Received: 10 September 2020 | Revised: 1 February 2021 | Accepted: 22 March 2021

DOI: 10.1002/sim.9025

RESEARCH ARTICLE

Statistics in Medicine WILEY

Minimum sample size for external validation of a clinical prediction model with a binary outcome

Richard D. Riley[1] | **Thomas P. A. Debray**[2] | **Gary S. Collins**[3,4] | **Lucinda Archer**[1] | **Joie Ensor**[1] | **Maarten van Smeden**[2] | **Kym I. E. Snell**[1]

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

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

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

[4]NIHR Oxford Biomedical Research Centre, John Radcliffe Hospital, Oxford, UK

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

Funding information
Cancer Research UK, Grant/Award Number: C49297/A27294; European Union's Horizon 2020 Research and Innovation Programme, Grant/Award Number: ReCoDID Grant Agreement no. 825746; National Institute for Health Research School for Primary Care Research (NIHR SPCR); Netherlands Organisation for Health Research and Development, Grant/Award Number: grant 91617050; NIHR Biomedical Research Centre, Royal Marsden NHS Foundation Trust/Institute of Cancer Research, Oxford; NIHR Biomedical Research Centre, Oxford

In prediction model research, external validation is needed to examine an existing model's performance using data independent to that for model development. Current external validation studies often suffer from small sample sizes and consequently imprecise predictive performance estimates. To address this, we propose how to determine the minimum sample size needed for a new external validation study of a prediction model for a binary outcome. Our calculations aim to precisely estimate calibration (Observed/Expected and calibration slope), discrimination (C-statistic), and clinical utility (net benefit). For each measure, we propose closed-form and iterative solutions for calculating the minimum sample size required. These require specifying: (i) target SEs (confidence interval widths) for each estimate of interest, (ii) the anticipated outcome event proportion in the validation population, (iii) the prediction model's anticipated (mis)calibration and variance of linear predictor values in the validation population, and (iv) potential risk thresholds for clinical decision-making. The calculations can also be used to inform whether the sample size of an existing (already collected) dataset is adequate for external validation. We illustrate our proposal for external validation of a prediction model for mechanical heart valve failure with an expected outcome event proportion of 0.018. Calculations suggest at least 9835 participants (177 events) are required to precisely estimate the calibration and discrimination measures, with this number driven by the calibration slope criterion, which we anticipate will often be the case. Also, 6443 participants (116 events) are required to precisely estimate net benefit at a risk threshold of 8%. Software code is provided.

KEYWORDS
binary outcomes, calibration, discrimination, external validation, minimum sample size, multivariable prediction model, net benefit

1 | INTRODUCTION

Each year in the medical literature, hundreds of prediction models are developed to predict health outcomes in individuals.[1,2] During the COVID-19 pandemic, for example, a systematic review identified 145 prediction models aiming to inform the diagnosis or prognosis of individuals with (suspected) COVID-19 within only a few months since the start of the pandemic.[3] Many proposed models, including those for COVID-19, are developed using poor methods and small low-quality datasets, and thus are likely to perform poorly in practice.[1,4,5] Therefore, before consideration for use in clinical practice, it is important to evaluate a model's performance in separate data independent to that used for model development. This process is known as *external validation* and helps decide whether an existing model is fit for purpose in the target population and settings of interest.[6,7]

External validation studies should aim to precisely estimate key measures of model performance, in order to draw strong conclusions about whether the model is fit for purpose. In particular, calibration (agreement between observed and predicted outcome risks), discrimination (separation of predicted risks between those with and without events), and clinical utility (eg, net benefit of a model when used to guide clinical decision based on thresholds of predicted risk) are all relevant types of performance measures to investigate.[8-10]

Simulation and resampling studies conducted by Vergouwe et al,[11] Collins et al,[12] and Van Calster et al[13] suggested external validation studies should have at least 100 events and 100 nonevents to ensure accurate and precise estimates of performance measures, and even larger sample sizes (a minimum of 200 events and 200 nonevents) to derive flexible calibration curves.[12,13] However, existing validation studies often fall short of meeting such requirements, leading to reported estimates of prediction model performance that are highly inaccurate and often misleading.

Rules-of-thumb for sample size are themselves problematic, as they are not specific to the model or validation setting. Indeed, Snell et al showed that the rule-of-thumb of having at least 100 events and 100 nonevents does not always produce precise estimates of a model's predictive performance measures.[14] This is because the precision of the performance estimates also depends on factors other than the number of events and nonevents, including the variance of the model's linear predictor values in the validation population, the expected performance of the model upon validation, and the anticipated magnitude of any miscalibration. This has also been demonstrated for validation of prediction models with a continuous outcome.[15] To address this for binary outcomes, Snell et al proposed using a simulation-based approach to identify the sample size required. In brief, they suggested simulating an external validation dataset of a particular sample size under assumed conditions (eg, anticipated outcome event proportion, model's anticipated linear predictor distribution, whether or not the model is well calibrated, etc),[14] and using the data to calculate the precision of performance estimates for calibration, discrimination and clinical utility. This is repeated many (eg, 1000) times, so that the average precision of performance estimates within datasets of a particular sample size can be identified. This whole process is then repeated for gradually increasing sample sizes, until a minimum sample size is identified that ensures the average precision (eg, across 1000 simulated datasets of that size) is deemed adequate for each performance measure.

Simulation based calculations offer a high degree of flexibility but may take considerable time and are difficult to apply for those less familiar to setting up and running a simulation. To address this, in this article we derive simpler iterative or closed-form sample size calculations for identifying the *minimum* sample size required for external validation. We focus on models for binary outcomes, which are by far the most common in the medical literature (eg, as derived using logistic regression). We considered continuous outcomes in a previous paper.[15] Our goal is to ensure external validation studies are large enough to precisely estimate a model's calibration, discrimination, and clinical utility.

The paper outline is as follows. In Section 2 we introduce key performance measures for calibration, discrimination and clinical utility of a binary outcome prediction model. Then, for external validation studies that aim to precisely estimate these performance measures, Section 3 derives formulae for calculating the minimum sample size required. Section 4 provides an illustrative example, and Section 5 concludes with discussion.

2 | KEY PERFORMANCE MEASURES UPON EXTERNAL VALIDATION OF A PREDICTION MODEL FOR A BINARY OUTCOME

Consider that we wish to externally validate an existing prediction model for a binary outcome ($Y_i = 0$ or 1), and that this model can be used to calculate the probability (p_{PREDi}) of the outcome event (ie, the predicted risk of $Y_i = 1$,) for

an individual participant (i). As the outcome is binary, the existing prediction model equation will usually be in the form of a logistic regression containing an intercept (γ_0), and predictor effects (γ_1, γ_2, γ_3 etc) corresponding to predictor variables (X_{1i}, X_{2i}, X_{3i}, etc). A simple example of an existing prediction model equation with three predictors equation is:

$$\text{logit}(p_{\text{PREDi}}) = \text{LP}_i = \gamma_0 + \gamma_1 X_{1i} + \gamma_2 X_{2i} + \gamma_3 X_{3i,} \quad (1)$$

The right-hand side of the model equation (the *linear predictor*, LP_i) may be far more complex than shown in Equation (1), for instance with more than three predictors and potential interactions and nonlinear terms (eg, defined by splines or polynomials). The model might also relate to a machine learning approach (eg, neural network), and so not have an equation as such, but can still be validated if it can be used to produce predicted risks for each individual in the validation dataset.

For validation, we require a representative dataset containing N participants from the target population of interest, and for each participant ($i = 1$ to N) we know whether they had the outcome event ($Y_i = 1$) or not ($Y_i = 0$), and their values of predictors included in the model. We now describe how to quantify the prediction model's performance in such a validation dataset.

2.1 | Calibration

Calibration examines the agreement between predicted and observed outcome event risks, and should be examined across the whole spectrum of predicted risks.[16] It is recommended that calibration should be visualized graphically and include a smoothed nonlinear curve generated using a loess smoother or splines at the participant-level.[13,17] Alongside calibration plots, the magnitude of (mis)calibration can also be quantified by statistical measures including the calibration slope (ideal value of 1) and the observed/expected ratio (O/E, ideal value of 1) or conversely the E/O ratio. Other useful measures include the Estimated (ECI) or Integrated (ICI) Calibration Index,[18,19] which respectively measure an average of the squared or absolute differences between predicted risks from the model and observed risks from a calibration curve. Our sample size calculations for calibration focus on O/E and the calibration slope, so we now discuss these in more detail.

2.1.1 | O/E

Calibration-in-the-large measures the overall calibration between observed and predicted risks in the validation dataset.[19] A fundamental measure of calibration-in-the-large is O/E, which is the ratio of the total number of observed outcome events, divided by the total number of expected (predicted) outcome events. Equivalently, O/E is the observed risk of having the outcome event in the entire validation dataset (ie, O/N) divided by the average risk predicted by the model $\left(\frac{\sum_{i=1}^{N} p_{\text{PREDi}}}{N}\right)$. Thus, the ideal value of O/E is 1. Values less than 1 indicate the model is over-predicting the total number of outcome events in the dataset, while values above 1 indicate the model is under-predicting the total number of outcome events in the dataset.

Calibration-in-the-large can also be measured by the estimated intercept in a calibration model that constrains the calibration slope to be 1 (see Equation (2)).[20] However, we focus on O/E in the remainder of the article as it is easier to interpret.

2.1.2 | Calibration slope

The calibration slope is a measure of the association between observed and predicted risk of the outcome event across the whole range of predicted risks.[20,21] It quantifies how a unit increase in logit predicted risk corresponds with a unit increase in logit observed risk. When a model is developed using traditional estimation techniques (eg, unpenalized maximum likelihood estimation), the observed (apparent) calibration slope will always be 1 in the original development data. However, upon validation in new data, a model's calibration slope will typically deviate from 1 due to lack of generalizability of model predictions.

For binary outcomes, the calibration slope can be estimated in the validation dataset using the following logistic regression,

$$\text{logit}(p_i) = \alpha + \beta \text{LP}_i, \quad (2)$$

where p_i is the underlying risk for individual i in the validation dataset, β is the calibration slope, and LP_i is the linear predictor value derived from the prediction model for individual i (ie, $\text{LP}_i = \text{logit}(p_{\text{PRED}i})$).

