





Individual participant data meta-analysis for external validation, recalibration, and updating of a flexible parametric prognostic model

Joie Ensor¹  | Kym I. E. Snell¹  | Thomas P. A. Debray^{2,3}  |
Paul C. Lambert^{4,5} | Maxime P. Look⁶ | Mamas A. Mamas^{1,7} | Karel G. M. Moons^{2,3} |
Richard D. Riley¹ 

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

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

³Cochrane Netherlands, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands

⁴Biostatistics Research Group, Department of Health Sciences, University of Leicester, Centre for Medicine, Leicester, UK

⁵Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

⁶Department of Medical Oncology, Erasmus MC Cancer Institute, Erasmus University Medical Center, Rotterdam, The Netherlands

⁷Department of Cardiology, Royal Stoke University Hospital, Stoke-on-Trent, UK

Correspondence

Joie Ensor, Centre for Prognosis Research, School of Medicine, Keele University, Keele, Staffordshire, UK.
Email: j.ensor@keele.ac.uk

Individual participant data (IPD) from multiple sources allows external validation of a prognostic model across multiple populations. Often this reveals poor calibration, potentially causing poor predictive performance in some populations. However, rather than discarding the model outright, it may be possible to modify the model to improve performance using recalibration techniques. We use IPD meta-analysis to identify the simplest method to achieve good model performance. We examine four options for recalibrating an existing time-to-event model across multiple populations: (i) shifting the baseline hazard by a constant, (ii) re-estimating the shape of the baseline hazard, (iii) adjusting the prognostic index as a whole, and (iv) adjusting individual predictor effects. For each strategy, IPD meta-analysis examines (heterogeneity in) model performance across populations. Additionally, the probability of achieving good performance in a new population can be calculated allowing ranking of recalibration methods. In an applied example, IPD meta-analysis reveals that the existing model had poor calibration in some populations, and large heterogeneity across populations. However, re-estimation of the intercept substantially improved the expected calibration in new populations, and reduced between-population heterogeneity. Comparing recalibration strategies showed that re-estimating both the magnitude and shape of the baseline hazard gave the highest predicted probability of good performance in a new population. In conclusion, IPD meta-analysis allows a prognostic model to be externally validated in multiple settings, and enables recalibration strategies to be compared and ranked to decide on the least aggressive recalibration strategy to achieve acceptable external model performance without discarding existing model information.

KEYWORDS

external validation, IPD Meta-analysis, model recalibration, model updating, time-to-event models

1 | INTRODUCTION

Prognostic models estimate the risk of future outcomes in individuals to aid clinical decision making for both clinicians and patients.¹⁻³ External validation studies aim to evaluate an existing model's performance in new data distinct from that used for model development, in populations with similar or increasingly different case-mix.²⁻⁴ Evidence suggests that there is currently much research waste where prognostic models are concerned, with many newly published models but relatively few validations of existing models.²⁻⁵ Good performance at external validation is an important step toward the uptake of a model in practice.^{1-3,6-8} Often external validation reveals poor model performance, however, rather than discarding the model outright, it may be possible to modify the model to improve performance by identifying and addressing the reasons for poor performance. Poor external validity may be due to various reasons including a poor model development strategy, the external population itself (eg, differences in measurement and definition of outcomes or predictors, or differences in case-mix between external and development samples), or simply due to chance variation, with many validation studies having small sample sizes.⁹⁻¹¹

Development of a new model solely due to poor external performance of an existing model is counterintuitive to evidence based medicine, as it discards useful information gleaned from previous patient populations.^{9,12-18} Recalibration offers a potential solution to poor performance of existing models in new patients; instead of developing a new model researchers can recalibrate or update existing models, combining information from both previous and new patients.^{9,18} Methodology in this area is well developed for both binary and time-to-event outcomes.^{7,12,13,15,17} In particular, approaches to model recalibration and updating for time-to-event models have been described by van Houwelingen et al, and are based on parametric models as these allow estimation (and thus recalibration) of the baseline hazard, an essential part in obtaining absolute risk predictions from a survival model.^{13,15}

Usually, only one dataset is available for external validation and recalibration of a prognostic model. However, increasingly multiple datasets (eg, from multiple studies or sources) are available for external validation, meaning model performance can be assessed on multiple occasions and summarized using meta-analysis methods.¹¹ This allows potential comparison of different recalibration methods across a set of validation studies of differing case-mix. Previous work has proposed the use of individual participant data (IPD) meta-analysis to compare model implementation strategies focusing on logistic regression models, and multivariate meta-analysis.^{14,16,19} Here, our focus is on external validation of time-to-event models with IPD from multiple studies, specifically flexible parametric (FP) prognostic models. Royston and Parmar proposed FP models to extend standard parametric survival models to allow more flexibility to capture realistic baseline hazard shapes.²⁰⁻²² First, we show that this flexibility allows us to apply different types of recalibration of the baseline hazard, allowing further options to tailor the model without discarding existing information from the published model. Second, we propose an IPD meta-analysis methodology that allows us to compare and rank the impact of these recalibration methods on the predictive performance of a model in new populations, with the aim of identifying the simplest method to achieve good model performance while re-estimating as little as possible from the existing model.

The remainder of this article is structured as follows. Section 2 introduces a motivating example where IPD are available from multiple studies to externally validate a prognostic model in breast cancer. As we assume a prognostic model already exists, this section also briefly describes the development of an artificial existing model to be used in the remainder of the article. Section 3 describes the methods and advantages of FP modeling using the Royston-Parmar approach, and the methods for synthesis of model performance statistics from multiple validation studies. Section 4 then introduces options for model recalibration, and considers how IPD meta-analysis allows these methods to be compared and ranked. Section 5 applies the methods to the motivating example data, and Section 6 concludes with some discussion.

2 | MOTIVATING EXAMPLE IN BREAST CANCER

To illustrate the methods proposed in this article, we use an IPD meta-analysis dataset that contains 5978 breast cancer patients, with follow-up ranging from 1 to 10 years (for more information on the original study see Look et al²³). It was formed by pooling datasets from eight centers (hereafter referred to as "studies" for simplicity) across six countries, with Rotterdam having the largest patient numbers (see Table 1). While the dataset is not publicly available, a closely related dataset for the Rotterdam data is available at <https://www.uniklinik-freiburg.de/imbi/stud-le/multivariable-model-building.html#c134656>.

TABLE 1 Summary statistics for Look et al dataset

Model phase	Used for external validation								
	Development	Rotterdam	Sweden	Lille	Nijmegen	St Cloud	Switzerland	Denmark1	Denmark2
Total sample size	2627		621	552	293	499	620	444	322
Recurrence/Death [#]	1224 (47%)		137 (22%)	150 (27%)	80 (27%)	168 (34%)	150 (24%)	211 (48%)	123 (38%)
Follow up (Months)*	63.64		106.84	60.71	52.90	97.84	42.71	48.34	55.36
Min	1.18		3.04	1.08	1.28	4.01	1.71	1.02	1.02
Max	120.00		120.00	120.00	120.00	120.00	83.45	120.00	100.17
Predictors									
Age (Years)									
Mean (SD)	56.46 (13.28)		58.47 (11.64)	57.00 (11.19)	56.68 (13.01)	58.82 (12.56)	57.94 (11.31)	53.86 (10.85)	56.21 (10.73)
Lymph nodes [#]									
0	1371 (52.19)		226 (36.39)	381 (69.02)	153 (52.22)	233 (46.69)	357 (57.58)	299 (67.34)	152 (47.2)
1 to 3	684 (26.04)		243 (39.13)	96 (17.39)	89 (30.38)	177 (35.47)	165 (26.61)	95 (21.4)	89 (27.64)
4 to 10	422 (16.06)		125 (20.13)	57 (10.33)	33 (11.26)	72 (14.43)	56 (9.03)	46 (10.36)	57 (17.7)
> 10	150 (5.71)		27 (4.35)	18 (3.26)	18 (6.14)	17 (3.41)	42 (6.77)	4 (0.9)	24 (7.45)
Menopausal status [#]									
Pre	1076 (40.96)		191 (30.76)	193 (34.96)	103 (35.15)	146 (29.26)	206 (33.23)	220 (49.55)	104 (32.3)
Post	1551 (59.04)		430 (69.24)	359 (65.04)	190 (64.85)	353 (70.74)	414 (66.77)	224 (50.45)	218 (67.7)
Tumour size [#]									
≤ 20mm	1177 (44.8)		217 (34.94)	302 (54.71)	102 (34.81)	211 (42.28)	298 (48.06)	179 (40.32)	96 (29.81)
20-50 mm	1296 (49.33)		396 (63.77)	242 (43.84)	165 (56.31)	271 (54.31)	306 (49.35)	232 (52.25)	192 (59.63)
>50 mm	154 (5.86)		8 (1.29)	8 (1.45)	26 (8.87)	17 (3.41)	16 (2.58)	33 (7.43)	34 (10.56)
Adjuvant treatment [#]									
No	1998 (76.06)		142 (22.87)	278 (50.36)	172 (58.7)	177 (35.47)	125 (20.16)	310 (69.82)	132 (40.99)
Yes	629 (23.94)		479 (77.13)	274 (49.64)	121 (41.3)	322 (64.53)	495 (79.84)	134 (30.18)	190 (59.01)

Note: RFS, Recurrence free survival; * Median; # Number and percentage.

