Statistics in Medicine

WILEY

Statistics in Medicine

RESEARCH ARTICLE OPEN ACCESS

Instability of the AUROC of Clinical Prediction Models

Florian D. van Leeuwen[1] | Ewout W. Steyerberg[1] | David van Klaveren[2,3] | Ben Wessler[3] | David M. Kent[3] | Erik W. van Zwet[1]

[1]Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands | [2]Department of Public Health, Erasmus University Medical Center, Rotterdam, Netherlands | [3]Predictive Analytics and Comparative Effectiveness Center, Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Boston, Massachusetts, USA

Correspondence: Erik W. van Zwet (e.w.van_zwet@lumc.nl)

Received: 12 June 2024 | **Revised:** 4 December 2024 | **Accepted:** 18 January 2025

Funding: The authors received no specific funding for this work.

Keywords: clinical prediction models | CPM | empirical Bayes | heterogeneity | meta-analysis

ABSTRACT

Background: External validations are essential to assess the performance of a clinical prediction model (CPM) before deployment. Apart from model misspecification, also differences in patient population, the standard of care, predictor definitions, and other factors influence a model's discriminative ability, as commonly quantified by the AUC (or c-statistic). We aimed to quantify the variation in AUCs across sets of external validation studies and propose ways to adjust expectations of a model's performance in a new setting.

Methods: The Tufts-PACE CPM Registry holds a collection of CPMs for prognosis in cardiovascular disease. We analyzed the AUC estimates of 469 CPMs with at least one external validation. Combined, these CPMs had a total of 1603 external validations reported in the literature. For each CPM and its associated set of validation studies, we performed a random-effects meta-analysis to estimate the between-study standard deviation τ among the AUCs. Since the majority of these meta-analyses have only a handful of validations, this leads to very poor estimates of τ. So, instead of focusing on a single CPM, we estimated a log-normal distribution of τ across all 469 CPMs. We then used this distribution as an empirical prior. We used cross-validation to compare this empirical Bayesian approach with frequentist fixed and random-effects meta-analyses.

Results: The 469 CPMs included in our study had a median of 2 external validations with an IQR of [1–3]. The estimated distribution of τ had a mean of 0.055 and a standard deviation of 0.015. If $\tau = 0.05$, then the 95% prediction interval for the AUC in a new setting has a width of at least $+/-$ 0.1, no matter how many validations have been done. When there are fewer than 5 validations, which is typically the case, the usual frequentist methods grossly underestimate the uncertainty about the AUC in a new setting. Accounting for τ in a Bayesian approach achieved near nominal coverage.

Conclusion: Due to large heterogeneity among the validated AUC values of a CPM, there is great irreducible uncertainty in predicting the AUC in a new setting. This uncertainty is underestimated by existing methods. The proposed empirical Bayes approach addresses this problem which merits wide application in judging the validity of prediction models.

1 | Introduction

Clinical prediction models may provide care-givers and patients with quantitative estimates of risk and prognosis, which can inform clinical decision-making [1]. Before deployment of a newly developed CPM, it is crucial that its performance is carefully and repeatedly validated. If the performance of a CPM is assessed with the same data that was used to develop it, then it is important to account for some degree of overfitting. Common approaches for internal validation include cross-validation and bootstrap resampling [2]. Beyond internal validation, external validation refers to the assessment of performance in a new setting (a plausibly related population [3].) While internal validation quantifies reproducibility, external validation assesses the generalizability of CPMs [3–5].

Here we study the Tufts-PACE CPM Registry which is a unique, carefully curated set of external validations of CPMs in the field of cardiovascular medicine [6]. We focus on discrimination as a key aspect of performance in external validation studies, commonly quantified in terms of the Area Under the Receiver Operating Curve (AUROC, AUC) or the c-statistic. Large variation among the validations of the same CPM would be problematic because it implies that there is great uncertainty about the AUC when we want to deploy that CPM in a new setting. Therefore, our main goal is to assess the amount of heterogeneity among the validations of a CPM and propose ways to adjust expectations of a model's performance in a new setting. Moreover, as we will demonstrate, the usual frequentist methods severely underestimate this uncertainty when there are few (fewer than 5) validations.

The paper is organized as follows. In the next section, we introduce our data set, provide the relevant background information, and introduce the problem with two examples. In Section 3, we describe our statistical model and propose an empirical Bayes approach for predicting the AUC in a new setting. In Section 4, we present our results. We provide an estimate of the heterogeneity and use cross-validation to compare our empirical Bayes approach to the usual (frequentist) methods. We end the paper with a brief discussion.