A $\beta < 1$ indicates that in some probability ranges (which can directly be viewed from the calibration plot) the model predictions are too extreme (ie, predictions close to 1 are too high, and predictions close to 0 are too low) and a $\beta > 1$ indicates model predictions are too narrow (ie, predictions close to 1 are too low, and predictions close to 0 are too high). A $\beta < 1$ is often observed in external validation studies, consistent with a lack of adjustment for over-fitting (optimism) of the model when it was developed.[22-24] In this article, we refer to a prediction model as "well calibrated" if $\beta \approx 1$ and $O/E \approx 1$ upon external validation (and thus also $\alpha \approx 0$ conditional on $\beta \approx 1$). Even so, this may only be considered as a "weak" level of acceptable calibration,[13] and so forms the minimum that we should aim to detect, as explained in the next subsection.

2.1.3 | Why O/E and calibration slope are the bare minimum to consider for calibration

Sample size calculations must consider both O/E and calibration slope. O/E may be 1 even when there is still substantial miscalibration; that is, on average predictions may appear well calibrated, but there can be under-prediction in some predicted probability ranges which cancels out over-prediction (or vice versa) in other ranges. This stresses the need for also providing calibration slope. Conversely, systematic over- or under-prediction is still possible even when the calibration slope is 1, and thus it is important to consider both O/E and calibration slope to quantify calibration performance.

Reporting O/E and calibration slope corresponds to assessing "weak" level calibration,[13] as they provide an important but incomplete assessment of a model's calibration upon external validation. In particular, they do not fully reveal the calibration across the spectrum of predicted risks.[13] They should be presented alongside calibration plots with loess smoothed curves, which will visually reveal miscalibration in particular regions of predicted values even when O/E and calibration slope are both 1. However, precisely estimating calibration curves and (mis)calibration in particular ranges (eg, in those with predicted risks between 0 and 0.2) requires larger sample sizes than needed to quantify O/E and the calibration slope in the entire dataset. Hence, by focusing only on O/E and calibration slope, our sample size guidance for calibration in Section 3 forms the minimum required to evaluate calibration in external validation studies.

2.2 | Discrimination

Discrimination refers to how well a model's predictions separate between individuals who do and do not have the outcome event of interest. The most popular and widely used statistic for quantifying model discrimination is the Concordance (C) statistic,[21,25] where a value of 1 indicates the model has perfect discrimination, while a value of 0.5 indicates the model discriminates no better than chance. For binary outcomes, it is equivalent to the area under the receiver operating characteristic curve. It gives the probability that for any randomly selected pair of individuals, one with and one without the outcome event, the model assigns a higher probability to the individual with the outcome event.

2.3 | Clinical utility

When used to direct decision-making, a prediction model should also be evaluated for its overall (net) benefit on participant and health care outcomes; also known as its clinical utility.[26-28] For example, for binary outcomes, if the model gives a predicted risk above a certain threshold value, then the patient and healthcare professional may decide on some clinical action (eg, above current clinical care), such as administering a particular treatment, monitoring strategy, or life-style change.

The overall consequences of using a prediction model for clinical decisions can be measured using net benefit, which requires only the weighing of the benefits (eg, improved patient outcomes) against the harms (eg, worse patient outcomes,

additional costs).[29,30] It requires the researchers to choose a probability threshold p_t, such that if an individual's predicted risk of the outcome event is $\geq p_t$ there will be a clinical action (eg, onset of treatment, referral to specialist, etc). Based on the chosen probability threshold, the net benefit is the difference between the number of true-positive (TP) results and the number of false-positive (FP) results, relative to the total sample size (N), and weighted by a factor $\frac{p_t}{1-p_t}$. The weighting factor is essentially the odds of the outcome event at the chosen probability threshold value (the probability of the outcome event divided by the probability of not having the outcome event), which can equivalently be considered to represent an acceptable harm to benefit ratio. In other words, the chosen p_t should reflect where the expected benefit of clinical action is equal to the expected benefit of avoiding clinical action.[30]

Net benefit (NB_{p_t}) at probability threshold p_t can be calculated as,

$$\begin{aligned}\text{NB}_{p_t} &= \frac{\text{TP}}{N} - \left(\frac{\text{FP}}{N} \times \frac{p_t}{1-p_t}\right) = \frac{\text{TP} - \left(\text{FP} \times \frac{p_t}{1-p_t}\right)}{N} \\ &= (\text{sensitivity} \times \phi) - \left((1 - \text{specificity}) \times (1-\phi) \times \frac{p_t}{1-p_t}\right), \end{aligned} \tag{3}$$

where ϕ is the observed outcome event proportion in the entire dataset. Positive values of the net benefit indicate the model has clinical utility, as the benefits outweigh the harms. It is helpful to multiply the net benefit by 1000, so it can be interpreted as the additional number of true cases identified for treatment (or some clinical action) without increasing the number treated unnecessarily per 1000 individuals. The net benefit is zero if the benefit compensates the harm, and negative if harm surpasses benefit.

The maximum possible value of the net benefit is ϕ, therefore it is bounded below 1.[29] The standardized net benefit ($s\text{NB}_{p_t}$) is defined as NB_{p_t}/ϕ, and the standardization ensures that the maximum value is 1 regardless of the validation setting. Often it is helpful to compare a model's (standardized) net benefit to that of a "treat none" strategy, which is by definition zero. Similarly, a comparison to a "treat all" strategy can be made, where $\text{NB}_{p_t}(\text{treat all}) = \phi - \left[(1-\phi) \times \frac{p_t}{1-p_t}\right]$.

3 | DERIVATION OF SAMPLE SIZE FORMULA FOR EXTERNAL VALIDATION STUDIES

We now derive formulae for calculating the sample size required in an external validation study to target precise estimation of O/E, calibration slope, the C-statistic, and net benefit. For each performance statistic, we derive an iterative or closed-form expression for calculating the minimum sample size (N) required to target a SE (and thus confidence interval width) of a particular magnitude. We assume that confidence intervals are to be derived using the standard Wald-based approach; that is, estimate $\pm$ (1.96 $\times$ SE), where SE is the standard error of the estimate of interest.

3.1 | *O/E*

Debray et al considered various strategies for deriving SEs and confidence intervals for the O/E statistic.[31] One approach uses the delta method,[32] and gives:

$$\text{SE}\left(\ln\left(\frac{O}{E}\right)\right) \approx \sqrt{\frac{1-\phi}{O}}. \tag{4}$$

Note that the SE is for $\ln(O/E)$ and so the confidence interval is derived on the $\ln(O/E)$ scale and then back-transformed to the O/E scale. Recognizing that $O = N\phi$, we can rearrange Equation (4) to give

$$N = \frac{(1-\phi)}{\phi\,(\text{SE}(\ln(O/E)))^2} \tag{5}$$

Hence, by specifying the anticipated outcome event proportion (ϕ) in the external validation population, and a target value for $\text{SE}(\ln(O/E))$ that would be acceptable, we can calculate the minimum sample size (N) required to meet this criterion. Stata code to implement this is provided in Appendix S1.

3.1.1 | Illustrative examples

Assume that ϕ is 0.5 and O/E is 1 in the external validation population, and that we aim for a 95% confidence interval width of 0.2 for O/E to precisely demonstrate that the model has a good calibration-in-the-large. This targeted confidence interval on the O/E scale is about 0.905 to 1.105, and corresponds to using a SE on the $\ln(O/E)$ scale of about 0.051. Hence, using Equation (5) the sample size required is:

$$N = \frac{(1-\phi)}{\phi\,(\mathrm{SE}(\ln(O/E)))^2} = \frac{(1-0.5)}{0.5 \times 0.051^2} = 384.5.$$

Thus, 385 participants (and about 193 events) are required to target a 95% confidence interval width of 0.2 for O/E. Assuming O/E is 1 in the external validation population is a sensible starting point, as it is important to know that a true O/E of 1 can be estimated precisely.

Note that even changing to an assumed $O/E < 1$ or > 1, the required sample size will still be the same if $\mathrm{SE}(\ln(O/E))$ is kept fixed. For instance, let us assume we want to detect evidence of miscalibration, taking an assumed value of O/E of 1.105, which recall was the upper end of the targeted confidence interval when previously assuming an O/E of 1. Then 385 participants are still required if $\mathrm{SE}(\ln(O/E))$ is kept at 0.051, and this corresponds to a targeted confidence interval of 1.00 to 1.22, with the lower bound at the value of perfect calibration-in-the-large.

The choice of the confidence interval width (target lower and upper values), and thus the value of $\mathrm{SE}(\ln(O/E))$ to use, is subjective and context specific, as it depends on the anticipated outcome event proportion and the maximum width of the confidence interval for O/E that is deemed acceptable by the researchers planning the external validation study. Confidence interval limits for O/E of 0.9 and 1.1 represent a difference of 10% in the observed and expected outcome event proportions. For example, if the expected outcome event proportion is 0.5, an O/E of 1.1 corresponds to an observed outcome event proportion of $0.5 \times 1.1 = 0.55$, and thus a 0.05 absolute difference in the observed and expected outcome event proportions. Such an absolute difference is perhaps the maximum difference that would be considered acceptable.

However, if the anticipated outcome event proportion is further away from 0.5 and closer to 0 (as is often the case), a confidence interval width for O/E of 0.2 may be too stringent. For example, when the outcome event proportion is 0.1, Equation (5) identifies that 3461 participants (and about 346 events) are required to target a confidence interval width of 0.2 (assuming O/E is 1). This corresponds to a small absolute difference of 0.01 in the observed and expected outcome event proportions. If we relax the target confidence interval width for O/E to 0.7 (as found when using $\mathrm{SE}(\ln(O/E) = 0.175)$, then assuming O/E is 1, this corresponds to an expected confidence interval of about 0.71 to 1.41. The upper bound of 1.41 corresponds to an absolute difference of 0.041 in the observed and expected outcome event proportions, which is less stringent than before, but might still represent a reasonably small difference (though this should be determined in the context of the clinical setting). Now applying Equation (5) identifies that the required sample size is,

$$N = \frac{(1-\phi)}{\phi\,(\mathrm{SE}(\ln(O/E))^2} = \frac{(1-0.1)}{0.1 \times 0.175^2} = 293.9,$$

and thus at least 294 participants (about 29 events) are required to satisfy this criterion.