TABLE 2 Predictor effect estimates for the existing model

Predictor	Beta	SE(Beta)	HR	Lower 95% CI	Upper 95% CI
Age	-0.019	0.004	0.981	0.973	0.989
Tumour size					
≤ 20 mm	ref				
20-50 mm	0.472	0.111	1.602	1.399	1.835
>50 mm	0.833	0.271	2.299	1.825	2.897
Lymph nodes					
0	ref				
1-3	0.647	0.171	1.909	1.602	2.275
4-10	1.260	0.311	3.525	2.965	4.191
>10	1.682	0.620	5.376	4.289	6.739
Menopausal status					
Pre	ref				
Post	0.236	0.130	1.267	1.036	1.548
Adjuvant treatment					
No	ref				
Yes	-0.454	0.052	0.635	0.540	0.747

Abbreviations: Beta, regression coefficients; SE, standard error; HR, hazard ratio; CI, confidence interval.

The focus of this article is on using IPD from multiple studies to externally validate and recalibrate an existing prognostic model that estimates an individual's probability of recurrence-free survival (RFS) over time from surgery, defined as the time to recurrence or death from any cause. Though there is also methodological interest in how to *develop* a prognostic model utilizing more than one study,^{14,24} this article focuses solely on how to perform external validation and model recalibration of an *existing* (ie, previously published) prognostic model using IPD from multiple studies. To this end, for illustrative purposes (and to retain most studies for external validation), we selected just one study (Rotterdam) as a derivation dataset in which to build an artificial "existing" prognostic model. This ensured that the remaining seven studies were available for use within the external validation exercise.

For brevity we present only the essential information regarding the artificial existing model here, with extensive details provided as online supplementary material including: the model development process, sample size considerations, estimated baseline hazard, how the model would be used to estimate individual risk predictions, and the apparent performance of the model. Details of the estimated predictor effects for the existing model are provided in Table 2.

We now take this developed model as our existing prognostic model, and rather focus our attention on undertaking external validation and recalibration of the model in the seven remaining IPD studies. The number of events in these studies ranged from 80 to 211 per study, with a total of 1019 events across all validation studies (see Table 1). In terms of validation, current evidence suggests that at least 100-200 events and non-events are required for external validation,²⁵⁻²⁷ meaning that the breast cancer data is likely sufficient both as a whole (1019 in total) and in each study separately, perhaps apart from the Nijmegen study which has sample size below 100 events.²⁸ Given the shortest follow-up was 83 months in the Switzerland study, we truncated follow-up at 72 months so as to reduce the possible influence of long-term survivors on parameters.²⁹

It is helpful to assess how similar the development and validation cohorts are to identify which validation cohorts may allow us to measure model transportability or model reproducibility.⁴ There are a number of ways to assess this relatedness from simple comparison of summary measures as presented in Table 1, comparing baseline hazard shape and magnitude, to comparisons of the prognostic index spread and location between cohorts. Such explorations are included in our online supplementary material.

3 | METHODS FOR EXAMINING PERFORMANCE OF A FP MODEL USING IPD META-ANALYSIS

In this section, we outline the methods for prognostic modeling using Cox and FP models, and discuss the advantages of the FP approach for prognostic modeling. We also introduce methods for synthesis of model performance statistics from multiple validation studies.

3.1 | Prognostic models with time-to-event outcomes

3.1.1 | Prognostic models using the Cox proportional hazards model

The Cox proportional hazards model is commonly used to develop prognostic models for time-to-event outcomes. The model can be written as follows,

$$h(t|\mathbf{x}_i) = h_0(t) \exp(PI_i) \tag{1}$$

where the hazard function, $h(t|\mathbf{x}_i)$, is dependent on the prognostic index, $PI_i = \beta_1x_{i1} + \beta_2x_{i2} + \dots + \beta_px_{ip}$ (where x_{ip} correspond to predictor values for subject i and β_p represent the corresponding predictor effects (log hazard ratios)) and the baseline hazard function, $h_0(t)$, which is equivalent to $h(t|\mathbf{x}_i = 0)$. The Cox model is referred to as a semiparametric model because it does not assume (nor estimate) a specific shape for the baseline hazard function.

However, the cumulative hazard function $H(t|\mathbf{x}_i)$ is crucial to calculating individual survival predictions and so, after fitting the Cox model, a non-parametric approach or smoothing methods such as splines can be used to estimate the baseline cumulative hazard function ($\hat{H}(t|\mathbf{x}_i = 0) = \hat{H}_0(t)$ for brevity). Then the baseline survival function can be obtained by:

$$\hat{S}_0(t) = \exp(-\hat{H}_0(t)) \tag{2}$$

This can be combined with the $\hat{\beta}$'s from maximum likelihood estimation of the Cox model to calculate survival probabilities at time t for an individual with covariate values x_{ip} using the survival function:

$$\hat{S}(t|\mathbf{x}_i) = \hat{S}_0(t)^{\exp(\hat{PI}_i)} \tag{3}$$

3.1.2 | Prognostic models using a FP framework

Unlike the Cox model, FP models directly model the shape of the log baseline hazard, removing the need for post-estimation smoothing.²⁰⁻²² Standard parametric models assume distributional shapes for the baseline hazard, but are restricted and often unable to capture realistic hazard functions which may rise and fall over time. Using FP models allows for increased flexibility in the shape of the baseline hazard through the use of restricted cubic splines to model the log baseline cumulative hazard function, which leads to the following model for $H(t|\mathbf{x}_i)$:

$$\ln(H(t|\mathbf{x}_i)) = \text{spline}(\ln(t)|\gamma, \mathbf{n}) + PI_i \tag{4}$$

where $\text{spline}(\ln(t)|\gamma, \mathbf{n})$ is the restricted cubic spline function for the baseline cumulative hazard and the PI_i as before. Equation (4) is typically estimated using maximum likelihood (for example using the `stpm2` module in Stata) with suitable starting values for the coefficients, β_p , derived from a Cox model with covariates x_{ip} .^{21,22,30}

Individual estimates of survival probability can be calculated as before using Equations (2) and (3). We now briefly explain the derivation of the restricted cubic spline function in Equation (4).

3.1.3 | Restricted cubic splines

FP models utilize restricted cubic splines to flexibly model the baseline hazard on the log-cumulative hazard scale. By fitting cubic splines between given joining points over time, known as knots, FP models can better capture fluctuations

in the baseline hazard. A smooth function is obtained by constraining the function and its first and second derivatives to zero at the knots. Restricted cubic splines are used in preference over cubic splines to force the function to also be linear before the first and after the last knot, so as to ensure a more biologically plausible function in the tails of the distribution where there is more likely to be sparse data.³⁰ We define a vector of n knots, then the spline function can be written in terms of parameters γ and the derived variables $z_i(t)$ as below,

$$\text{spline}(\ln(t)|\gamma, \mathbf{n}) = \gamma_0 + \gamma_1 z_1(t) + \cdots + \gamma_{n-1} z_{n-1}(t) \quad (5)$$

$$\begin{aligned} z_1(t) &= \ln(t) \\ z_i(t) &= (\ln(t) - k_i)_+^3 - \lambda_i (\ln(t) - k_1)_+^3 - (1 - \lambda_i) (\ln(t) - k_n)_+^3 \end{aligned} \quad (6)$$

where the $+$ notation indicates as follows,

$$(u)_+ = \begin{cases} u, & u > 0 \\ 0, & u \leq 0 \end{cases}$$

and where λ_i may be calculated using the following formula for $i = 2, \dots, n-1$;

$$\lambda_i = \frac{k_n - k_i}{k_n - k_1} \quad (7)$$

3.1.4 | Advantages of FP models for prognosis

There are several benefits to using FP models over other more traditional parametric or semi-parametric models for the purpose of prognostic modeling. In particular, semi-parametric models such as the Cox model do not explicitly model the baseline hazard, providing only relative effects rather than absolute risk estimates. Being able to parameterize and estimate the baseline hazard is essential for prognostic model research; first, in order to obtain individualized absolute risk predictions over time and second, for out-of-sample prediction enabling external validation. We examine the benefits of using FP models further in the online supplementary material, where we also consider the use of fractional polynomials as an alternative to spline-based parametrizations of the baseline hazard.³¹

The focus of this article here onward is on how to validate the predictive performance of an existing FP model using IPD from multiple studies. The following sections introduce important model performance measures for this purpose, and explain potential strategies for recalibration of a model to improve performance in external populations.