2 | Background and Problem Statement

According to the flow chart in the Appendix A, we included 469 different CPMs that have at least 2 external validations in the Tufts-PACE CPM Registry. Each of these CPMs can be used to calculate the risk for a binary cardiovascular outcome of future patients. As an introduction to our data set, we plot the external validation AUCs (or c-statistics) versus the associated development AUCs (Figure 1). We added a regression curve (a natural spline with 3 degrees of freedom) and note that the AUCs at development were systematically higher than AUCs at validation. This may be due to optimism that is not always fully accounted for at internal validation. Moreover, validation populations may be more or less heterogeneous than the development population. We also note a substantial variability across validation AUCs.

As an example, we consider the CRUSADE prediction model for patients with angina pectoris [7]. This model was externally validated one year after development [8]. The external validation resulted in an estimated AUC of 0.82 with a 95% confidence interval from 0.77 to 0.87. This would seem to imply that if we use this CPM in a new setting, we can be quite confident that the AUC will be at least 0.77. Unfortunately, that is not the case at all.

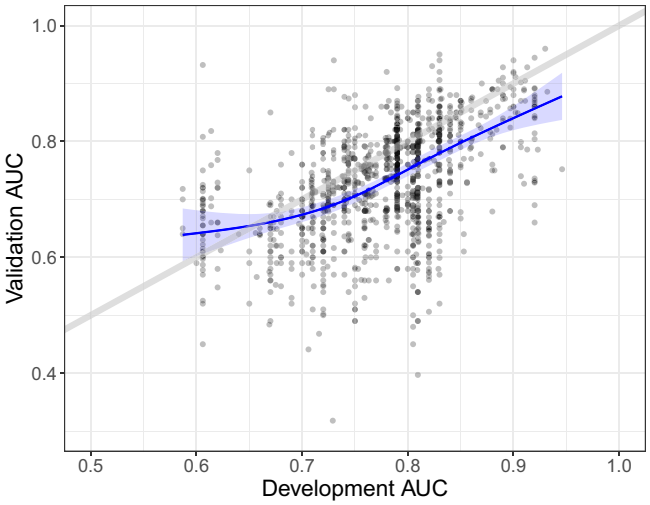

FIGURE 1 | Relation between development AUCs and validation AUCs in the Tufts-PACE CPM Registry. The regression curve shows that the validation AUCs tend to be lower than the development AUCs.

After the first external validation of the CRUSADE model, 8 more validations were performed. We show the cumulative results in Figure 2 as a forest plot. We used the R package metafor [9] to do a standard random-effects meta-analysis of all 9 external validations. We estimate the pooled AUC to be 0.69 with a 95% confidence interval from 0.63 to 0.76. Remarkably, this confidence interval excludes the entire confidence interval after the first validation.

The large uncertainty about the pooled AUC is due to the large heterogeneity between validation studies (Figure 2). We quantify this heterogeneity as the between-study standard deviation τ, and in the case of the CRUSADE model we estimate $\tau = 0.09$. The large heterogeneity may be due to many factors including differences in population, standard of care, and variations in predictor and outcome definitions and assessment [10], in addition to model misspecification.

In the case of meta-analysis of clinical trials, prediction intervals for the effect of the treatment in a new study are recognized as important [11]. We consider the 95% prediction interval for the AUC of a prediction model in a new setting to be more relevant than the 95% confidence interval for the pooled AUC. We find that the prediction interval based on the 9 external validations is centered at 0.69 and extends from 0.5 to 0.89—a range of discriminatory performance that spans from useless to what most would consider very good [12]. Thus, even after 9 external validations, the performance in a new setting remains highly uncertain.

As a further illustration, consider the logistic EuroSCORE CPM for patients undergoing major cardiac surgery [13]. This model has 83 external validations. In Figure 3, we show the results of "cumulative" fixed- and random-effects meta-analyses. That is,

