calculations for external validation of a clinical prediction model with a time-to-event outcome. *Statistics in Medicine*. 2022;41(7):1280-1295. doi: 10.1002/sim.9275