3.2 | Quantifying the predictive performance of a FP model upon external validation

We now consider how to examine the predictive performance of a FP prognostic model upon external validation, in terms of calibration and discrimination.³² Various statistics can be calculated for these measures.³³ In this article, overall calibration was assessed using the expected (E) and observed (O) survival probabilities, to calculate the ratio (E/O) at specific time points, as well as overall and time-specific calibration plots across the risk spectrum to summarize calibration performance over time.^{1,34} If the E/O statistic at a particular time-point is greater than, or less than one, there is indication of either systematic over or under prediction, respectively. An E/O statistic of one indicates perfect agreement, either overall or within specific subgroups.

The discrimination performance of a FP model can be measured using Harrell's C-statistic,³⁵ and Royston's D-statistic.³⁶ Harrell's C-statistic reveals the separation of predicted risks for those with and without the outcome.³⁵ Royston's D-statistic is a measure of separation in survival curves related to the standard deviation of the PI; it can be interpreted as the log hazard ratio between two groups defined by dichotomizing the PI at the median.³⁶ Royston's R^2_D gives a measure of the proportion of explained variation based on the D-statistic, which is therefore similar to the interpretation of R^2 in linear regression models.³⁶

3.3 | External validation in multiple studies with meta-analysis of performance

Given IPD from multiple studies, the predictive performance of an existing model can be evaluated multiple times. This leads to multiple estimates for each validation statistic of interest (eg, Harrell’s C-statistic), and so naturally lends itself to a formal meta-analysis to summarize performance across studies, and examine the magnitude and potential sources of heterogeneity.^{14,16,37,38}

Let Y_i be the estimate of a particular performance statistic of interest where i represents the validation study and let S_i^2 be the associated variance of Y_i , then a random-effects meta-analysis could be used:

$$Y_i \sim N(\theta_i, S_i^2) \quad \theta_i \sim N(\theta, \tau^2) \tag{8}$$

This model can be estimated using restricted maximum likelihood (REML) estimation, and assumes that Y_i follows a normal distribution around the i^{th} study’s true performance, θ_i and that θ_i is also normally distributed around an average performance, θ and a between-study variance τ^2 . The latter allows for heterogeneity in predictive performance, which is expected because of changes in case-mix variation, outcome incidence and clinical policies across different populations and settings.¹¹ A key assumption is normality of the performance estimate in each study, for which appropriate transformations are needed, in particular the logit(C-statistic), logit(R^2_D), and ln(E/O) best meet this assumption.^{19,37,39}

Following the meta-analysis, usually of most interest will be the summary (average) performance, θ , and its 95% confidence interval, which is usually derived by $\theta \pm 1.96 SE(\theta)$, where $SE(\theta)$ is the standard error of θ . Alternative methods for deriving confidence intervals have been proposed to improve the coverage of confidence intervals in meta-analysis and could be adopted here also, such as the Hartung-Knapp modification to additionally account for uncertainty in the between-study variance estimate.⁴⁰⁻⁴² Also of interest may be an approximate prediction interval, such as a 95% prediction interval.

$$\theta \pm t_{k-2, 0.975} \sqrt{\tau^2 + SE(\theta)^2} \tag{9}$$

where k is the number of validation studies in the meta-analysis. This prediction interval infers the potential model performance in a new population/setting similar to those included in the meta-analysis.^{43,44} A narrower prediction interval implies more consistent performance in new external populations, and is thus desirable if the model is to be generalizable outside of a few local settings. Stata code to perform the meta-analysis methods described above is provided as online supplementary material, for Stata version 15.1.⁴⁵

4 | RECALIBRATION STRATEGIES IN A SINGLE VALIDATION STUDY

If an existing model’s calibration performance is sub-optimal or considered inadequate upon external validation, then—rather than discarding the model outright—recalibration strategies may be considered to improve it. In this section, four possible methods are proposed for recalibrating a FP prognostic model within a *single* validation study, and then we suggest how IPD meta-analysis methods can be used to compare and rank each of these across multiple studies.

4.1 | Recalibration options

The following recalibration methods progressively increase the extent to which the model is adjusted, and closely link to previously selected methods for a binary outcome example.⁴⁶ Stata code is provided as online supplementary material, for each of the methods.

4.1.1 | Recalibration Method 1: Keep the same predictor effects and baseline hazard shape, but change the overall magnitude of the baseline hazard

For this approach, the existing model is applied to participants in the validation study but with the constant term (γ_0 in Equation (5)) re-estimated within the validation data, to give γ_{0New} so that Equation (4) can be rewritten using,

$spline^\#(\ln(t) | \gamma, \mathbf{n})$ in Equation (10) below. Other parameters are as estimated for the existing model and are not altered.

$$\ln(H(t|\mathbf{x}_i)) = spline^\#(\ln(t)|\gamma, \mathbf{n}) + \widehat{PI}_i = \gamma_{0New} + \widehat{\gamma}_1 z_1(t) + \widehat{\gamma}_2 z_2(t) + \cdots + \widehat{PI}_i \quad (10)$$

This is the simplest form of recalibration, allowing the baseline hazard in the existing model to be shifted by a constant factor to better represent the validation population's baseline hazard. This kind of recalibration is useful when the baseline hazard rate differs substantially between derivation and validation samples,¹⁸ but has the same or similar shape over time. In logistic regression the constant term can be related to prevalence of disease and so can be easily altered in new patient populations.^{14,47} In survival data, it represents a part of the baseline hazard function and so could be interpreted as a constant shift increasing or decreasing the hazard at any particular time; that is, with different parameterizations, it could be interpreted as a hazard ratio comparing the hazard rate in the development data to the validation data, under the assumption that covariates effects do not change.

4.1.2 | Recalibration Method 2: Keep the same predictor effects but change the entire baseline hazard (shape and magnitude)

This approach extends Method 1 to allow adjustment for any differences between the derivation and validation populations in terms of the overall magnitude (Method 1) and additionally the shape of the baseline hazard. So, for example, the validation population may have a much earlier and sharper peak in their baseline hazard and a long flattened tail, and so recalibration of the FP model to capture this shape could improve the performance of the model in external populations. To implement this method, while keeping predictor effects fixed at their original values, the baseline hazard shape and magnitude in Equation (4) are completely re-estimated in the validation sample giving a new baseline hazard term, $spline(\ln(t) | \gamma, \mathbf{n})_{New}$ as below;

$$\ln(H(t | \mathbf{x}_i)) = spline(\ln(t) | \gamma, \mathbf{n})_{New} + \widehat{PI}_i \quad (11)$$

When re-estimating the baseline hazard using Method 2, we selected the same number of knots as used in the existing model, and also forced the same knot locations. However, the method is not restricted to this assumption, and if the shape of the baseline was more complex in the validation data allowing different numbers of knots or knot locations could be justified (we provide a sensitivity analysis as online supplementary material).

4.1.3 | Recalibration Method 3: Keep the same baseline hazard shape but change the overall magnitude of the baseline hazard and adjust the prognostic index as a whole by a constant (ϕ)

In this approach, Method 1 is extended to additionally adjust the PI by a constant scalar term, ϕ , which is estimated in the validation data, so that Equation (10) is adjusted as follows,

$$\ln(H(t | \mathbf{x}_i)) = spline^\#(\ln(t) | \gamma, \mathbf{n}) + (\phi \times \widehat{PI}_i) = \gamma_{0New} + \widehat{\gamma}_1 z_1(t) + \widehat{\gamma}_2 z_2(t) + \cdots + (\phi \times \widehat{PI}_i) \quad (12)$$

As such, Method 3 allows the overall magnitude (but not shape) of the baseline hazard to be corrected, and additionally scales the predictor effects as a whole by some constant scalar. For example, if predictor effects are systematically too large (eg, due to uncorrected overfitting during the development of the existing model⁷), then ϕ will be < 1 and predictor effects will be shrunk.

4.1.4 | Recalibration Method 4: Keep the baseline hazard shape and the prognostic index fixed, but change the overall magnitude of the baseline hazard and re-estimate the effect of particular predictors

Poor calibration performance in a validation study may relate specifically to changes in the effect of one or a few predictors. For example, the strength of a single predictor effect such as β_1 may differ in the new population, or the functional form

of the predictor may have been modeled poorly at development (eg, continuous predictor modeled as categorical).⁴⁸ In this case, Method 4 builds on Method 1 by additionally re-estimating β_1 in the validation data to give β_{1New} , keeping all other terms fixed so that Equation (10) is adjusted as follows,

$$\ln(H(t | \mathbf{x}_i)) = spline^\#(\ln(t) | \gamma, \mathbf{n}) + \widehat{PI}_i^\dagger + \beta_{1New}x_1 = \gamma_{0New} + \widehat{\gamma}_1z_1(t) + \widehat{\gamma}_2z_2(t) + \dots + \widehat{PI}_i^\dagger + \beta_{1New}x_1 \quad (13)$$

where \widehat{PI}_i^\dagger now excludes β_1x_1 . Method 4 could be extended to allow additional predictor effects to be re-estimated within the validation data.