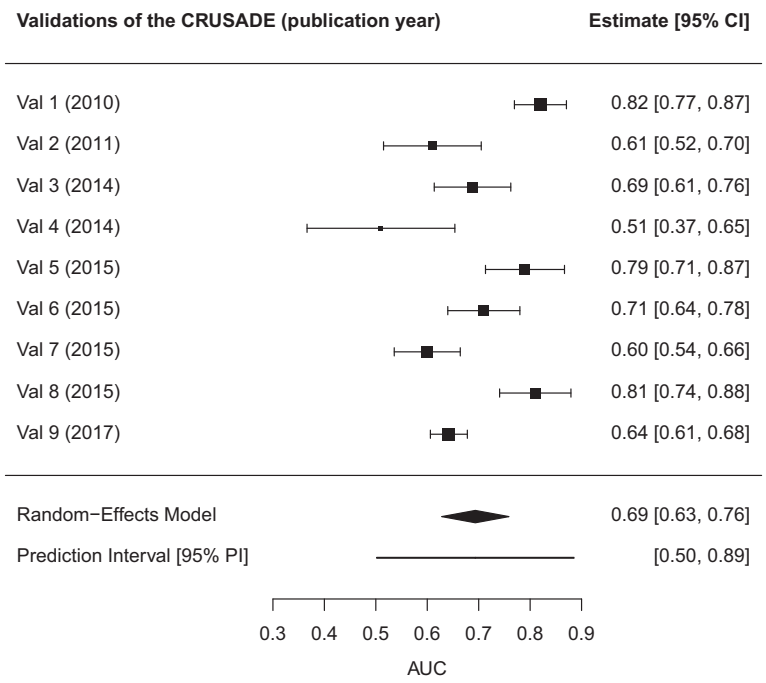

FIGURE 2 | Forest plot and random-effects meta-analysis of the AUC estimates of validations for the CRUSADE CPM. The black diamond is the 95% confidence interval for the mean AUC across all validations. The solid line at the bottom represents the 95% prediction interval for the true AUC in a new study.

we show the 95% confidence intervals for the mean AUC and the 95% prediction intervals for the AUC in a new setting based on the first $1, 2, 3, \ldots, 83$ validation studies. In the left panel, we show the results of fixed-effects meta-analyses. In that case, we assume that τ is zero and therefore the confidence interval for the pooled AUC and the prediction interval for the AUC in a new study are equal. After about 50 validations, the location of the intervals has stabilized and their width has become negligible.

In the right panel, we show the results of random-effects meta-analyses where we used the REML method to estimate the heterogeneity τ. When we have just one validation, it is not possible to estimate τ and it is set to zero. When we have few validations, the width of the intervals varies considerably because the estimates of τ are very noisy. Eventually, we see the width of the intervals stabilizing and then gradually shrinking. While the width of the confidence interval will tend to zero, the width of the prediction interval will not. In fact, it will tend to $2 \times 1.96 \times \tau$. Thus, no matter how many validations have been done, there will always remain substantial uncertainty about the AUC of the EuroSCORE model in a new setting.

It is obvious from Figure 3 that it is inappropriate to assume that τ is zero. This will lead to gross underestimation of the uncertainty for the AUC. To make confidence intervals or prediction intervals with the correct coverage, we need accurate estimates of τ. Unfortunately, most CPMs have very few validation studies. Of the CPMs included in our study, 239/469 (51%) have only one external validation. The median number of external validations is 2 with an IQR from 1 to 3. Clearly, this is insufficient to estimate τ with good accuracy. Even worse, the usual methods (such as REML or the well-known method of DerSimonian and Laird [14]) have a tendency to estimate τ at zero. This happens because the variation between the observed AUCs consists of within- and between-study variation (heterogeneity). If the observed variation can be explained by the within-study variation alone, then τ will be estimated at zero [15]. As we will demonstrate, this will often lead to severe undercoverage of confidence and prediction intervals. This is the problem we want to address.

In the next section, we set up hierarchical (or multi-level) models to study the 469 CPMs and their validations. In particular, we estimate the distribution of τ across the CPMs. We also estimate the distribution of the pooled AUCs. Next, we implement two (empirical) Bayesian models. The first has a flat prior for the average AUC, and an informative prior for τ, and the second has informative priors for both. We also have a "poor man's" Bayesian method where we set τ equal to a fixed (non-zero) value which can easily be done with the metafor package [9]. To evaluate and compare the frequentist and Bayesian methods, we use leave-one-study-out cross-validation.

3 | Methods

We use the observed AUC values of cardiovascular Clinical Prediction Models (CPMs) from the Tufts PACE CPM Registry [6]. This is a publicly available compilation of models predicting outcomes for patients at risk for, or already having, cardiovascular disease. The inclusion criteria of the registry require the CPM to predict a binary cardiovascular outcome, presented in a way that enables patient risk prediction. The search strategy considered CPMs that were developed and published between 1990 and March 2015 case. Next, a SCOPUS citation search on March 22, 2017, identified external validations of the CPMs, defined as reports studying the same model in a new population. In total, the