3.2 | Calibration slope

Borenstein et al and Demidenko considered sample size calculations for a simple logistic regression model with one predictor,[33,34] which is akin to the calibration model of Equation (2). As summarized by Novikov et al,[35] their methods utilize SEs as derived from the inverse of Fisher's information matrix ($\mathbf{I}$) for the maximum likelihood estimates of α and β in Equation (2), where β is the calibration slope. Applying their approach to Equation (2), $\mathbf{I}$ can be expressed as,

$$\mathbf{I} = N \begin{pmatrix} E\left(\frac{e^{\alpha+\beta \mathrm{LP}_i}}{(1+e^{\alpha+\beta \mathrm{LP}_i})^2}\right) & E\left(\frac{\mathrm{LP}_i\, e^{\alpha+\beta \mathrm{LP}_i}}{(1+e^{\alpha+\beta \mathrm{LP}_i})^2}\right) \\ E\left(\frac{\mathrm{LP}_i\, e^{\alpha+\beta \mathrm{LP}_i}}{(1+e^{\alpha+\beta \mathrm{LP}_i})^2}\right) & E\left(\frac{(\mathrm{LP}_i)^2\, e^{\alpha+\beta \mathrm{LP}_i}}{(1+e^{\alpha+\beta \mathrm{LP}_i})^2}\right) \end{pmatrix}, \quad (6)$$

where N is the total sample size and $E(.)$ denotes the expectation of the term in the brackets.

For simplicity, let us abbreviate Equation (6) to,

$$\mathbf{I} = N\begin{pmatrix} I_\alpha & I_{\alpha,\beta} \\ I_{\alpha,\beta} & I_\beta \end{pmatrix},$$

such that the I terms correspond to the entries of the matrix shown in Equation (6).

The variance of the estimated calibration slope (β) is found by taking the [2,2] element of the inverse of $\mathbf{I}$, and so the SE of the estimate of β can be expressed as:

$$\text{SE}(\beta) = \sqrt{\frac{I_\alpha}{N(I_\alpha I_\beta - I_{\alpha,\beta}^2)}}.$$

Rearranging, we can specify a sample size formula for precisely estimating the calibration slope:

$$N = \frac{I_\alpha}{\text{SE}(\beta)^2(I_\alpha I_\beta - I_{\alpha,\beta}^2)}. \tag{7}$$

Hence, to calculate the required sample size to precisely estimate the calibration slope, researchers need to apply Equation (7) after specifying a target value for SE(β), such as 0.051 to correspond to a 95% confidence interval width of about 0.2 (eg, 0.9 to 1.1 if $\widehat{\beta} = 1$), and also the anticipated values of I_α, $I_{\alpha,\beta}$, and I_β for the external validation population. The I_α, $I_{\alpha,\beta}$, and I_β values correspond to expected values (Equation (6)), and depend on the assumed true values of α and β and the assumed distribution of linear predictor values (ie, LP_i values) in the external validation population. The more variability of the linear predictor, the lower the required sample size. We now describe how to implement this.

3.2.1 | Implementing the approach

Box 1 describes a five-part process (parts (A)-(E)) to calculate I_α, $I_{\alpha,\beta}$, and I_β needed for the sample size calculation. Essentially the approach requires a large dataset (eg, containing a million participants with linear predictor and outcome values) to be generated based on the assumed distribution of the prediction model's linear predictor, either in the whole population or in each of the outcome groups (events and nonevents) separately. Then, I_α, $I_{\alpha,\beta}$, and I_β can be calculated by applying Equation (6) after choosing values for α and β, and then Equation (7) is applied to calculate the required sample size (N) for a targeted SE(β). Stata code is provided in the Appendix S1 that automates this process in just a few seconds or minutes, depending on the distribution sampled from. The most challenging part is defining α, β and the linear predictor distribution. Guidance is now provided.

BOX 1 Process to calculate I_α, $I_{\alpha,\beta}$, and I_β in the sample size calculation for calibration slope. Statistical code is provided in the Appendix S1 to implement the method

Part (A): For the calibration model of Equation (2), specify the anticipated values of α and β in the external validation population, where β is the calibration slope.

- Begin by assuming $\alpha = 0$ and $\beta = 1$, such that the model is assumed to be well calibrated (reflecting "weak" level calibration[13]). If assuming $\beta \neq 1$, the value of α must be chosen to ensure the generated dataset has the anticipated outcome event proportion; see Notes below.

Part (B): Specify the anticipated distribution of the model's linear predictor (LP_i) values (in the external validation population); either one distribution for the whole population, or separate distributions for the events and nonevents groups separately. Guidance is given in the main text for identifying the distribution(s), for example based on the model development study. Also see Notes below.

Part (C): Generate a dataset containing a large number of hypothetical participants (eg, 1 000 000), and randomly generate a LP_i value from the distribution (or distributions) specified in part (B) for each participant. When using a distribution for the whole population, ensure that it corresponds to the anticipated outcome event proportion for the validation population (see Notes below). If generating LP_i values from different distributions for the event and nonevent groups, firstly fix which participants in the generated dataset do or do not have an event, so that the dataset matches the anticipated outcome event proportion, and then generate LP_i values for each participant from the corresponding distribution according to whether they have the event or not.

Part (D): For each participant, use their generated value of LP_i from part (C) and the chosen values of α and β from part (A) to calculate values for each of the following:

$$a_i = \frac{e^{\alpha+\beta LP_i}}{(1+e^{\alpha+\beta LP_i})^2} \quad b_i = \frac{LP_i\ e^{\alpha+\beta LP_i}}{(1+e^{\alpha+\beta LP_i})^2} \quad c_i = \frac{(LP_i)^2\ e^{\alpha+\beta LP_i}}{(1+e^{\alpha+\beta LP_i})^2}$$

Part (E): Using all participants in the dataset, calculate the mean value of a_i, the mean value of b_i, and the mean value of c_i, which correspond to I_α, $I_{\alpha,\beta}$, and I_β, respectively. These can then be plugged in to Equation (7) to obtain the total sample size (N) to estimate calibration slope precisely.

NOTES:

- It is important to check that the distribution(s) used correspond to the anticipated outcome event proportion in the overall population for validation. Assuming the model is well calibrated, this can be done by simulating values of the linear predictor (eg, 1 000 000 LP_i values), converting each value back to a predicted probability (p_{PREDi}), randomly generating outcome (0 or 1) values according to Bernoulli(p_{PREDi}) distribution, and finally calculating the proportion of outcome values that equal 1 in the whole dataset. This should be very close to the anticipated outcome event proportion, if the assumed distributions are appropriate.
- If assuming $\beta \neq 1$, the value of α must be chosen to ensure the generated dataset has the anticipated outcome event proportion. This can be checked by generating 1 000 000 revised LP_i values by $\alpha + \beta LP_i$ converting each value back to a revised predicted probability (p_{PREDi}), randomly generating outcome (0 or 1) values according to Bernoulli (p_{PREDi}) distribution, and finally calculating the proportion of outcome values that equal 1. Again, this should closely match the anticipated outcome event proportion.

3.2.2 | Choice of α and β

In terms of choosing the anticipated values of α and β for the calibration model of Equation (2), we recommend assuming true values of $\alpha = 0$ and $\beta = 1$. As the calibration slope (β) is often <1 upon validation (consistent with overfitting during model development, such that the model's predictor effects were too large and not subject to penalization or shrinkage), situations with $\beta < 1$ might also be considered, such as $\beta = 0.8$. In that situation, the value of α should be chosen to ensure that the overall outcome event proportion remains the same as that anticipated in the target population. Nonetheless, based on previous findings,[14,15] the calculation assuming a slope of 1 will usually lead to a larger sample size required. An example is given in Section 4.2.

3.2.3 | Choice of linear predictor distribution

The sample size calculation (ie, Box 1 followed by Equation (7)) utilizes values of the model's linear predictor (LP_i) values. Gauging the anticipated distribution of the model's linear predictor in the external validation population may be nontrivial, and it is important to take care over this choice. If the external validation population is the same or similar to that used for model development, then a sensible starting point is to consider the linear predictor distribution in the development data; for example, the user may identify a distribution that closely matches the distribution presented (eg, in a histogram)