Note that if all recalibration methods were combined, this would be equivalent to developing an entirely new model (while retaining the selected predictors and their functional form from the original model¹), and so we do not consider this.

One possible method to identify predictor effects that may benefit from recalibration would be to measure potential heterogeneity in the predictor effects when estimated in different studies using meta-analysis (as shown in online supplementary material). The original model could be fitted within each validation study and, separately for each predictor, the parameter estimates could be pooled in a random-effects meta-analysis using Equation (8). By synthesizing the effect estimates from each validation study, a summary effect estimate can be obtained, and importantly an estimate of between-study heterogeneity (τ^2 as in Equation (8)). If τ^2 is large, or similarly if the prediction interval for a predictor's effect is wide, then this may signal that calibration would be improved if the predictor effect was re-estimated.

4.2 | IPD meta-analysis to summarize performance of recalibration strategies

The recalibration methods described above can be subsequently evaluated and summarized using IPD meta-analysis to identify the best recalibration strategy¹⁶; that is, to find the simplest strategy that leads to the necessary improvement in the model's predictive performance within and across studies. When assessing strategies for model recalibration, the aim is to achieve good model performance on average as indicated by the summary performance statistic $\widehat{\theta}$, as well as small between-study heterogeneity ($\widehat{\tau}^2$) when fitting Equation (8). Importantly the 95% prediction interval (Equation (9)) can also be used to assess the generalizability of the model in new populations similar to those included in the meta-analysis.^{43,44}

4.3 | Comparing and ranking recalibration strategies using multivariate IPD meta-analysis

While we can formally compare the different recalibration strategies separately for each performance measure of interest, we may also wish to make decisions based on multiple criteria allowing for the correlation between such measures.¹⁶ For example, we might wish to assess the probability that a recalibration strategy could ensure a C-statistic above 0.65 and an E/O statistic between 0.95 and 1.05, indicating "good" overall performance.

Snell et al previously proposed a method which can be used to rank such strategies in terms of the joint probability of achieving a particular level of performance in a new setting or population.¹⁶ Extending Equation (8) to a multivariate meta-analysis allows the joint synthesis of all performance statistics of interest by accounting for both the within- and between-study correlations. Say for example we wish to jointly synthesize the E/O and C-statistics, then the multivariate meta-analysis model presented by Snell et al. can be simplified to a bivariate meta-analysis as given by;

$$\begin{aligned} \begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix} &\sim N \left[\begin{pmatrix} \theta_{i1} \\ \theta_{i2} \end{pmatrix}, \mathbf{S}_i \right], & \mathbf{S}_i &= \begin{pmatrix} S_{i1}^2 & S_{i1}S_{i2}\rho_{wi(1,2)} \\ S_{i1}S_{i2}\rho_{wi(1,2)} & S_{i2}^2 \end{pmatrix} \\ \begin{pmatrix} \theta_{i1} \\ \theta_{i2} \end{pmatrix} &\sim N \left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \boldsymbol{\Sigma} \right], & \boldsymbol{\Sigma} &= \begin{pmatrix} \tau_1^2 & \tau_1\tau_2\rho_{B(1,2)} \\ \tau_1\tau_2\rho_{B(1,2)} & \tau_2^2 \end{pmatrix} \end{aligned} \quad (14)$$

Here, θ_{i1} and θ_{i2} represent the true E/O and C-statistic, respectively, for the i th study, \mathbf{S}_i is the within-study variance-covariance matrix for the i th study (assumed known) containing the variances of the estimates (S_{i1}^2 and S_{i2}^2) and

Study	C-statistic	R ² _D	D-statistic
Sweden	0.73 (0.68, 0.77)	0.34 (0.23, 0.44)	1.46 (1.12, 1.81)
Lille	0.64 (0.59, 0.69)	0.13 (0.05, 0.23)	0.79 (0.48, 1.11)
Nijmegen	0.66 (0.59, 0.72)	0.18 (0.06, 0.30)	0.94 (0.53, 1.35)
St Cloud	0.65 (0.60, 0.70)	0.15 (0.07, 0.24)	0.86 (0.58, 1.16)
Switzerland	0.72 (0.67, 0.76)	0.30 (0.21, 0.40)	1.36 (1.06, 1.66)
Denmark 1	0.61 (0.57, 0.65)	0.11 (0.04, 0.19)	0.71 (0.42, 1.00)
Denmark 2	0.67 (0.62, 0.71)	0.21 (0.12, 0.32)	1.06 (0.75, 1.40)
Meta-analysis			
Summary estimate (95% CI)	0.67 (0.63, 0.70)	0.20 (0.14, 0.28)	1.02 (0.80, 1.24)
95% PI	(0.56, 0.76)	(0.06, 0.49)	(0.34, 1.70)
τ^2	0.03	0.22	0.06
I ² (%)	67.75	67.2	68.28

Abbreviations: CI, confidence interval; PI, prediction interval.

TABLE 3 Performance of the breast cancer model when applied to the external validation studies in terms of the C-statistic, D-statistic, and R²_D

their covariances in the off-diagonal (for example, $S_{i1}S_{i2}\rho_{Wi(1,2)}$ is the within-study covariance for the E/O and C-statistics, where $\rho_{Wi(1,2)}$ is their within-study correlation caused by estimates derived from the same patients), μ_1 and μ_2 represent the means for the E/O and C-statistics, and Σ is the between-study variance-covariance matrix containing the between-study variances (τ_1^2 and τ_2^2) and their between-study covariances in the off-diagonal as before (where $\rho_{B(1,2)}$ is the between-study correlation induced by differences in study populations and settings).

Joint inferences across multiple performance measures can then be made by accounting for their correlation, using the multivariate approach as in Equation (14). One can obtain joint probabilistic inferences if we assume the multivariate t-distribution (with $k - 2$ degrees of freedom) is an approximate posterior distribution (that is, we assume it is obtained from a Bayesian analysis with uninformative priors, and give it means, variances and covariances obtained from REML estimation of Equation (14)) as described by Snell et al.¹⁶ For example, one can sample from this assumed joint distribution many times and calculate the proportion of times performance lies within given parameters (say the C-statistic will be above 0.65 and E/O will be between 0.95 and 1.05), to derive the joint probability of such performance in a new population. This probability can then be used to rank recalibration strategies, with the highest probability ranked best.

5 | APPLICATION TO THE BREAST CANCER EXAMPLE

The process of external validation and recalibration using IPD meta-analysis is now applied to the breast cancer example introduced in Section 2. Recall that—for the sake of illustration—one of the studies (Rotterdam) was used to create an existing model, and thus there were seven studies available for external validation (see Table 1). First, we describe the performance of the existing model upon external validation, and then work through the model recalibration methods and compare their external validation performance.

5.1 | IPD meta-analysis of external validation performance of existing model

Table 3 gives the summary results from a random-effects meta-analysis (Equation (8)) of the model's performance in the seven validation studies, in terms of discrimination (with appropriate transformation of performance statistics as required see Section 3.3).

Discrimination performance was moderate, with observed C-statistics ranging from 0.61 to 0.73 across validation studies. An approximate 95% prediction interval suggests the actual C-statistic for the model could range from 0.56 to 0.76 in a single population/setting, most likely due to differences in case-mix distribution across studies (as shown in online supplementary material). The summary R²_D suggests the model explains 20% variation on average, with a 95% prediction interval of 0.06 to 0.49.