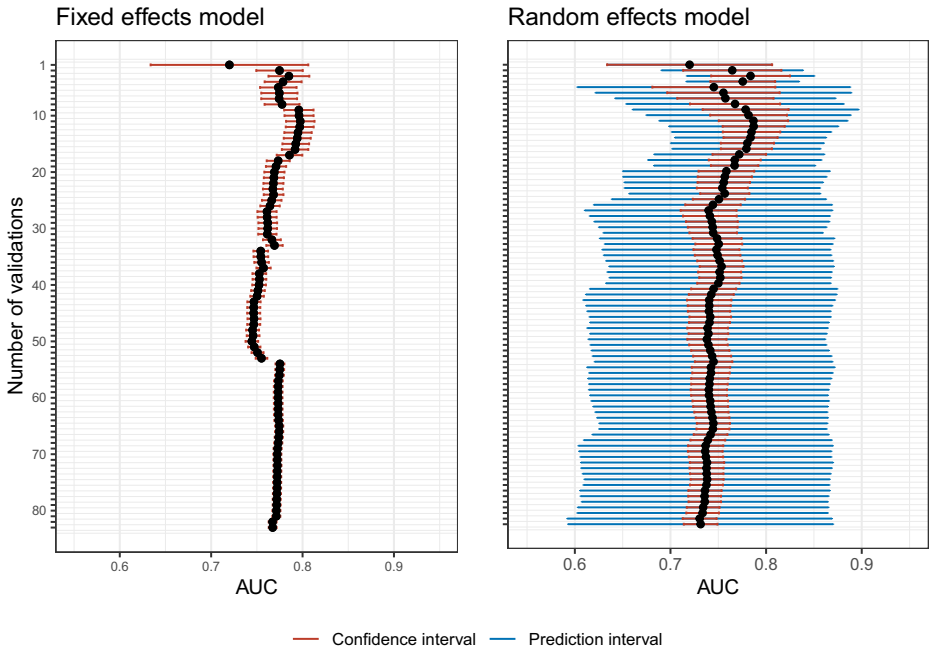

FIGURE 3 | Cumulative confidence and prediction interval of fixed and random-effects meta-analyses for the EuroScore model based on the first 1,2,3, . . . ,83 validation studies.

registry has 1382 CPMs and 2030 external validations. Most models are for patients with stroke ($n = 97$) and for patients undergoing cardiac surgery ($n = 46$). We selected CPMs with at least one external validation and complete information. Thus our data consists of 469 CPMs with 1603 external validations (see the flowchart in the Appendix A).

Since the validation AUCs are grouped within CPMs, we set up a collection of random-effects meta-analysis models [16]. So, for the j -th validation AUC of the i -th CPM, we assume:

$$\widehat{AUC}_{ij} \sim N(AUC_{ij}, s_{ij}^2) \tag{1}$$

$$AUC_{ij} \sim N(AUC_i, \tau_i^2) \tag{2}$$

where $i = 1, 2, \ldots, 469$, $j = 1, 2, \ldots, n_i$ and s_{ij} denotes the standard error of the observed $\widehat{AUC}_{ij}$. Despite the fact that AUCs are bounded between 0 and 1, we believe the normal distribution is appropriate because the observed values stay well away from the bounds (see Figure 1). As usual in meta-analyses, we will ignore the uncertainty about s_{ij} .

From the frequentist point of view, the AUC_i and τ_i are fixed parameters that are to be estimated. The defining feature of a fixed-effects meta-analysis is that τ_i is assumed to be zero. When τ_i is not assumed to be zero, the metafor package has 12 different methods to estimate it [9]. Here, we use the default REML method, but in the supplement, we also consider the method of Sidik and Jonkman [17] which tends to behave most differently from REML among the remaining 11 methods. In our case, however, the results turn out to be very similar to REML.

From the Bayesian perspective, we consider the AUC_i and τ_i to be random variables for which we need to specify prior distributions. We will assume a normal distribution for the AUC_i and a lognormal distribution for the τ_i:

$$AUC_i \sim N(\mu_{AUC}, \sigma_{AUC}^2) \tag{3}$$

$$\log(\tau_i) \sim N(\mu_\tau, \sigma_\tau^2) \tag{4}$$

This implies that the mean and variance of the τ_i are

$$\begin{aligned} \mathrm{E}(\tau_i) &= \exp\left(\mu_\tau + \frac{\sigma_\tau^2}{2}\right) \\ \mathrm{Var}(\tau_i) &= \left[\exp(\sigma_\tau^2) - 1\right] \exp(2\mu_\tau + \sigma_\tau^2) \end{aligned} \tag{5}$$

We will use the method of maximum likelihood to estimate the 4 parameters of our model ($\mu_{AUC}, \sigma_{AUC}, \mu_\tau$ and σ_τ). The likelihood does not have a closed form, so we use the R-package rstan to do the computation [18]. This package provides an R interface to the Stan platform for MCMC sampling to perform Bayesian inference. We specify uniform priors for each of the 4 parameters and then take their posterior modes as the MLEs. In terms of the estimated model parameters, the estimated mean of the τ_i is

$$\overline{\tau} = \exp\left(\hat{\mu}_\tau + \frac{\hat{\sigma}_\tau^2}{2}\right) \tag{6}$$

Our main goal is to predict the AUC in a new setting and to provide a 95% prediction interval. We use the `rma` function from the `metafor` package [9] to do 3 versions of frequentist meta-analyses:

1. fixed-effects model where we assume $\tau_i = 0$,
2. random-effects where we estimate the τ_i with REML,
3. random-effects model where we assume $\tau_i = \overline{\tau}$.