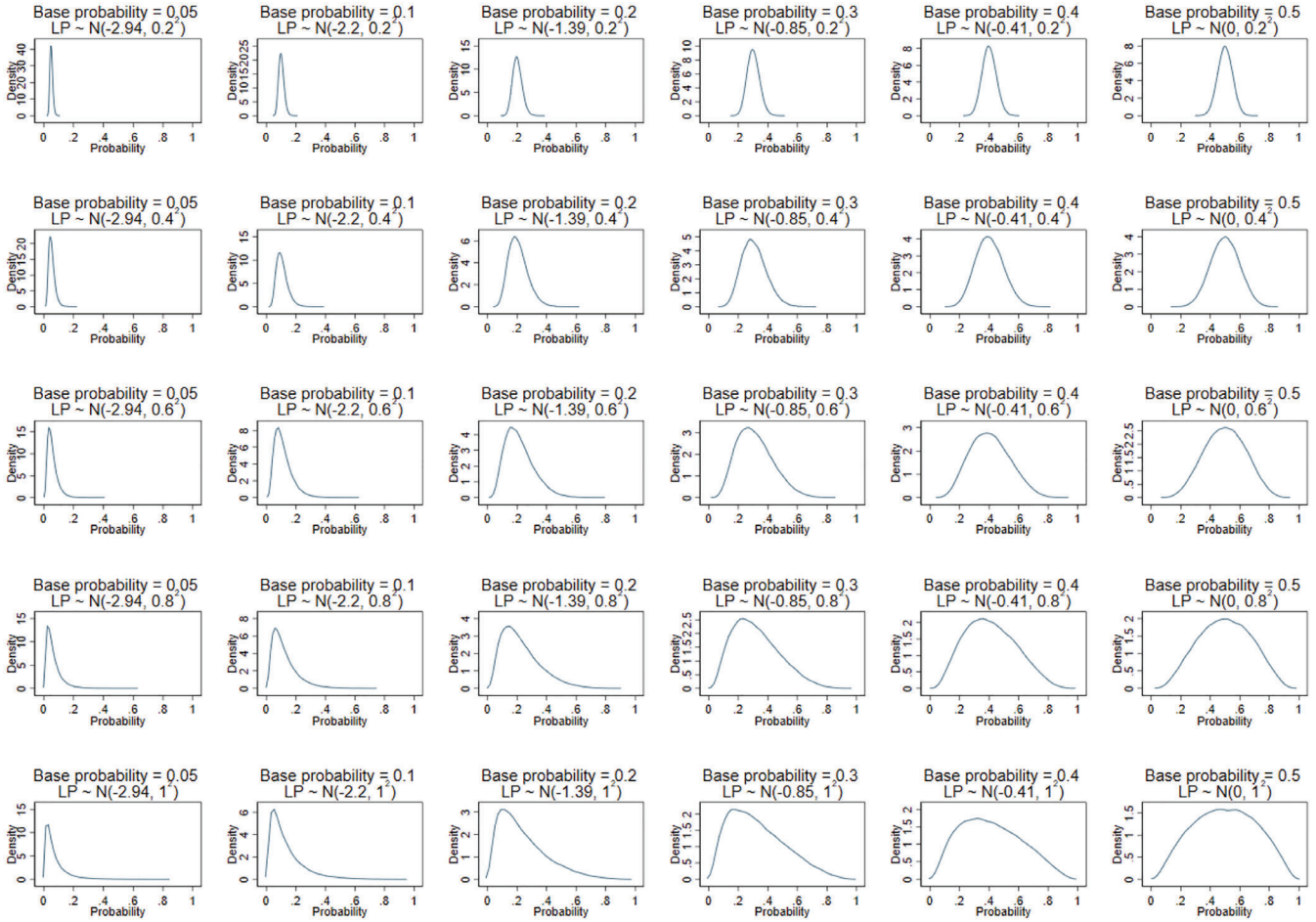

FIGURE 1 Example distributions* of p_{PREDi} values, based on an assumed normal distribution of the linear predictor (LP_i) values of the prediction model in the external validation population, as originally shown by Snell et al.[14] * "Base probability" refers to the probability of the outcome event for an individual with the mean LP value. The outcome event proportion for the distributions is about 0.06 in column 1, about 0.11 in column 2, about 0.22 in column 3, about 0.31 in column 4, about 0.41 in column 5, and 0.50 in column 6 [Colour figure can be viewed at wileyonlinelibrary.com]

in the model development publication or obtained from the model developers directly. An example is given in Section 4, in which a skewed normal distribution is used.

Sometimes a study will present (eg, in a histogram underneath a calibration plot) the distribution of predicted risks (ie, p_{PREDi} values). Then, a suitable distribution on the 0 to 1 scale (eg, a beta distribution) can be used that approximates this distribution, and the logit of sampled values provides the LP_i values to be used in the process described in Box 1. Figure 1 provides various examples of distributions on the risk scale, which correspond to assuming normality of the linear predictor distribution. However, we stress that these are only illustrative examples, and other distributions are possible. Indeed, it is inappropriate to simply assume the linear predictor distribution is normally distributed by default as, for example, it may be skewed, bimodal, or even discrete (eg, if the linear predictor is derived from only binary or categorical covariates).

Sometimes the development study may report or present the linear predictor distribution for each of the outcome event and nonevent groups separately; indeed, these may be very different to one another. In that situation, the researcher should specify separate distributions for each group, which can then be used in the process of Box 1 to generate linear predictor values for event and nonevent individuals in the sample size calculation. If no linear predictor distributions are reported or presented then the reported *C*-statistic can also be used to infer the distributions (Section 3.2.3.1), although this requires strong assumptions. It is important to check that the distribution(s) used correspond to the overall outcome event proportion (ϕ) anticipated for the target population (as described in the Notes within Box 1). If the outcome event proportion is different, then the sample size calculation will be inaccurate (Section 4.6 provides an example of this issue).

To address differences in the event proportion from development and external validation populations, researchers might consider shifting the linear predictor distribution observed in the development study; for example, they could shift the distribution toward lower values (eg, lower outcome event proportion), or toward higher values (eg, higher outcome event proportion).

However, if large differences are anticipated in the linear predictor distribution between development and validation populations (eg, due to expected changes in both outcome event proportion and case-mix), such that the mean, variance, and even shape of the distribution may be affected, then this is very challenging. If such a situation is anticipated before collection of data, or if no information at all is available to inform the linear predictor distribution, then we recommend a pilot study to better gauge the distribution. Such pilot data could still be included later in the final sample used for validation.

If a dataset is already available for validation (with a fixed number of participants), then the distribution of the linear predictor can be observed and therefore used directly to inform whether the sample size is sufficient.

3.2.4 | Special case assuming the calibration slope is 1 and the linear predictor distribution is approximately normally distributed with a common variance in event and nonevent groups

The relationship between the *C*-statistic and the distribution of the linear predictor has been previously described,[36] with solutions described when the linear predictor distributions in the events and nonevents groups are normal with a common variance. Assuming the calibration slope for the existing model is 1 in the development dataset (as it typically should be), the variance of the linear predictor in each group is,

$$s^2 = 2(\Phi^{-1}(C))^2, \tag{8}$$

where C is the C-statistic in the model development dataset and Φ^{-1} is the inverse of the standard normal distribution (ie, the z-score that corresponds to a probability of C of sampling a value less than z). Furthermore, again assuming the calibration slope is 1 in the development dataset, the difference in the mean for the events (μ_1) and nonevents (μ_2) groups is

$$\mu_1 - \mu_2 = s^2 = 2(\Phi^{-1}(C))^2, \tag{9}$$

and therefore $\mu_1 = \mu_2 + 2(\Phi^{-1}(C))^2$. Based on these results, we can define the two linear predictor distributions in the model development dataset as:

$$\begin{aligned} \text{Events group} &: \text{LP}_i \sim N(\mu_2 + s^2, s^2) \\ \text{Nonevents group} &: \text{LP}_i \sim N(\mu_2, s^2) \\ \text{where } \text{s}^2 &= 2(\Phi^{-1}(C))^2 \end{aligned} \tag{10}$$

We can also utilize Equation (10) to define the linear predictor distributions upon calibration, by plugging in the value of C as the anticipated C-statistic upon validation; this again assumes the calibration slope will be 1 upon validation and that the two linear predictor distributions are normal with a common variance. The value of μ_2 must ensure a mixture of the two distributions in Equation (10) has the correct outcome event proportion anticipated for the whole validation population. Box 2 describes how to identify the value of μ_2.

Once s^2 and μ_2 (and thus μ_1) are defined, the linear predictor distributions defined in Equation (10) can be used in the sample size calculation for the calibration slope, according to the process described in Box 1 followed by Equation (7). However, we emphasize that this approach makes strong assumptions of normality and equal variances in each outcome group, which may be unrealistic in most situations (especially the common variance assumption). We investigate this within an example described in Section 4.6.3. Stata code to implement the approach is provided in supplementary material S3.

BOX 2 How to identify the value of μ_2 in Equation (10) to define the means of the linear predictor distributions when they are assumed normal with a common variance in each of the two outcome groups

- Simulate a dataset with a large (eg, 1 000 000) number of participants
- Randomly generate event status of each participant from a Bernoulli (ϕ) distribution, where ϕ is the anticipated outcome event proportion in the validation population.
- Choose a value for μ_2, and then generate the prediction model's linear predictor (LP_i) values for each participant conditional on their designated group, and the normal distributions defined in Equation (10) based on the reported *C*-statistic.
- Randomly generate an observed outcome (0 or 1) for each participant based on a Bernoulli($\exp(LP_i)/(1+\exp(LP_i))$ distribution.
- Check that the proportion of observed outcome events in the entire dataset matches ϕ. If so, then the chosen μ_2 is correct; otherwise, repeat the process with a different μ_2 value.

Stata code to implement this is provided in Supplementary Material S3 of Appendix S1.

3.2.5 | Illustrative example

To illustrate the sample size calculation for the calibration slope, we now consider a published model to predict having a diagnosos of deep vein thrombosis (DVT), developed using data from 1295 individuals with about 285 (ie, $\phi = 0.22$) having the outcome event.[7] Overfitting was not a major concern as the sample size was much larger than that needed to minimize overfitting according to recent sample size guidance for prediction model development.[37,38] The model's linear predictor distribution was reported to be approximately $N(-1.75, 1.47^2)$ in the development dataset. Simulating values from this distribution, and converting back to predicted probabilities, the average probability is 0.22, which is the same as the outcome event proportion (ϕ) reported for the development dataset. This reassures us that the $N(-1.75, 1.47^2)$ is appropriate.

We plan to externally validate this model in the same underlying population, and assume that the model will be well calibrated. Hence, for part (A) in Box 1 we chose $\alpha = 0$ and $\beta = 1$, and for part (B) we assumed LP_i follows a Normal(-1.75, 1.47^2) distribution. In part (C) we generated LP_i values from this distribution for 1 000 000 individuals, and in part (D) we calculated a_i, b_i, and c_i for each individual. Lastly, in part (E), the expected values of a_i, b_i, and c_i were calculated from the mean values for a_i, b_i, and c_i in the dataset, which gave $I_\alpha = 0.1289$, $I_{\alpha,\beta} = -0.1237$ and $I_\beta = 0.2784$.