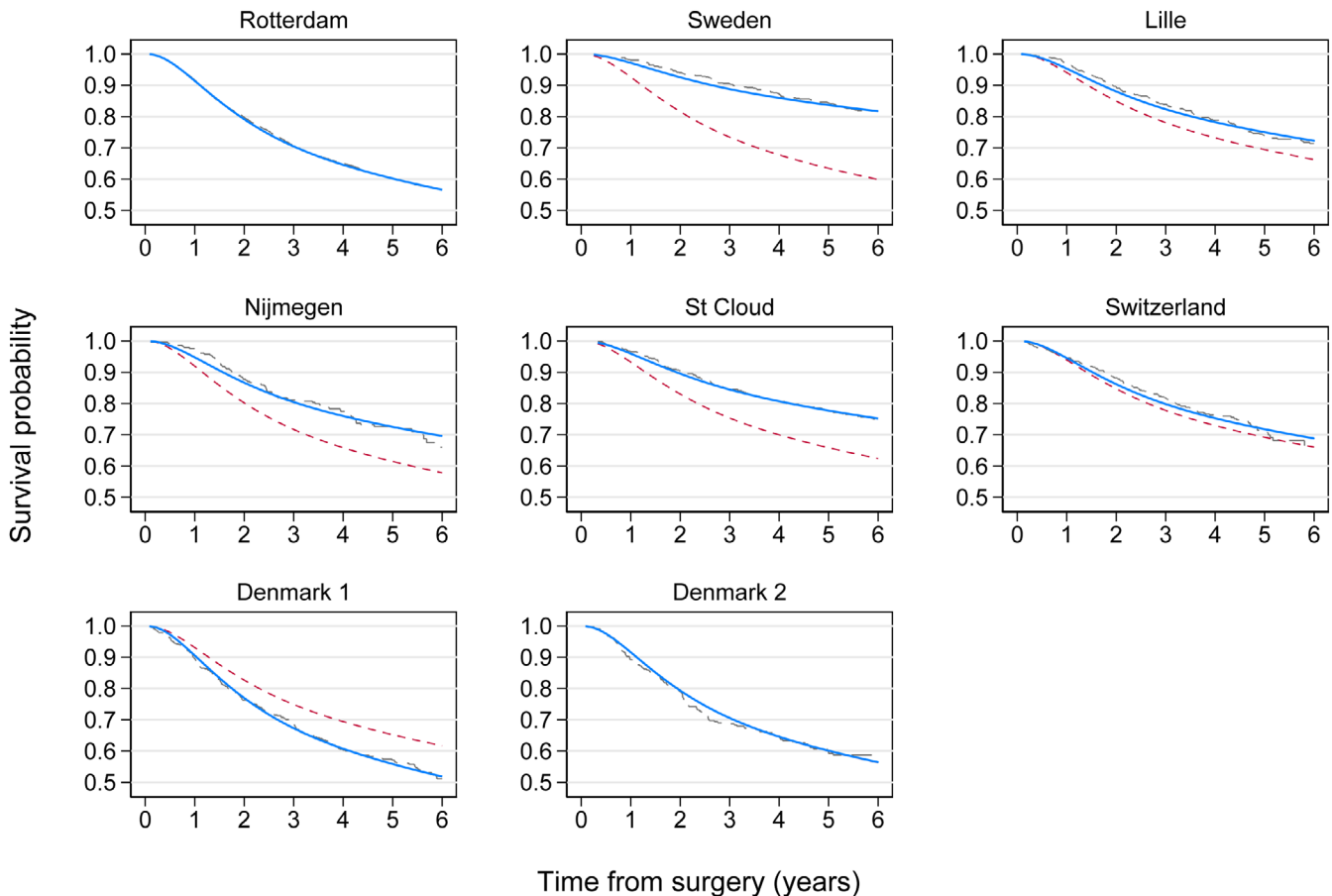


FIGURE 1 Calibration plot showing performance of the existing prediction model (at the population-level) before and after recalibration via Method 1 in the seven validation studies. Dashed grey lines = observed Kaplan-Meier curve. Solid blue lines = prediction after recalibration using Method 1. Short-dashed red lines = original model predictions before recalibration [Color figure can be viewed at wileyonlinelibrary.com]

In terms of overall calibration of predicted survival probabilities in each dataset, Figure 1 shows good observed calibration in one study (“Denmark 2”), but systematic under prediction in five studies (most notably in Sweden) and over prediction in the “Denmark 1” study. Therefore, there is considerable between-study heterogeneity in the calibration performance of the model. These differences are again likely due to differences in case-mix distribution, and baseline hazard across studies, which is explored further within the online supplementary material.

Figure 2 shows the summary calibration performance based on univariate random-effects meta-analysis of E/O at various time points since surgery. It is clear that, on average across studies, the model’s under-prediction of outcome risk increases over the first 4 years, with E/O about 0.9 at 4 years onward. Furthermore, the between-study heterogeneity in calibration performance increases over time as reflected by a widening prediction interval for E/O, for example of 0.7 to 1.3 by 4 years.

5.2 | IPD meta-analysis of external validation performance after recalibration

The previous section showed that the breast cancer model had large heterogeneity in predictive performance across the seven external validation studies. Average performance was also shown to be suboptimal, particularly at increased years from surgery, and so the model is unlikely to be robust in all populations of potential application. Therefore, we now consider methods for recalibration of the existing model to improve model performance, and aim to identify the best recalibration strategy using meta-analysis methods.

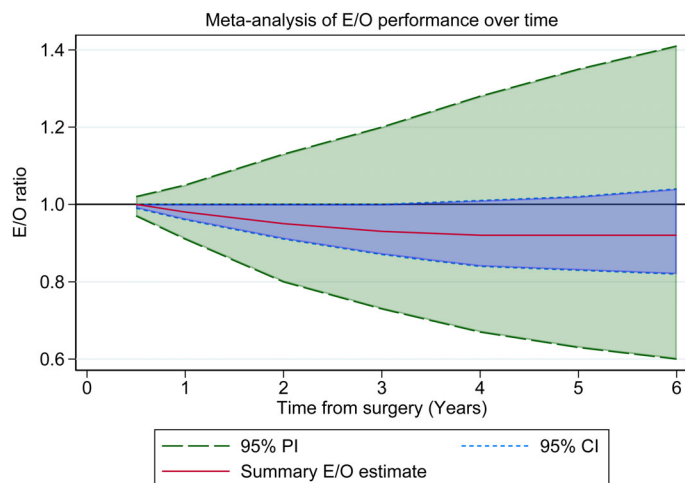


FIGURE 2 Calibration performance of the existing model over time based on univariate random-effects meta-analysis of the E/O statistic (at years 1 to 6 separately), showing summary (average) E/O in the seven external validation studies, alongside 95% confidence (CI) and prediction intervals (PI) [Color figure can be viewed at wileyonlinelibrary.com]

5.2.1 | Recalibration Method 1 calibration performance

Calibration performance was substantially improved through use of recalibration Method 1, where the intercept of the baseline cumulative hazard's spline function was re-estimated for each study (see Figure 1). The updated model's predictions using Method 1 are, overall across the entire set of individuals, extremely close to the observed Kaplan-Meier curve in each study (Figure 1, solid lines), in contrast to the (sometimes large) miscalibration for predictions from the existing model before recalibration (Figure 1, short-dashed lines).

A forest plot for the E/O statistic measured at 3 years post-surgery using all methods is given in Figure 3, showing poor calibration of the existing model, and stark improvement after applying Method 1. For example, the E/O statistic at 3 years for the Sweden validation study is 0.81, which is equivalent to a patient's predicted probability of survival being 19% lower than the truth. However after recalibration of the overall magnitude of the baseline hazard using Method 1, this under prediction is reduced to just 2%, as E/O is 0.98 (see Figure 3).