One could argue that the first and third models are actually Bayesian models with extremely strong priors for the

TABLE 1 | Prediction intervals for the true and observed AUC in the next study. Based on the first n studies, $\widehat{AUC}_{1:n}$ is the estimate of the pooled AUC, $s_{1:n}$ is the associated standard error, and $\hat{\tau}_{1:n}$ is the estimate of τ. In the Bayesian approach, AUC_{post} and SD_{post} are the posterior mean and standard deviation of the pooled AUC. s_{n+1} is the standard error of the observed AUC in the $(n+1)$-th study.

Model	Description	Prediction interval for AUC_{n+1}	Prediction interval for $\widehat{AUC}_{n+1}$
FE	$\tau = 0$	$\widehat{AUC}_{1:n} \pm 1.96\sqrt{s^2_{1:n}}$	$\widehat{AUC}_{1:n} \pm 1.96\sqrt{s^2_{1:n} + s^2_{n+1}}$
RE	Estimate τ	$\widehat{AUC}_{1:n} \pm 1.96\sqrt{s^2_{1:n} + \hat{\tau}^2_{1:n}}$	$\widehat{AUC}_{1:n} \pm 1.96\sqrt{s^2_{1:n} + \hat{\tau}^2_{1:n} + s^2_{n+1}}$
RE	$\tau = \overline{\tau}$	$\widehat{AUC}_{1:n} \pm 1.96\sqrt{s^2_{1:n} + \overline{\tau}^2}$	$\widehat{AUC}_{1:n} \pm 1.96\sqrt{s^2_{1:n} + \overline{\tau}^2 + s^2_{n+1}}$
Bayes	Prior	$\widehat{AUC}_{1:n} \pm 1.96\sqrt{SD^2_{post}}$	$\widehat{AUC}_{1:n} \pm 1.96\sqrt{SD^2_{post} + s^2_{n+1}}$

τ_i and a non-informative prior for the AUC_i. We use the R-package `baggr` [19] to do two versions of (empirical) Bayesian meta-analyses:

4. Bayesian meta-analysis with a non-informative prior for AUC_i, and an informative prior for τ_i,
5. Bayesian meta-analysis with informative priors for both AUC_i and τ_i.

We used the default settings for `baggr` which means in particular that we run four independent chains of length 2000 steps with 1000 steps burn-in.

To evaluate and compare the performance of these five methods, we use a leave-one-study-out cross-validation approach. We fix a number n of validation studies ($n = 1, 2, \ldots, 5$) and then we use $((\widehat{AUC}_{i,1}, s_{i,1}), , \ldots, (\widehat{AUC}_{i,n}, s_{i,n})$ and $s_{i,n+1}$ to predict $\widehat{AUC}_{i,n+1}$. We also form a 95% prediction interval for $\widehat{AUC}_{i,n+1}$. We do this by forming the 95% prediction interval for the true $AUC_{i,n+1}$, and then accounting for the sampling error. We show the formulas in Table 1.

We make sure that there are at least $n+1$ studies in the meta-analysis, so that we can check how often the observed AUCs of the left-out studies fall within the prediction interval. Hence, only CPMs with at least 2 validations are used in the cross-validation. If the coverage of the observed AUCs is 95% then we conclude that the coverage of the prediction interval for the true AUC is also 95%. Finally, we also compute the root mean squared prediction error (RMSE) for the observed AUC in a new study.

When we have only two validations and we leave one study out, then we are left with only one study to run the meta-analysis. If we then attempt to use REML to estimate the heterogeneity, the `rma` function will simply estimate zero. For approaches 1 and 3 where we set the heterogeneity to a fixed value and for the Bayesian approaches 4 and 5, having only one study is of course not an issue.

4 | Results

Recall from formulas (3) and (4) that our model has four parameters, namely the mean and standard deviation of the $\log(\tau_i)$ and the mean and standard deviation of the AUC_i. We consider two variants to set these parameters. In the first variant, we set the mean of the AUC_i to zero and their standard deviation to a large value to obtain an essentially flat or "non-informative" prior. We estimate the mean and standard deviation of the $\log(\tau_i)$ using maximum likelihood as described in the previous section. In the second variant, we estimate all four parameters. We provide the estimates of both variants in Table 2.