Assuming that we want a 95% confidence interval width of 0.2, and so target a SE(β) of 0.051, we can now apply Equation (7) to obtain the required sample size of:

$$N = \frac{I_\alpha}{\mathrm{SE}(\beta)^2(I_\alpha I_\beta - I_{\alpha,\beta}^2)} = \frac{0.1289}{0.051^2 \times ((0.1289 \times 0.2784) - (-0.1237)^2)} = 2406.6$$

Hence, 2407 participants are required to externally validate the DVT prediction model, which corresponds to about 529 events based on an anticipated proportion of participants with DVT of 0.22. In this particular example, the minimum required sample size is much larger than suggested by rules-of-thumb of 100 or 200 events.

3.3 | *C*-statistic

Various strategies have been proposed for deriving the SE of the *C*-statistic (SE(*C*)).[31,39] Here, we utilize the formula proposed by Newcombe,[40] because it is derived from a nonparametric approach (Mann-Whitney based method), and so does not make assumptions about the underlying distribution of the prediction model's linear predictor. The formula is,

$$\mathrm{SE}(C) \approx \sqrt{\frac{C(1-C)\left(1+\left(\frac{N}{2}-1\right)\left(\frac{1-C}{2-C}\right)+\frac{\left(\frac{N}{2}-1\right)C}{1+C}\right)}{N^2\phi(1-\phi)}}, \tag{11}$$

where the values of C and ϕ are those observed in the validation dataset.

Debray et al examined this SE solution in a range of empirical scenarios and showed that it generally performs well.[39] Feng et al concluded from a simulation that using this SE to derive Wald-based confidence intervals for the C-statistic is a good approach as "coverage probabilities are very close to the nominal level",[41] and generally perform better than other nonparametric alternatives. SE solutions for the C-statistic could alternatively be based on parametric assumptions about the distribution of the linear predictor in the outcome event and nonevent groups (eg, assuming the linear predictor distribution is normally distributed[41]). However, Feng et al concluded that the nonparametric approach of Equation (11) is preferable when the underlying parametric distribution is questionable, which will typically be the case when planning the external validation study.[41] Hence, Equation (11) seems a sensible solution to base a sample size equation on for the C-statistic, to avoid distributional assumptions for the linear predictor (which is unavoidable for the calibration slope as described in Section 3.2).

Equation (11) provides a formula for the SE of the C-statistic that depends on C, N, and ϕ. For a sample size calculation, we would ideally like to rearrange the equation to isolate N, but this is not possible. However, an iterative approach can be used to identify the value of N that gives a targeted value for SE(C), given values for the anticipated C and ϕ in the external validation population. Stata code for this iterative approach is provided in Supplementary Material S1 of Appendix S1, which takes only a few seconds to run. Based on the approach, sample sizes for estimating the C-statistic with a targeted confidence interval width of 0.1 (SE(C) = 0.0255) are shown in Figure 2, for a range of assumed outcome event proportions and C-statistics.

How should the researcher anticipate the C-statistic of the prediction model upon external validation? If the validation population is similar to that used for model development, we suggest assuming C-statistics that are the same or about 0.05 lower than the C-statistic reported for model development, and then taking the largest required sample size identified. Lower C-statistic values are often observed upon validation,[4] as prediction models are often subject to overfitting upon development, which (if not adjusted for) leads to optimistic performance estimates. In a review by Collins et al,[4] the mean C-statistic was 0.787 upon development and the mean was 0.757 upon external validation, and thus a mean reduction of 0.03. We suggest using a more conservative value of 0.05 to err on the side of caution. However, if the population for validation is expected to have a narrower case-mix than that used for development (eg, if validation done in only females or an older age group, when the development data included males and females across a wide age range), then it is essential to consider even lower values of the C-statistic, as narrower case-mix can substantially reduce the C-statistic.

3.3.1 | Illustrative example

As an example, consider that the C-statistic and outcome event proportion are anticipated to be 0.7 and 0.1, respectively, in the external validation. Then an iterative approach to solve Equation (11), using the Stata code provided in Supplementary Material S1 of Appendix S1, finds that a total sample size of 1154 participants (about 115 events) are required for an expected SE(C) of 0.0255. This SE corresponds to an expected 95% confidence interval of about 0.65 to 0.75, and thus a width of 0.1, which is reasonably precise. If the C-statistic and outcome event proportion are anticipated to be 0.8 and 0.5, respectively, then a total sample size of 302 participants (151 events) is required to target the same confidence interval width. Hence, as for O/E and calibration slope, again we see that the number of participants and events required for external validation is context specific, as it depends on the anticipated discrimination performance of the model and the outcome event proportion in the validation population.

3.4 | Net benefit

If a prediction model is to be considered for guiding clinical decisions at a particular risk threshold (p_t) then, alongside calibration and discrimination, it is also important to estimate the model's net benefit (NB_{p_t}) precisely in the external

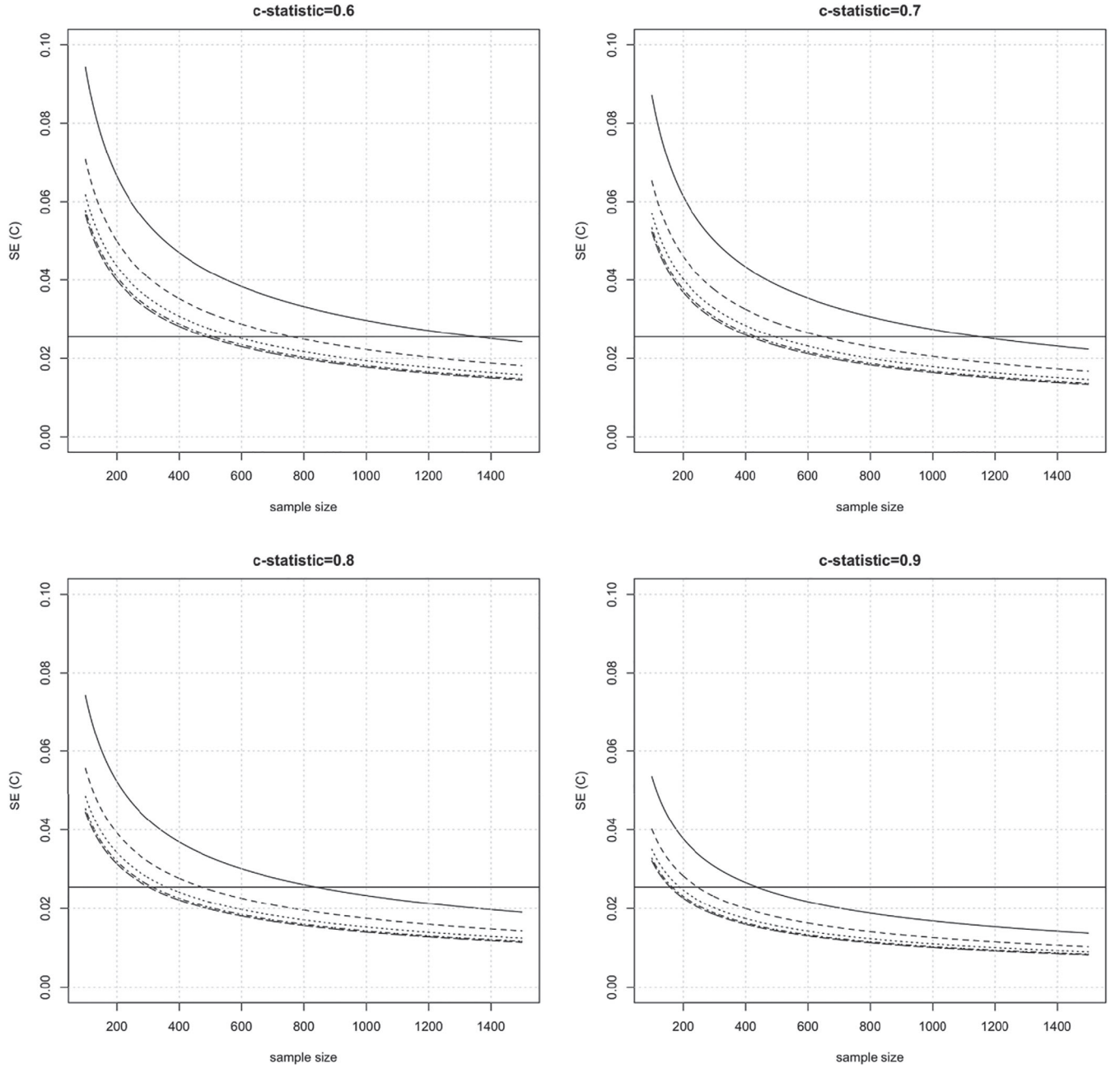

FIGURE 2 Values of the SE of the *C*-statistic (SE(*C*)) calculated using Equation (11) for a range of *C*-statistic and outcome event proportions, with the solid horizontal line corresponding to a targeted confidence interval with of 0.1 (SE(*C*) = 0.0255). Moving from top to bottom, the five curves on each plot correspond to outcome event proportions of 0.1, 0.2, 0.3, 0.4, and 0.5

validation dataset. Equation (3) shows that the estimate of NB_{p_t} is a function of ϕ, and the sensitivity and specificity of the model's classification at the chosen threshold.

Confidence intervals for NB_{p_t} are usually derived using bootstrapping.[42] However, Marsh et al proposed a Wald-based approach (ie, $sNB_{p_t} \pm (1.96 \times SE(sNB_{p_t}))$ based on asymptotic theory and using an empirical closed-form estimator for $SE(sNB_{p_t})$, the SE of the standardized NB_{p_t}.[43] Indeed, their concurrent aim was to inform sample size calculations for validation studies, and so the following simply summarizes their approach.

Marsh et al suggested that,[43]

$$SE(sNB_{p_t})^2 = \frac{1}{N}\left(\frac{\text{sensitivity}(1-\text{sensitivity})}{\phi} + \frac{w^2\ \text{specificity}\ (1-\text{specificity})}{1-\phi} + \frac{w^2(1-\text{specificity})^2}{\phi(1-\phi)}\right), \quad (12)$$

where $w = \frac{(1-\phi)}{\phi}\frac{p_t}{1-p_t}$. The three terms that are summed within the large bracket of Equation (12) account for the variance of estimates for sensitivity, specificity and outcome event proportion, respectively. The simulation study of Marsh et al showed this closed-form approach performs comparably to bootstrapping in most scenarios,[43] giving coverage close to 95% unless one of the three constituents (sensitivity, specificity, and outcome event proportion) is close to 0 or 1, but even then the coverage is still reasonable (close to 93%). Further, the required sample size is unlikely to be small, and so the asymptotic approach of Marsh et al is likely to be a good approximation for sample size purposes.