5.2.2 | Recalibration Method 2 calibration performance

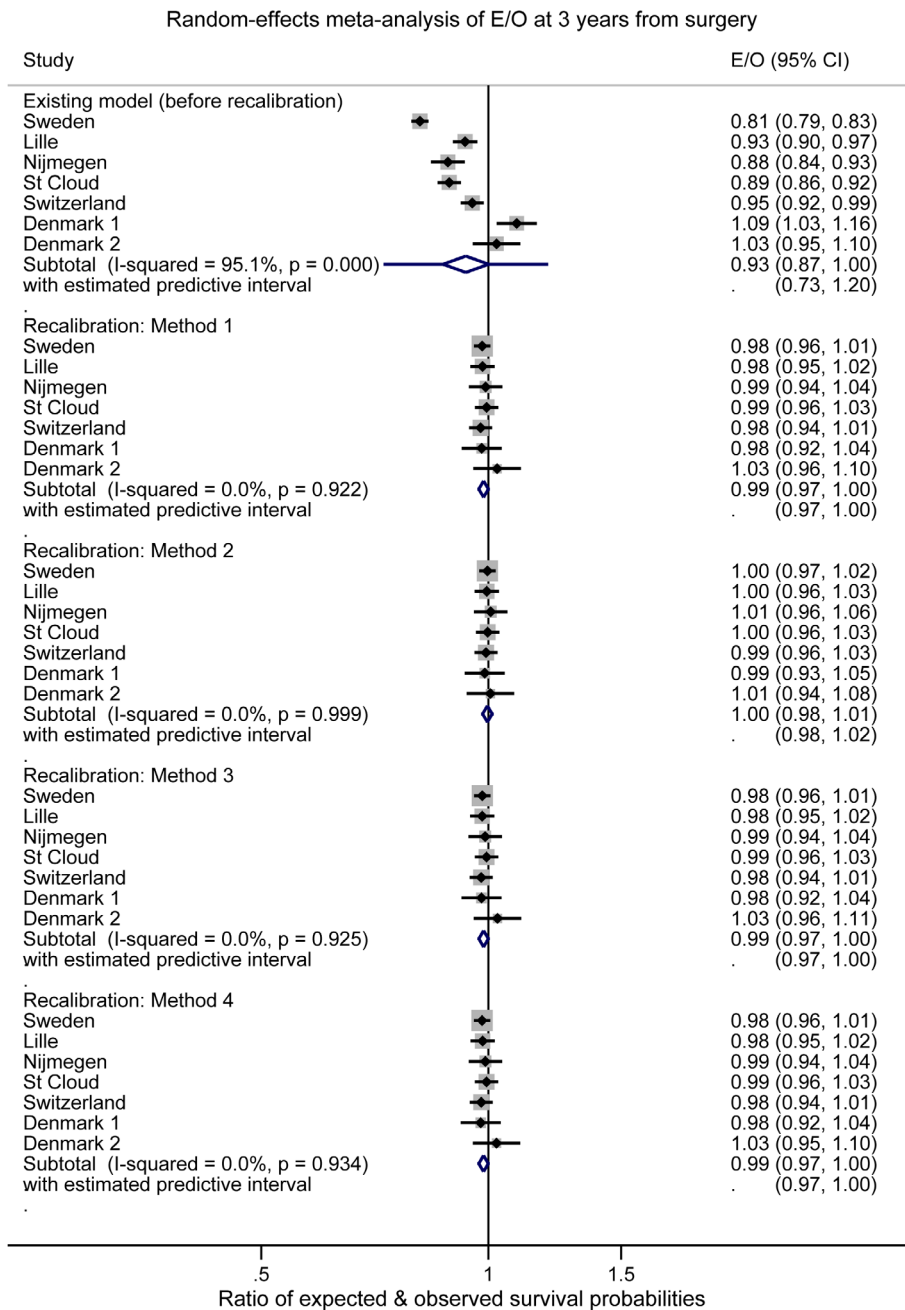
Recalibration Method 2 builds on Method 1 by additionally allowing the shape of the baseline cumulative hazard to be re-estimated, which leads to further improvements compared to the existing model's performance. For example, after recalibration of the magnitude of the baseline hazard using Method 1 the summary E/O statistic was 0.990 (0.981, 0.999) at 1 year post-surgery equivalent to under prediction of 1%, with heterogeneity across validation studies performance ($I^2 = 48\%$). However, after additional recalibration of the whole baseline cumulative hazard (magnitude and shape) using Method 2, the models' predictions are (at the population-level) perfect with a summary E/O = 0.998 (0.992, 1.005) and no heterogeneity ($\tau^2=0$, $I^2 = 0\%$; see Table 4).

The prediction intervals calculated from the random effects meta-analysis were much narrower after Methods 1 (95% PI: 0.966, 1.016) and 2 (95% PI: 0.989, 1.007) compared with the existing model (95% PI: 0.908, 1.052), indicating very little potential variation in calibration performance in new populations where recalibration can be implemented. There was evidence of heterogeneity in calibration performance at all time points post-surgery ($\tau^2 = [0.001, 0.024]$), observed for the existing model, and this was reduced to zero heterogeneity after use of recalibration Method 2 in particular (see Table 4).

5.2.3 | Recalibration Method 3 calibration performance

Method 3 builds on Method 1 as described in Section 4, and thus calibration performance remained superior to the existing model. Notably Method 3 appeared to negate the additional improvement made by Method 2, pulling performance back in line with that seen under Method 1, such that there is still some miscalibration in E/O statistics (see Figure 3).

FIGURE 3 Random-effects meta-analysis of calibration performance (E/O at 3 years post-surgery) of the existing model in all validation studies split by recalibration method. Top panel shows performance of the original existing model in the validation studies (ie, before any recalibration) [Color figure can be viewed at wileyonlinelibrary.com]



5.2.4 | Recalibration Method 4 calibration performance

Recalibration Method 4 is the same as Method 1 but also allows re-estimation of a particularly heterogeneous predictor effect (see Section 4). To identify such a predictor in the existing model, we applied a separate random-effects meta-analysis of each predictor’s effects in the external validation dataset. That is, we re-fitted the existing model in each validation study separately and all estimates and standard errors of the regression coefficients were stored. Next a random-effects meta-analysis of log hazard ratio estimates was performed for each predictor in turn to quantify the heterogeneity in the predictor’s effect across the validation studies using Equation (8). This process identified the tumor size predictor had a notably heterogeneous effect, with a between-study variance of $\hat{\tau}^2 = 0.242$ and $I^2 = 62.5\%$ (see Figure 4), and it was therefore re-estimated while holding other predictor effects fixed (ie, values kept the same as in the original model).

As Method 4 builds on Method 1, the magnitude of the baseline cumulative hazard is also updated alongside the re-estimated predictor effect for tumor size. The revised model has calibration performance similar to that for Method 3, with increased miscalibration in the Denmark 2 study compared to Method 2 (see Figure 3).

E/O (1 year)	Summary effect (CI)	95% PI	τ^2	I ² (%)
Existing model	0.977 (0.957, 0.999)	(0.908, 1.052)	0.001	89.695
Method 1	0.990 (0.981, 0.999)	(0.966, 1.016)	0	47.726
Method 2	0.998 (0.992, 1.005)	(0.989, 1.007)	0	0
Method 3	0.990 (0.981, 0.999)	(0.966, 1.014)	0	45.394
Method 4	0.990 (0.981, 0.999)	(0.966, 1.015)	0	47.380
E/O (3 year)	Summary effect (CI)	95% PI	τ^2	I ² (%)
Existing model	0.934 (0.871, 1.001)	(0.726, 1.200)	0.008	95.140
Method 1	0.986 (0.971, 0.999)	(0.967, 1.005)	0	0
Method 2	0.997 (0.983, 1.012)	(0.978, 1.016)	0	0
Method 3	0.986 (0.971, 1.000)	(0.967, 1.005)	0	0
Method 4	0.985 (0.971, 0.999)	(0.967, 1.004)	0	0
E/O (6 year)	Summary effect (CI)	95% PI	τ^2	I ² (%)
Existing model	0.919 (0.816, 1.036)	(0.598, 1.414)	0.024	95.817
Method 1	1.005 (0.983, 1.027)	(0.976, 1.035)	0	0
Method 2	0.998 (0.976, 1.019)	(0.969, 1.027)	0	0
Method 3	1.006 (0.984, 1.029)	(0.977, 1.036)	0	0
Method 4	1.005 (0.983, 1.027)	(0.976, 1.034)	0	0
C-statistic	Summary effect (95% CI)	95% PI	τ^2	I ² (%)
Existing model	0.667 (0.634, 0.699)	(0.563, 0.760)	0.001	68.60
Method 1	0.667 (0.634, 0.699)	(0.563, 0.760)	0.001	68.60
Method 2	0.667 (0.634, 0.699)	(0.563, 0.760)	0.001	68.60
Method 3	0.667 (0.634, 0.699)	(0.563, 0.760)	0.001	68.60
Method 4	0.667 (0.635, 0.699)	(0.570, 0.760)	0.001	68.20

Abbreviations: CI, confidence interval; PI, prediction interval.

TABLE 4 Comparison of univariate random effects meta-analysis results for each recalibration method in terms of both calibration (E/O statistic) and discrimination performance (C-statistic)

5.2.5 | Discrimination performance after recalibration

Table 4 also shows the discrimination performance of the model after applying each recalibration method. It is clear that discrimination performance was unaffected by recalibration Methods 1 to 3, as expected because these methods do not alter the relative ranking of the PI.⁹ Discrimination is strongly influenced by case-mix in the validation cohort, and Method 4 directly adjusts the PI by re-estimating an individual predictor effect, but in this example we considered adjustment of only one predictor (tumor size). Harrell's C-statistic varied by 0.01 in the 95% CI and PI, but did not alter in terms of the summary performance estimate even after recalibration with Method 4 (see Table 4 and online supplementary material for additional results). Further, there was no change in the heterogeneity in discrimination performance across validation studies after any recalibration method.