TABLE 2 | The estimated priors of the τ_i and AUC_i.

Model	μ_τ	σ_τ	μ_{AUC}	σ_{AUC}
Non-informative for AUC_i, informative for τ_i	−2.94	0.27	0	10
Informative AUC_i and τ_i	−2.89	0.21	0.73	0.07

We used formulas (5) to compute the mean of τ in the first variant as 0.055 with a standard deviation of 0.15. The mean and standard deviation of τ in the second variant are very similar at 0.057 and 0.12. For our fixed-effects meta-analysis with non-zero heterogeneity, we set $\overline{\tau} = 0.055$.

When the prediction intervals are based on only one study, both the fixed-effects model and the random-effects model with REML estimation can only set τ equal to zero which results in severe undercoverage (Figure 4). The fixed-effects model will continue to undercover even when we base the prediction intervals on more studies, but the coverage of the random-effects model will increase to the nominal level. When we base the prediction intervals on 5 or more studies, the coverage of the random-effects model becomes close to nominal. However, only a small minority of CPMs (69/469; 15%) have 5 or more external validations. The two Bayesian models and the model where we set $\tau = 0.055$ always had near nominal coverage. The slight undercoverage that remained may be expected from Wald-type intervals which ignore the uncertainty about the standard errors of the observed AUCs.

Turning to the Root Mean Squared Prediction Error, we note the relatively poor performance of the fixed-effects model, which is due to the inefficient weighing of the individual studies (Figure 5). We also note the superior performance of the Bayesian model with an informative prior for the AUC_i which is due to the shrinkage towards the overall average of the AUCs at 0.734. When we use a single validation to predict the AUC in a new setting, the error of the Bayesian model is on average about 1 percentage point less than the other methods. When we use more validations, this advantage decreases.

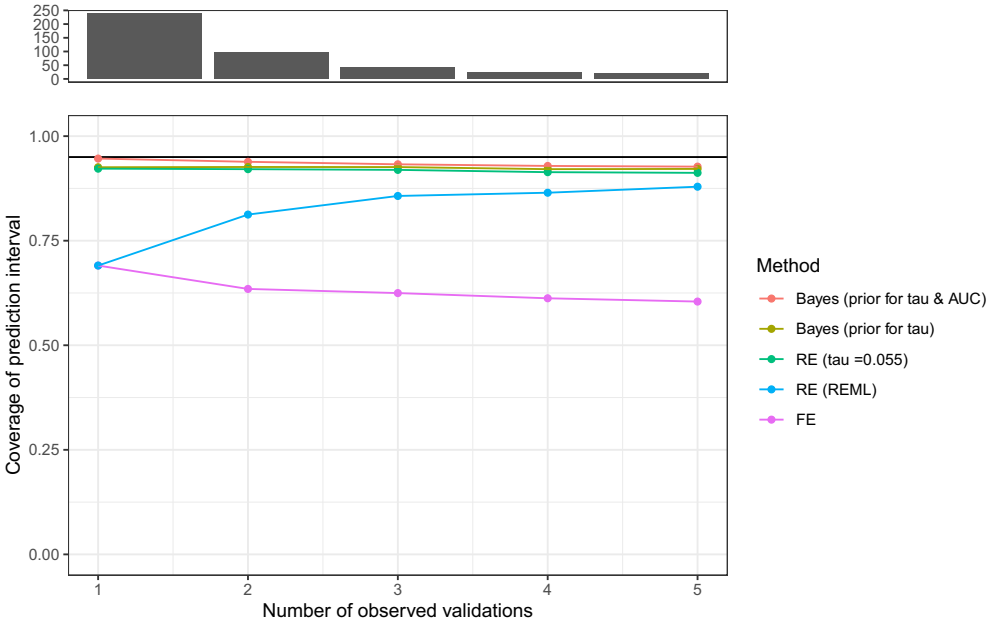

FIGURE 4 | Top panel: The number of CPMs with exactly 1, 2, . . . , 5 external validations. Bottom panel: Coverage of the prediction intervals for the observed AUC in the next study.