Rearranging Equation (12) we obtain:

$$N = \frac{1}{\text{SE}(s\text{NB}_{p_t})^2}\left(\frac{\text{sensitivity}(1-\text{sensitivity})}{\phi} + \frac{w^2\,\text{specificity}\,(1-\text{specificity})}{1-\phi} + \frac{w^2(1-\text{specificity})^2}{\phi(1-\phi)}\right). \quad (13)$$

Equation (13) allows us to calculate the sample size required for net benefit, conditional on prespecifying the target value for $\text{SE}(s\text{NB}_{p_t})$, and the anticipated values in the external validation population for the outcome event proportion, and sensitivity and specificity of the prediction model at threshold p_t. If there are a range of thresholds of interest, then the calculation should be repeated for each, and the largest required sample size across thresholds identified. Sensitivity and specificity can be inferred from the anticipated distribution of the linear predictor. Section 4.4 illustrates this.

3.4.1 | Illustrated example

Marsh et al[43] considered the external validation of a prediction model with an anticipated outcome event proportion of 0.72, a sensitivity of 0.60 and a specificity of 0.88 at a risk threshold of 0.80. Using Equation (3), this corresponds to an anticipated net benefit of,

$$\begin{aligned}\text{NB}_{p_t} &= (\text{sensitivity}\times\phi) - \left((1-\text{specificity})\times(1-\phi)\times\frac{p_t}{1-p_t}\right)\\ &= (0.60\times 0.72) - \left((1-0.88)\times(1-0.72)\times\frac{0.80}{1-0.80}\right) = 0.298,\end{aligned} \quad (14)$$

and so a standardized net benefit of $0.298/0.72 = 0.41$. If we target a SE of 0.051 (such that the 95% confidence interval width is about 0.2), then using Equation (13) and noting that $w = \frac{(1-\phi)}{\phi}\frac{p_t}{1-p_t} = \frac{(1-0.72)\times 0.80}{0.72\times(1-0.80)} = 1.556$, we find that the sample size required for the external validation study is,

$$\begin{aligned}N &= \frac{1}{\text{SE}(s\text{NB}_{p_t})^2}\left(\frac{\text{sensitivity}(1-\text{sensitivity})}{\phi} + \frac{w^2\,\text{specificity}\,(1-\text{specificity})}{1-\phi} + \frac{w^2(1-\text{specificty})^2}{\phi(1-\phi)}\right)\\ N &= \frac{1}{0.051^2}\left(\frac{0.60\times(1-0.60)}{0.72} + \frac{1.556^2\times 0.88\times(1-0.88)}{1-0.72} + \frac{1.556^2\times(1-0.88)^2}{0.72\times(1-0.72)}\right)\\ &= 545.5,\end{aligned} \quad (15)$$

and thus 546 participants are required (with about 393 events).

It is difficult to give a general recommendation about what constitutes a precise confidence interval for $s\text{NB}_{p_t}$. But at a minimum, we suggest to at least target a confidence interval whose lower bound is at the value of $s\text{NB}_{p_t} = 0$, as this would target strong evidence that the model (at the chosen threshold) has a higher net benefit than "treat none." Additionally, the researcher might also target a confidence interval whose lower bound is at the $s\text{NB}_{p_t}$ of the "treat all" strategy; if $s\text{NB}_{p_t}$ is greater than 0 for "treat all" this will require a narrower confidence interval and so a larger sample size than when comparing to "treat none." Also, care is needed when deciding which risk threshold(s) to consider, as the acceptable threshold(s) may vary between settings and hence also between the development and validation setting.

4 | SUMMARY OF SAMPLE SIZE PROPOSAL AND APPLIED EXAMPLE

We now summarize our sample size approach by providing a step-by-step application. We consider the sample size needed to externally validate the prediction model developed by Pavlou et al,[44] which estimates a patient's risk of failure of a mechanical heart valve within about 12 months. The prediction model was a logistic regression model with 10 predictors and was obtained using penalized estimation via the lasso. The model's linear predictor was:

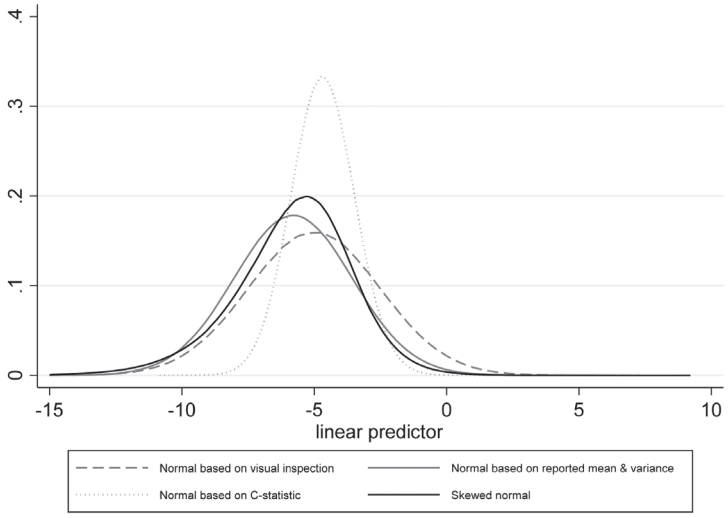

FIGURE 3 Comparison of the linear predictor distributions assumed for the prediction model in the applied example of Section 4

$$\begin{aligned}\text{LP}_i = &-6.65 - 0.16(\text{if female sex}) - 0.05(\text{age in years}) + 1.75(\text{body surface area in m2})\\ &+ 0.61(\text{if aortic size 23,27,29 or 31 mm}) + 0.43(\text{if mitral size 23-27 mm})\\ &+ 1.13(\text{if mitral size 29 mm}) + 1.77(\text{if mitral size 31 mm}) + 1.73(\text{if mitral size 33 mm})\\ &+ 0.64(\text{if fracture in batch}) + 1.22(\text{if date of manufacture after 1981})\end{aligned}$$

The model development dataset had a small outcome event proportion of 56/3118 = 0.018, and the reported *C*-statistic was 0.80 (95% CI: 0.78-0.82). The linear predictor distribution was shown in Figure 1 of the Pavlou et al article,[44] and on visual inspection appears approximately normally distributed with a mean of −5 and SD of about 2.5. However, when we used this distribution to simulate linear predictor values for a million participants, this gave an outcome event proportion of about 0.05, and not 0.018. On closer examination, the distribution was slightly skewed. By using the *sknor* package in Stata, and through trial and error, we identified that a skewed normal distribution with a mean of −5.8, a variance of 5, a skewness parameter of −0.5, and a kurtosis parameter of 4 was a better approximation, as it gave a similar distributional shape to that shown by Pavlou et al and an outcome event proportion of 0.018 in a large simulated dataset (Figure 3).

We now consider the sample size required for a new study to externally validate this prediction model, assuming that the validation population is similar to that used for model development. Corresponding Stata code is provided in Supplementary Materials S1 and S3 of Appendix S1.

4.1 | Step (i): Sample size for *O*/*E*

The first step is to consider the sample size for *O*/*E*. We assumed that the outcome event proportion will be 0.018 (as in the model development dataset), and targeted a confidence interval width of 1.0 for *O*/*E*. This corresponds to a desired SE(ln(*O*/*E*)) of 0.245 and, assuming *O*/*E* is 1, an expected confidence interval for *O*/*E* of about 0.62 to 1.62. The confidence interval may seem wide, but the upper bound of 1.62 still corresponds to a small absolute difference of 0.011 in the observed and expected outcome event proportions, assuming the latter is 0.018. Using Equation (5), the required sample size is,

$$N = \frac{(1-\phi)}{\phi\,(\text{SE}(\ln(O/E))^2} = \frac{(1-0.018)}{0.018 \times (0.245)^2} = 908.9,$$

and thus a minimum of 909 participants (about 16 events) are needed to target a precise estimate for *O*/*E* in this example. This is the minimum required, as other more stringent situations could be considered. Note, though, that if we rather assumed some miscalibration of, say, an *O*/*E* of 1.62 (which, as mentioned, corresponds to an observed outcome event risk of 0.029 compared to the expected value of 0.018), then a targeted confidence interval for *O*/*E* of 1.0 to 2.62 also requires a SE(ln(*O*/*E*)) of about 0.245, and hence 909 participants (16 events).

4.2 | Step (ii): Sample size for calibration slope

The second step is to consider the sample size for the calibration slope (β in Equation (2)). We assumed the prediction model will have at least "weak" level calibration in the validation population, such that $\alpha = 0$ and $\beta = 1$ for the calibration model of Equation (2).[13] Assuming the linear predictor ($\text{logit}(p_{\text{PREDi}})$) values had the skewed normal distribution described above, we applied the process described in Box 1 and this gave $I_\alpha = 0.0151$, $I_{\alpha,\beta} = -0.0408$, and $I_\beta = 0.1494$ (shown rounded to four decimal places here, but full values were used in the calculations that follow, and so readers will get small differences to the final sample size if trying to replicate by hand).

We targeted a SE(β) of 0.051, to aim for a 95% confidence interval width of about 0.2. This corresponds to 0.9 to 1.1 when $\widehat{\beta} = 1$, which we considered to represent strong evidence that the model's calibration slope is very close to 1.[37,38] Then, the required sample size to precisely estimate the calibration slope was obtained by applying Equation (7), which gave:

$$\begin{aligned} N &= \frac{I_\alpha}{\text{SE}(\beta)^2(I_\alpha I_\beta - I_{\alpha,\beta}^2)} \\ &= \frac{0.0151}{0.051^2 \times ((0.0151 \times 0.1494) - (-0.0408 \times -0.0408))} \\ &= 9834.5. \end{aligned}$$

Thus, a minimum of 9835 participants (about 177 events) are needed to target a precise estimate of the calibration slope in this example, assuming the calibration slope will be 1.