5.3 | Overall ranking of recalibration strategies

We now fit a multivariate meta-analysis jointly synthesizing the E/O and C-statistics to compare the recalibration strategies as described in Section 4.3. Post-estimation, the joint probability of achieving a C-statistic > 0.65 and an E/O statistic between 0.95 and 1.05 (at 1 year) was calculated for each recalibration method in turn by drawing 10 000 samples from a multivariate t-distribution. The ranking of these joint probabilities is presented in Figure 5, where all recalibration methods showed high probabilities, close to one, whereas fitting the existing model without recalibration had a much lower predicted probability of 0.713 of achieving the specified performance criteria. The highest predicted probability was for

FIGURE 4 Random-effects meta-analysis of tumour size regression coefficient [ln(HR)] from each validation study [Color figure can be viewed at wileyonlinelibrary.com]

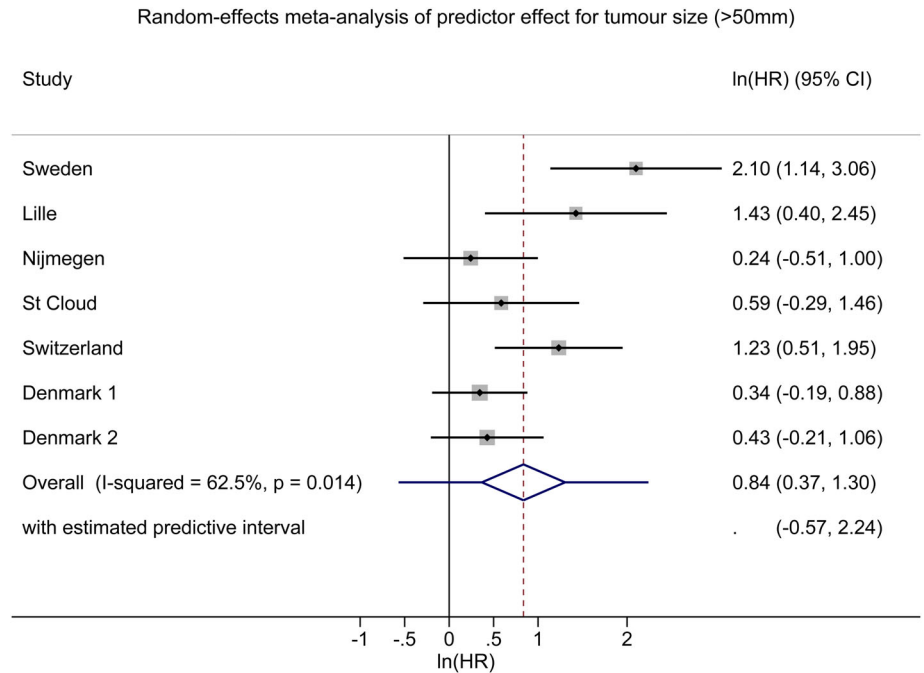
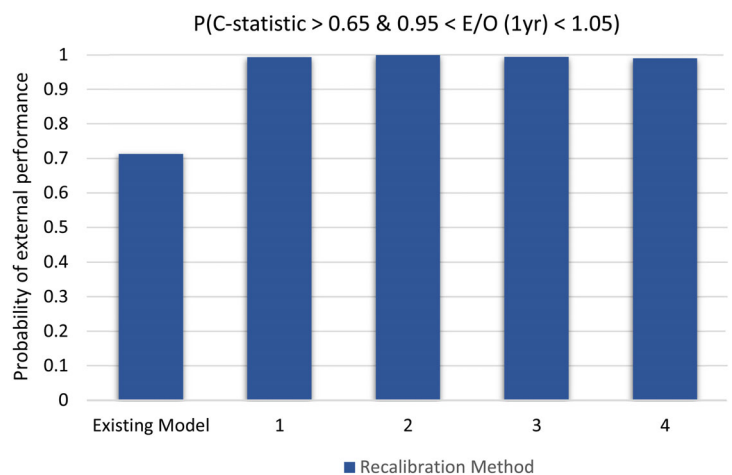


FIGURE 5 Ranking of recalibration methods by the expected probability of good performance in a new population, defined as a C-statistic > 0.65 & 0.95 < E/O < 1.05. Based on 10,000 draws from a multivariate t-distribution using parameters from a multivariate meta-analysis of C-statistic and E/O performance across the validation studies available in the breast cancer dataset [Color figure can be viewed at wileyonlinelibrary.com]



Method 2 with a probability of almost 1. Crucially, however, the simpler Method 1 gives practically the same improvement, and so is the one we recommend for applying the model in practice (as it only requires the magnitude of the baseline hazard to be re-estimated in each country of application).

5.4 | Further evaluations of calibration at the individual level using pseudo-values

Throughout the manuscript, we have focused on population level calibration of the model; however, it is also important that the model performs well across the risk spectrum, that is, the model must calibrate well for patients at high and low risk, not only on average. To assess calibration across the risk spectrum, we examined time-specific calibration plots using both groupings and individual event probabilities to discern more subtle miscalibrations, using Stata packages “pmcalplot” and “stcoxcal”.^{34,49} We used pseudo-values to derive the observed event probabilities for individuals at a specific timepoint, though alternative approaches have recently been proposed.^{34,50} Examining these plots showed a similar pattern of miscalibration across the risk spectrum to that seen at the population level in Figure 1, with recalibration improving performance substantially (further details provided in online supplementary material).

6 | DISCUSSION

We examined the use of recalibration methods for improving the performance of an existing survival model at external validation, when multiple validation studies are available. In the applied example, when making predictions using the existing model directly, discrimination performance was moderate, but calibration was often poor with substantial over and under prediction across studies, as is common in new patient populations, due to changes in case-mix variation and baseline hazard rate (as explored in the online supplementary material).^{8,9,51,52} Various recalibration strategies were evaluated to address this using IPD meta-analysis methodology. The findings showed that recalibration of the magnitude of the baseline hazard gave large improvements in the calibration performance of the model. The method is simple, as it only requires re-estimation of the intercept in the validation study. Similar improvements have been shown through recalibration of the intercept in logistic prediction model examples.^{9,12,47} The other methods investigated (recalibration of the baseline hazard shape, prognostic index, or individual predictors), showed little additional improvement on adjustment of the intercept alone. Discrimination performance was unaltered by such methods as adjustment of baseline risk or hazard does not materially alter the relative rankings of the prognostic index.⁹

The premise of our work lies in identifying which recalibration method is most appropriate in new settings (ie, greatest improvement in performance with simplest model updating method), as such other applications may find another recalibration method more beneficial than that identified for this case study. Previous work has proposed the use of a closed testing procedure to identify the best recalibration method to improve performance while avoiding overfitting for logistic prediction models⁴⁶; such an approach could also be evaluated for time-to-event models in further work. However, here we instead used a novel ranking system based on the joint probability of achieving “good” summary performance (derived after a multivariate meta-analysis) in terms of both calibration and discrimination of the model in the external population. This adds to the proposed methodology by allowing for the multivariate nature of model performance, where both discrimination and calibration are equally important for a model to be useful. In our case study, the rankings reinforced conclusions drawn from the IPD meta-analysis results, highlighting that simpler methods of recalibration could achieve good model performance without needing to throw away information from the existing model, and prevents researchers from simply developing an entirely new model.

One limitation of recalibration methods is the level of information required about the existing model for recalibration to be possible. Here FP models were used because they naturally allow flexibility to investigate the recalibration of various parts of the model and at various time-points; however in practice many published prognostic models are Cox models. This hinders the use of even simple recalibration methods because it is very rare that an estimate of the baseline hazard (or baseline survival) is published,⁵³ and this is essential to individualized prediction and recalibration.^{1,30} Despite this, Royston and Altman have laid out detailed methods for external validation of a Cox model elsewhere.¹ For example, even if authors have only presented $S_0(t)$ at a few specific time points, it is possible to make predictions and gain some assessment of calibration performance (though this is rarely reported).¹ The strategies proposed here could be applied to Cox model examples with the exception of Method 1, instead recalibration would need to begin with Method 2 allowing complete recalibration of the baseline hazard in the validation sample.

In the motivating example, IPD was available from seven validation studies with clustering based on countries. However, the methods discussed here are equally applicable to any clustering variable, for example when using e-health databases which often contain clustering by center or region.¹¹ In the example, recalibration of the intercept to specific clusters provided the required improvement in model performance (while re-estimating fewest parameters from the original model). Practically, this means that for every new population a new intercept must be calculated for the existing model (to achieve best performance), and so the relevant data must be available and large enough to estimate the intercept reliably.⁵⁴⁻⁵⁶ Previous research has also found that models using study-specific intercepts may give greatest performance; for example, Wynants et al recommended the use of conditional model predictions in clustered data as they perform well at both the population and center level in terms of both calibration and discrimination.^{14,16,57} Our four methods are not exhaustive, and other options could be considered. In particular, temporal recalibration may be important in fields of rapid medical advancement, allowing recalibration of the baseline hazard using a specific recent time window.^{58,59}

Importantly recalibration in any form could strictly be viewed as a new model, which itself requires external validation in new data, in potentially different patient populations, new settings, different geographical locations or at different times.^{1,8,32,60} This could mean a never-ending cycle of recalibration followed by external validation, with then further recalibration, and so on. Further research of this issue is needed, but is likely dependent on the purpose of the model in practice. For example, recalibration methods such as those discussed here may work best when the aim is to tailor a model

to a specific practice in which the model will then be used routinely, making an emphasis on further external validation in new settings potentially less important.

In conclusion, we have presented methods for the recalibration of FP prognostic models, and proposed methodology using IPD meta-analysis to examine the ability of the various recalibration methods, to improve an existing prognostic model's performance in external validation datasets. Summary measures including summary discrimination and calibration statistics, heterogeneity in performance statistics, prediction intervals, and joint probability ranking can inform selection of the best recalibration strategy for particular external populations. In particular this methodology can be used to decide on the least aggressive recalibration strategy to achieve acceptable external model performance without discarding existing model information. Stata code is available to perform the proposed methodology.