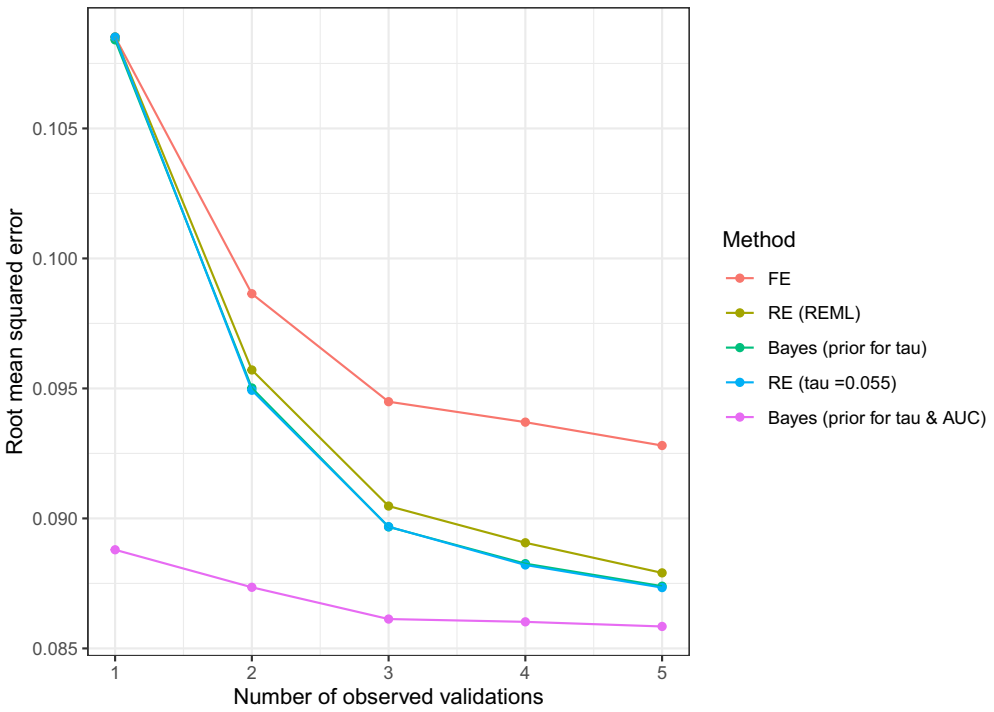

FIGURE 5 | Root Mean Squared Prediction Error for the observed AUC in the next study.

5 | Discussion

We noted considerable heterogeneity among the external validations of cardiovascular (CPMs). We estimated that the standard deviation τ is about 0.05 on average with a standard deviation of 0.01. Additionally, we estimated a normal distribution for the pooled AUCs with a mean of 0.73 and a standard deviation of 0.07. Using these distributions as an empirical prior substantially outperformed frequentist methods of meta-analysis in terms of prediction accuracy and coverage of the prediction interval for the next study. Especially when there were few validation studies (fewer than 5), frequentist methods showed severe undercoverage, while the empirical Bayes approach was very close to nominal. Our study illustrates the usefulness of empirical Bayes approaches for meta-analyses in general, where estimation of heterogeneity is unreliable unless a large number of studies are analyzed.

If τ is 0.05, then the 95% prediction interval for the AUC in a new setting will have a width of at least $+/-$ 0.1, no matter how many

validations have been done. In this sense, our findings support the claim of Van Calster et al. (2023) [10] that "there is no such thing as a validated prediction model".

Obviously, external validations should be taken into account before deployment of a CPM. However, most published CPMs have never been externally validated [20]. Even when external validations are available they do not provide a solid guarantee about the AUC in the next study. Therefore the discriminatory performance in a new setting should be monitored after deployment. While many researchers may understand the AUC as an intrinsic measure of CPM quality, in fact AUC is an extrinsic property of a CPM that emerges only when a model is applied to a specific population.

There are two broad reasons for variation in AUC when transporting a model from one setting to another: (1) differences in the heterogeneity of the sample; and (2) model misspecification. Regarding the first, more heterogeneous populations will generally result in larger AUC values. For example, at the extreme, a correctly specified model will have an AUC of exactly 0.5 if transported to a new population where each patient has the same value for each of the predictor variables. There are various ways to quantify patient heterogeneity, but one intuitive measure is the standard deviation of the linear predictor [21]. Unfortunately, such measures could not be calculated for our validations since we had no access to individual patient data. Secondly, model misspecification reflects differences in the associations of the predictor and outcome variables between the derivation and validation samples. This can arise for many reasons, including changes in how data are collected, changes in how predictors or outcomes are defined, changes in clinician and patient behavior, and changes with respect to the distribution of effect modifiers that are not included in the model [22].

In a previous analysis, we performed 158 validations of 108 published CPMs in the Tufts PACE registry [23]. We used publicly available data from randomized controlled trials for validation, where we expect less heterogeneity than in less-selected observational data sources as typically used for the development of CPMs. We found that the AUC differed substantially between model derivation (0.76 with interquartile range 0.73–0.78) and validation (0.64 with interquartile range 0.60–0.67). Indeed, approximately half of this decrease could be accounted for by the narrower case-mix (less heterogeneity) in the validation samples; the remainder could be attributed to model misspecification.