We also examined the required sample size assuming the calibration slope will be 0.8, which is consistent with the predictor effects being 20% too large in the model development dataset (eg, due to overfitting during model development). When fitting calibration model Equation (2), this corresponds to assuming values of $\beta = 0.8$ and $\alpha = -0.6$. The α value of -0.6 ensures the overall outcome event proportion is still 0.018, and was identified by trial and error by simulating a large dataset (as described in the Notes of Box 1). Under these conditions, the sample size required to estimate a precise calibration slope (ie, with targeted confidence interval of 0.7 to 0.9) is 7632 participants (137 events), which is lower than the 9835 participants (177 events) when assuming the calibration slope is 1.

4.3 | Step (iii): Sample size for *C*-statistic

The third step is to consider the sample size for the *C*-statistic. We anticipated that the *C*-statistic will be 0.80 in the validation population, as reported by Pavlou et al for their development population. We targeted a SE(C) of 0.0255, as this corresponds to an expected confidence interval width of 0.1 (ie, 0.75 to 0.85). We applied an iterative process to solve Equation (11), which identified that 4252 participants (about 77 events) are required to achieve the desired precision for the *C*-statistic. As a sensitivity analysis, we also considered *C*-statistics of 0.75 and 0.85, which suggested sample sizes of 5125 and 3271 participants, respectively. Therefore, to be conservative, a minimum sample size of 5125 (about 92 events) is recommended to precisely estimate the *C*-statistic.

4.4 | Step (iv): Sample size for net benefit

The fourth step is to consider the sample size for net benefit. Pavlou et al suggested their prediction model could be used to create four risk groups,[44] with the highest risk group defined by a probability threshold of $p_t = 0.08$. Hence, we assumed that net benefit at this threshold is of interest in the external validation study. No estimates of sensitivity and specificity were provided by Pavlou et al for this probability threshold, as the model's clinical utility was not the focus of their work. However, we used the following approach to obtain reasonable estimates for sensitivity and specificity at a probability threshold of $p_t = 0.08$:

- Simulate linear predictor (LP_i) values for a million participants from the skewed normal distribution previously described.

- Convert each linear predictor value back to a predicted probability (p_{PREDi}); assuming the model is well calibrated this is simply $\exp(\text{LP}_i)/(1+\exp(\text{LP}_i))$.
- Randomly generate outcome (0 or 1) values according to Bernoulli (p_{PREDi}) distribution for each participant.
- Classify participants as negative if their $p_{\text{PREDi}} < 0.08$, and positive otherwise.
- Calculate sensitivity as the proportion of individuals with an outcome value of 1 that were classed as positive.
- Calculate specificity as the proportion of participants with an outcome value of 0 that were classed as negative.

This process gave a sensitivity of 0.53 and a specificity of 0.96. The corresponding net benefit ($\text{NB}_{0.08}$) using Equation (3) is,

$$\begin{aligned}\text{NB}_{0.08} &= (\text{sensitivity} \times \phi) - \left((1-\text{specificity}) \times (1-\phi) \times \frac{p_t}{1-p_t}\right)\\ &= (0.53 \times 0.018) - \left((1-0.96) \times (1-0.018) \times \frac{0.08}{1-0.08}\right)\\ &= 0.0061,\end{aligned}$$

and the standardized net benefit ($s\text{NB}_{0.08}$) is $0.0061/0.018 = 0.34$.

For the sample size calculation, we targeted a confidence interval width for $s\text{NB}_{0.08}$ of 0.2, such that if $s\text{NB}_{0.08}$ is 0.34 then the expected confidence interval is 0.24 to 0.44. This confidence interval width was used by Marsh et al in their sample size proposal for net benefit,[43] and would target a confidence interval whose lower bound is well above the "treat none" $s\text{NB}_{0.08}$ of 0. This corresponds to targeting a $\text{SE}(s\text{NB}_{p_t})$ of 0.051, and applying Equation (13) the required sample size is,

$$\begin{aligned}N &= \frac{1}{\text{SE}(s\text{NB}_{p_t})^2}\left(\frac{\text{sensitivity}(1-\text{sensitivity})}{\phi} + \frac{w^2\ \text{specificity}\ (1-\text{specificity})}{1-\phi} + \frac{w^2(1-\text{specificity})^2}{\phi(1-\phi)}\right)\\ &= \frac{1}{0.051^2}\left(\frac{0.53(1-0.53)}{0.018} + \frac{4.744^2\ 0.96\ (1-0.96)}{1-0.018} + \frac{4.744^2(1-0.96)^2}{0.018(1-0.018)}\right)\\ &= 6442.2,\end{aligned}$$

and thus 6443 participants (about 116 events) are needed to target a precise estimate of net benefit in this example. Note that the net benefit for "treat all" is negative at the chosen risk threshold, and therefore is not as relevant a comparison as the "treat none" strategy.

4.5 | Step (vi): Determine the minimum required sample size

A summary of the results from steps (i) to (iv) is shown in Table 1. This suggests a study of at least 9835 participants (about 177 events) is required to target precise estimates for all four measures. This sample size is driven by calibration slope in this example, and also reflects the rarity of the outcome event. We emphasize that this is the minimum sample size required, as larger numbers are needed to go beyond "weak" level calibration assessments,[13] such as producing precise calibration curves or precise estimates of calibration performance in particular subgroups or regions of predicted risk.

4.6 | What if we had assumed an approximate normal distribution?

So far, we assumed the linear predictor values followed a skewed normal distribution which closely matched the distribution shown in Figure 1 of Pavlou et al,[44] and ensured the overall outcome event proportion was correct. However, such a distribution might not always be provided in the model development study. Hence, we now consider the impact on the sample size calculations had we simply assumed normal distributions for the linear predictor, based either on (i) visual inspection of Pavlou's Figure 1, (ii) a reported mean and variance of linear predictor values, or (iii) the reported *C*-statistic. The corresponding distributions are shown in Figure 3.

TABLE 1 Sample size and number of events required to target precise performance measures in an external validation study of the prediction model for mechanical heart valve failure,[44] with an assumed linear predictor that follows a skewed normal distribution and an outcome event proportion of 0.018

Performance measure	Assumed value	Targeted 95% CI width (corresponding SE)	Sample size: Number of participants (events) required
O/E	1	1 (0.245[a])	909 (16)
Calibration slope	1	0.2 (0.051)	9835 (177)
	0.8	0.2 (0.051)	7632 (137)
C-statistic	0.80	0.1 (0.0255)	4252 (77)
	0.75	0.1 (0.0255)	5125 (92)
Standardized net benefit	0.34	0.2 (0.051)	6443 (116)

[a] for SE of $\ln(O/E)$.

4.6.1 | Approximate normality based on visual inspection of the linear predictor distribution

Based on visual inspection of Figure 1 of Pavlou et al,[44] the prediction model's linear predictor distribution appears approximately normally distributed in the development dataset, with a mean of −5 and SD of about 2.5. Assuming this distribution and a calibration slope of 1, the sample size calculation (Box 1 followed by Equation (7)) suggests 4555 participants are required, about half of the 9835 participants previously identified when using the skewed normal distribution. This emphasizes the importance of understanding the potential linear predictor distribution as closely as possible. Here, the approximate normal distribution leads to a very different sample size, because it corresponds to an overall outcome event proportion of 0.05 (compared to the observed proportion of 0.018 in the development dataset), and has a wider distribution than the skewed normal distribution assumed in the previous sections, which leads to a lower sample size.

4.6.2 | Approximate normality based on reported mean and SD of the linear predictor distribution

Next, we consider the situation where we assume a normal distribution with the mean and SD as reported for the linear predictor in the development study article or as provided by the model developers. Linear predictor values generated from a (assumed to be true) skewed normal distribution have a mean of −5.799 and a SD of 2.237, so we take these as the values that would be reported or provided. Hence, assuming the linear predictor distribution to be $N(-5.799, 2.237^2)$, and assuming the calibration slope is 1, we repeated our sample size calculation for the calibration slope (Box 1 followed by Equation (7)). This suggested 8286 participants are required to estimate the calibration slope precisely. This is still somewhat lower than the 9835 participants based on the skewed normal distribution. However, it is a reasonable approximation, and is much closer than for the visual approach used in Section 4.6.1, as it corresponds to an outcome event proportion of 0.023 that is closer to the 0.018 reported for the model development dataset.

4.6.3 | Approximate normality based on reported *C*-statistic

Lastly, let us assume that only the *C*-statistic of 0.8 was reported from the model development study, and so no histograms, means or variances of the linear predictor distribution were available. As the Pavlou prediction model contains some continuous predictors, assuming the linear predictor is a continuous variable is appropriate. Given only the *C*-statistic, we might be tempted to apply Equation (10) and assume the linear predictor distribution in each outcome group is normal with a common variance of $s^2 = 2(\Phi^{-1}(C))^2 = 2(\Phi^{-1}(0.8))^2 = 2(0.8416)^2 = 1.417$. Applying the process in Box 2 identifies that an $\mu_2 = -4.7$ ensures a value of the overall outcome event proportion of 0.018, such that:

Event group: $LP_i \sim N(-4.7 + 1.417, 1.417)$
Nonevent group: $LP_i \sim N(-4.7, 1.417)$

Using these distributions in the sample size calculation for the calibration slope (Box 1 followed by Equation (7)) suggests that 17 049 participants (307 events) are required. This is almost twice the total number of participants required when we assumed the skewed normal distribution. This discrepancy is because their distributions are quite different (Figure 3), with the one based on the *C*-statistic much narrower. This suggests that using the *C*-statistic, while assuming normality and a common variance for the linear predictor distributions, may be a poor approximation in practice if these assumptions are not reliable. This emphasizes the potential importance of a pilot study in the absence of any other information, as discussed in Section 3.2.3.