ACKNOWLEDGEMENTS

We thank the Editor, Associate Editor, and Reviewers for constructive comments that improved this article upon revision.

DATA AVAILABILITY STATEMENT

The IPD used to provide examples is not routinely available for sharing.

ORCID

Joie Ensor  <https://orcid.org/0000-0001-7481-0282>

Kym I. E. Snell  <https://orcid.org/0000-0001-9373-6591>

Thomas P. A. Debray  <https://orcid.org/0000-0002-1790-2719>

Richard D. Riley  <https://orcid.org/0000-0001-8699-0735>

REFERENCES

1. Royston P, Altman DG. External validation of a Cox prognostic model: principles and methods. *BMC Med Res Methodol.* 2013;13:33.
2. Steyerberg EW, Moons KG, van der Windt DA, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med.* 2013;10(2):e1001381.
3. Wyatt JC, Altman DG. Commentary: prognostic models: clinically useful or quickly forgotten? *BMJ.* 1995;311(7019):1539-1541.
4. Debray TPA, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KGM. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol.* 2015;68(3):279-289.
5. Damen JAAG, Hooft L, Schuit E, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ.* 2016;353.
6. Steyerberg EW, Harrell FE Jr. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol.* 2016;69:245-247.
7. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating.* 1st ed. New York, NY: Springer-Verlag; 2009.
8. Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. *BMJ.* 2009;338:b605.
9. Toll DB, Janssen KJ, Vergouwe Y, Moons KG. Validation, updating and impact of clinical prediction rules: a review. *J Clin Epidemiol.* 2008;61(11):1085-1094.
10. Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KG. Internal and external validation of predictive models: a simulation study of bias and precision in small samples. *J Clin Epidemiol.* 2003;56(5):441-447.
11. Riley RD, Ensor J, Snell KIE, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *BMJ.* 2016;353.
12. Janssen KJ, Moons KG, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. *J Clin Epidemiol.* 2008;61(1):76-86.
13. van Houwelingen HC. Validation, calibration, revision and combination of prognostic survival models. *Stat Med.* 2000;19(24):3401-3415.
14. Debray TP, Moons KG, Ahmed I, Koffijberg H, Riley RD. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. *Stat Med.* 2013;32(18):3158-3180.
15. Van Houwelingen HC, Thorogood J. Construction, validation and updating of a prognostic model for kidney graft survival. *Stat Med.* 1995;14(18):1999-2008.
16. Snell KIE, Hua H, Debray TPA, et al. Multivariate meta-analysis of individual participant data helped externally validate the performance and implementation of a prediction model. *J Clin Epidemiol.* 2016;69:40-50.
17. Steyerberg EW, Borsboom GJ, van Houwelingen HC, Eijkemans MJ, Habbema JD. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Stat Med.* 2004;23(16):2567-2586.
18. Moons KG, Kengne AP, Grobbee DE, et al. Risk prediction models: II. external validation, model updating, and impact assessment. *Heart.* 2012;98(9):691-698.
19. van Klaveren D, Steyerberg E, Perel P, Vergouwe Y. Assessing discriminative ability of risk models in clustered data. *BMC Med Res Methodol.* 2014;14(1):5.

20. Royston P, Parmar MK. 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(15):2175-2197.
21. Royston P. Flexible parametric alternatives to the Cox model, and more. *Stata J*. 2001;1(1):1-28.
22. Royston P. Flexible parametric alternatives to the Cox model: update. *Stata J*. 2004;4(1):98-101.
23. Look MP, van Putten WL, Duffy MJ, et al. Pooled analysis of prognostic impact of urokinase-type plasminogen activator and its inhibitor PAI-1 in 8377 breast cancer patients. *J Natl Cancer Inst*. 2002;94(2):116-128.
24. Royston P, Parmar MK, Sylvester R. Construction and validation of a prognostic model across several studies, with an application in superficial bladder cancer. *Stat Med*. 2004;23:907-926.
25. 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.
26. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol*. 2016;74:167-176.
27. Vergouwe Y, Steyerberg EW, Eijkemans MJ, Habbema JD. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *J Clin Epidemiol*. 2005;58(5):475-483.
28. Steyerberg EW. Validation in prediction research: the waste by data-splitting. *J Clin Epidemiol*. 2018;103:131-133.
29. Valsecchi MG, Silvestri D, Sasieni P. Evaluation of long-term survival: use of diagnostics and robust estimators with Cox's proportional hazards model. *Stat Med*. 1996;15(24):2763-2780.
30. Royston P, Lambert PC. Flexible parametric survival analysis using stata: beyond the Cox model; 2006.
31. Royston P. *Estimating a Smooth Baseline Hazard Function for the Cox Model*. London, UK: Department of Statistical Science, University College London; 2011.
32. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMC Med*. 2015;13:1.
33. 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.
34. 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.
35. Harrell FE Jr, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *Jama*. 1982;247(18):2543-2546.
36. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. *Stat Med*. 2004;23(5):723-748.
37. Debray TPA, Damen JAAG, Snell KIE, et al. A guide to systematic review and meta-analysis of prediction model performance. *BMJ*. 2017;356.
38. Debray TP, Damen JA, 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(9):2768-2786.
39. Snell KIE, Ensor J, Debray TPA, Moons KGM, Riley RD. Meta-analysis of prediction model performance across multiple studies: Which scale helps ensure between-study normality for the C-statistic and calibration measures? *Stat Methods Med Res*. 2018;27(11):3505-3522.
40. Sidik K, Jonkman JN. Robust variance estimation for random effects meta-analysis. *Comput Stat Data Anal*. 2006;50(12):3681-3701.
41. Hedges LV, Tipton E, Johnson MC. Robust variance estimation in meta-regression with dependent effect size estimates. *Res Synth Methods*. 2010;1(1):39-65.
42. Hartung J, Knapp G. A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Stat Med*. 2001;20(24):3875-3889.
43. Higgins JP, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *J R Stat Soc Ser A Stat Soc*. 2009;172(1):137-159.
44. Riley RD, Higgins JP, Deeks JJ. Interpretation of random effects meta-analyses. *BMJ*. 2011;d549.342
45. StataCorp. *Stata Statistical Software: Release*. Vol 14. StataCorp LP: College Station, TX; 2015.
46. Vergouwe Y, Nieboer D, Oostenbrink R, et al. A closed testing procedure to select an appropriate method for updating prediction models. *Stat Med*. 2017;36(28):4529-4539.
47. Janssen KJ, Vergouwe Y, Kalkman CJ, Grobbee DE, Moons KG. A simple method to adjust clinical prediction models to local circumstances. *Can J Anaesth = Journal canadien d'anesthesie*. 2009;56(3):194-201.
48. Collins GS, Ogundimu EO, Cook JA, Manach YL, Altman DG. Quantifying the impact of different approaches for handling continuous predictors on the performance of a prognostic model. *Stat Med*. 2016;35(23):4124-4135.
49. PMCALPLOT. *Stata module to produce calibration plot of prediction model performance* [computer program]. Version S458486: Boston College Department of Economics; 2018.
50. 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.
51. Moons KGM, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *BMJ*. 2009;338.
52. 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.
53. Ng R, Kornas K, Sutradhar R, Wodchis WP, Rosella LC. The current application of the Royston-Parmar model for prognostic modeling in health research: a scoping review. *Diagn Progn Res*. 2018;2(1):4.

54. 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.
55. PMSAMPSIZE. *Stata module to calculate the minimum sample size required for developing a multivariable prediction model* [computer program]. Version S458569: Boston College Department of Economics; 2018.
56. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ*. 2020;368:m441.
57. Wynants L, Vergouwe Y, Van Huffel S, Timmerman D, Van Calster B. Does ignoring clustering in multicenter data influence the performance of prediction models? a simulation study. *Stat Methods Med Res*. 2018;0962280216668555;27(6):1723-1736.
58. Booth S, Riley RD, Ensor J, Lambert PC, Rutherford MJ. Temporal recalibration for improving prognostic model development and risk predictions in settings where survival is improving over time. *Int J Epidemiol*. 2020;49(4):1316-1325.
59. Hickey GL, Grant SW, Murphy GJ, et al. Dynamic trends in cardiac surgery: why the logistic EuroSCORE is no longer suitable for contemporary cardiac surgery and implications for future risk models. *Eur J Cardio-thorac Surg Offic J Europ Assoc Cardio-thorac Surg*. 2013;43(6):1146-1152.
60. Moons KM, 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.

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: Ensor J, Snell KIE, Debray TPA, et al. Individual participant data meta-analysis for external validation, recalibration, and updating of a flexible parametric prognostic model. *Statistics in Medicine*. 2021;1-19. <https://doi.org/10.1002/sim.8959>