5 | DISCUSSION

We have derived closed-form and iterative solutions for calculating the minimum sample size required for external validation of a clinical prediction model with a binary outcome. We focused on four key measures of model performance (*O*/*E*, calibration slope, *C*-statistic, and net benefit), which represent a model's calibration, discrimination and clinical utility. Calibration and net benefit are often neglected in current validation studies,[16] and we hope drawing attention to them at the stage of sample size calculation will encourage more researchers to evaluate them after their validation data is obtained. The largest sample size needed to precisely estimate all four key measures of model performance is the *minimum* recommended for the external validation study (Section 2.1.3).

Unlike blanket rules-of-thumb, our calculations allow sample size to be tailored to the actual prediction model of interest, as the equations require the user to specify the anticipated outcome event proportion, the model's predictive performance and the distribution of predicted values in the validation population. Obtaining such information is, like in any other sample size calculation (eg, for randomized trials), the stumbling block and may require a pilot study in the validation population. However, starting from the model development population characteristics will often be sensible, and the user could consider multiple plausible scenarios (eg, potential outcome event proportions, linear predictor distributions), and conservatively adopt the largest sample size identified. If the model to be validated was poorly developed (eg, without adjustment for overfitting), the performance in the validation sample is likely to be worse than reported for model development. This should be considered when investigating required sample size. Previous work suggests lower sample sizes are usually required to precisely estimate calibration slopes <1 than compared to when the slope is assumed to be 1[14,15]; hence, assuming a slope of 1 will usually be sufficient. However, the sample size for the *C*-statistic will usually be larger when the *C*-statistic assumed to be lower.

Defining the target precision for each performance measure is subjective, and potentially context specific, but our applied examples provide an illustration of potential choices. We anticipate that the final required sample size usually will be driven by the calibration slope calculation, but it is important to check the other three criterion for completeness. Despite leading to the largest sample sizes, the calibration slope criterion is still only aimed at precisely estimating "weak" level calibration,[13] and we anticipate larger sizes are needed to precisely estimate nonparametric calibration curves and calibration in subgroups, which represents "moderate" toward "strong" level calibration. This is illustrated in Supplementary Material S4 of Appendix S1, where a hypothetical validation study for the Pavlou model containing 9835 participants (ie, the number recommended by our sample size calculation) obtains a precise calibration slope but still relatively imprecise *O*/*E* estimates in subgroups. Indeed, rules of thumb of at least of 100 events and 100 nonevents (or 200) were originally driven by the desire to produce precise calibration curves.[12,13] Hence, if larger sample sizes are achievable than our calculations suggest, we recommend aiming for them still, as our recommendations represent the minimum required.

Stata code to implement our calculations is provided in Supplementary Material S1 of Appendix S1, and provides results within 2 seconds to 2 minutes, depending on the linear predictor distribution to be simulated. Supplementary Material S3 of Appendix S1 also provides Stata code to derive the linear predictor distribution from a reported *C*-statistic, under normality and other assumptions. We plan to embed this code within a formal package within Stata and R, to complement our other package, *pmsampsize*, which focuses on sample size for prediction model development.[37,38,45] Our closed-form solutions can be seen as complementary to the fully simulation-based proposal of Snell et al[14] The Supplementary Material S2 of Appendix S1 compares the two approaches in an applied example, and the findings are very similar. Our approach is clearly faster and potentially more accessible to a broader set of researchers, as it is based on closed-form and simpler iterative solutions. However, a fully simulation-based approach is more flexible, and can consider precision of measures for which closed-form solutions for precision are not available (eg, the ECI or ICI[18,19]); it may

also more accurately reflect sampling variability, and can display the variability (across all simulations) of confidence interval widths that could be observed for a particular target sample size.

We do not recommend basing sample size calculations on power calculations, such as the power to detect whether calibration or discrimination are different from a prespecified null hypothesis value (eg, calibration slope of 1). This may lead to imprecise estimates in some situations, which is unhelpful as the aim of a validation study is to establish a model's predictive performance reliably.

If an external dataset is already available (ie, sample size is fixed), our approach can be used to ascertain the expected precision for that known sample size and observed linear predictor distribution. This will help researchers (and potential funders) to ascertain if it is fit for purpose. Our calculations assume validation studies will utilize a cohort study design for prognostic prediction models, or a random or consecutive sample for diagnostic prediction models. Marsh also considers sample size for net benefit using a nested case-control design.[43]

In summary, we propose that precise performance estimates should be targeted when planning external validation studies of a clinical prediction model, and the minimum sample size required can be determined through the iterative and closed-form solutions presented in this paper. We hope this is helpful to researchers when designing their external validation studies or deciding whether an existing dataset is fit for purpose.

ACKNOWLEDGEMENTS

Gary Collins is supported by Cancer Research UK (program grant: C49297/A27294) and the NIHR Biomedical Research Centre, Oxford. Kym Snell is funded by the National Institute for Health Research School for Primary Care Research (NIHR SPCR). Thomas Debray is funded by the Netherlands Organisation for Health Research and Development (grant 91617050) and this project received funding from the European Union's Horizon 2020 research and innovation programme under ReCoDID grant agreement no. 825746. This publication presents independent research funded by the National Institute for Health Research (NIHR). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. We would like to thank two anonymous reviewers for their constructive feedback and suggestions that helped us to improve the article upon revision.

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
Thomas P. A. Debray https://orcid.org/0000-0002-1790-2719
Lucinda Archer https://orcid.org/0000-0003-2504-2613
Joie Ensor https://orcid.org/0000-0001-7481-0282
Maarten van Smeden https://orcid.org/0000-0002-5529-1541

REFERENCES

1. Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of covid19 infection: systematic review and critical appraisal. *BMJ*. 2020;369:m1328.
2. 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.
3. Steyerberg EW, Moons KG, van der Windt DA, et al. Prognosis research strategy (PROGRESS) 3: prognostic model research. *PLoS Med*. 2013;10(2):e1001381.
4. Wyatt J, Altman DG. Commentary: prognostic models: clinically useful or quickly forgotten? *BMJ*. 1995;311:1539-1541.
5. Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med*. 2000;19(4):453-473.
6. Debray TP, Vergouwe Y, Koffijberg H, et al. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol*. 2015;68(3):279-289.
7. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010;21(1):128-138.
8. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med*. 2015;162:55-63.
9. Moons KG, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med*. 2015;162(1):W1-W73.

10. Vergouwe Y, Steyerberg EW, Eijkemans MJ, et al. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *J Clin Epidemiol*. 2005;58(5):475-483.
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. Snell KIE, Archer L, Ensor J, et al. Sample size required for external validation of a clinical prediction model: simulation-based calculations are more flexible and reliable than rules-of-thumb. *J Clin Epidemiol*. 2021;135:79-89.
14. Archer L, Snell KIE, Ensor J, et al. Minimum sample size for external validation of a clinical prediction model with a continuous outcome. *Stat Med*. 2020;40:133-146.
15. Van Calster B, McLernon DJ, van Smeden M, et al. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17(1):230.
16. Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. *BMJ*. 2009;338:b605.
17. Van Hoorde K, Van Huffel S, Timmerman D, et al. A spline-based tool to assess and visualize the calibration of multiclass risk predictions. *J Biomed Inform*. 2015;54:283-293.
18. Austin PC, Steyerberg EW. The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. *Stat Med*. 2019;38:4051-4065.
19. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J*. 2014;35(29):1925-1931.
20. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. New York, NY: Springer; 2009.
21. 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.
22. Copas JB. Regression, prediction and shrinkage. *J R Stat Soc B Methodol*. 1983;45(3):311-354.
23. Copas JB. Using regression models for prediction: shrinkage and regression to the mean. *Stat Methods Med Res*. 1997;6(2):167-183.
24. Van Houwelingen JC. Shrinkage and penalized likelihood as methods to improve predictive accuracy. *Stat Neerl*. 2001;55:17-34.
25. 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.
26. Localio A, Goodman S. Beyond the usual prediction accuracy metrics: reporting results for clinical decision making. *Ann Intern Med*. 2012;157(4):294-295.
27. Moons KG, Altman DG, Vergouwe Y, et al. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *BMJ*. 2009;338:b606.
28. Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. *Ann Intern Med*. 2006;144(3):201-209.
29. 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.
30. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*. 2006;26(6):565-574.
31. Debray TPA, Damen JAAG, Riley RD, et al. A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes. *Stat Methods Med Res*. 2019;28:2768-2786.
32. Dorfman R. A note on the method for finding variance formulae. *Biometrics*. 1938;1:129-137.
33. Borenstein M, Rothstein H, Cohen J. *Power and Precision*. Englewood, NJ: Biostat Inc.; 2001.
34. Demidenko E. Sample size determination for logistic regression revisited. *Stat Med*. 2007;26(18):3385-3397.
35. Novikov I, Fund N, Freedman LS. A modified approach to estimating sample size for simple logistic regression with one continuous covariate. *Stat Med*. 2010;29(1):97-107.
36. Austin PC, Steyerberg EW. Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. *BMC Med Res Methodol*. 2012;12:82.
37. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ*. 2020;368:m441.
38. Riley RD, Snell KI, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: part II - binary and time-to-event outcomes. *Stat Med*. 2019;38(7):1276-1296.
39. Debray TP, Damen JA, Snell KI, et al. A guide to systematic review and meta-analysis of prediction model performance. *BMJ*. 2017;356:i6460.
40. Newcombe RG. Confidence intervals for an effect size measure based on the Mann-Whitney statistic. Part 2: asymptotic methods and evaluation. *Stat Med*. 2006;25(4):559-573.
41. Feng D, Cortese G, Baumgartner R. A comparison of confidence/credible interval methods for the area under the ROC curve for continuous diagnostic tests with small sample size. *Stat Methods Med Res*. 2017;26:2603-2621.
42. 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.
43. Marsh TL, Janes H, Pepe MS. Statistical inference for net benefit measures in biomarker validation studies. *Biometrics*. 2020;76(3):843-852.
44. Pavlou M, Ambler G, Seaman SR, et al. How to develop a more accurate risk prediction model when there are few events. *BMJ*. 2015;351:h3868.
45. Riley RD, Snell KIE, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: part I - continuous outcomes. *Stat Med*. 2019;38(7):1262-1275.

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, Debray TPA, Collins GS, et al. Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Statistics in Medicine*. 2021;40:4230–4251. https://doi.org/10.1002/sim.9025