The present study has two main limitations. First, the Tufts-PACE CPM Registry only has CPMs from cardiovascular medicine. As far as we know, the registry is unique in that it holds many different CPMs with multiple validations per CPM. Therefore, we can merely speculate about the generalizability of our findings to other medical specialties. The factors we discussed that influence the heterogeneity are not specific to cardiovascular medicine, and therefore we would expect qualitatively similar results. For example, a review of the IMPACT and CRASH models for patients with traumatic brain injury showed large variation in AUC values, similar to what we found for cardiovascular risk prediction models [24].

The second limitation is our exclusive focus on the AUC. Now, it can be argued that the AUC does not provide the most pertinent information about the usefulness of a CPM. The AUC is a measure of discrimination across all possible cut-offs and as such it is not directly meaningful when a particular cut-off is used in clinical practice to support decision making. Decision-analytic summary measures such as Net Benefit quantify clinical usefulness better [25]. Net Benefit depends on discrimination (higher with higher AUC) and calibration (highest with correct calibration at the decision threshold). Moreover, the clinical context is important, with higher Net Benefit if the decision threshold is in the middle of the risk distribution. Unfortunately, information about calibration is reported very poorly in most publications. In fact, almost none of the validations in the Tufts PACE CPM Registry reported a numerical summary of calibration. In our study of 158 external validations of selected models [23] we did consider several other performance measures besides the AUC, including calibration. However, based on the data we had access to, we were unable to validate the same model in multiple validations which would allow an analysis like the present paper. Further work is therefore necessary on quantifying calibration across validations of CPMs. A natural starting point is to quantify heterogeneity in summary measures for calibration in the large, where poor validity is commonly observed [26].

We conclude that if we want to use a CPM in a new setting, large uncertainty about the AUC will typically remain even when many external validation studies have been performed. Unfortunately, in many cases, only very few validations are available. We have proposed an empirical Bayes method provides a realistic assessment of the uncertainty even in those situations. The method is easy to use with existing software for Bayesian meta-analysis such as the `R` package `baggr` [19].

Finally, the only way to reduce the uncertainty about a CPM's performance both in terms of calibration and discrimination is to validate and update it in the setting where it is to be used. This is referred to as "targeted validation" [27], and related to continuous updating [28–31]. In light of the findings of the present study, we consider updating a crucial step before full-scale deployment.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The Tufts PACE CPM Registry is publicly available at https://www.pacecpmregistry.org. The data and code to reproduce all tables and figures are available as an online supplement.

References

1. E. Steyerberg, *Clinical Prediction Models* (New York: Springer, 2009).

2. F. E. Harrell, *Regression Modeling Strategies* (Switzerland: Springer International Publishing, 2015).

3. A. C. Justice, "Assessing the Generalizability of Prognostic Information," *Annals of Internal Medicine* 130, no. 6 (1999): 515.

4. D. G. Altman and P. Royston, "What Do We Mean by Validating a Prognostic Model?," *Statistics in Medicine* 19, no. 4 (2000): 453–473.

5. E. W. Steyerberg and F. E. Harrell, "Prediction Models Need Appropriate Internal, Internal-External, and External Validation," *Journal of Clinical Epidemiology* 69 (2016): 245.

6. B. S. Wessler, J. Nelson, J. G. Park, et al., "External Validations of Cardiovascular Clinical Prediction Models: A Large-Scale Review of the Literature," *Circulation. Cardiovascular Quality and Outcomes* 14, no. 8 (2021): e007858.

7. S. Subherwal, R. G. Bach, A. Y. Chen, et al., "Baseline Risk of Major Bleeding in Non–ST-Segment–Elevation Myocardial Infarction: The CRUSADE Bleeding Score," *Circulation* 119, no. 14 (2009): 1873–1882.

8. E. Abu-Assi, J. M. Gracía-Acuña, I. Ferreira-González, C. Peña-Gil, P. Gayoso-Diz, and J. R. González-Juanatey, "Evaluating the Performance of the Can Rapid Risk Stratification of Unstable Angina Patients Suppress Adverse Outcomes With Early Implementation of the ACC/AHA Guidelines (CRUSADE) Bleeding Score in a Contemporary Spanish Cohort of Patients With Non–ST-Segment Elevation Acute Myocardial Infarction," *Circulation* 121, no. 22 (2010): 2419–2426.

9. W. Viechtbauer, "Conducting Meta-Analyses in R With the Metafor Package," *Journal of Statistical Software* 36, no. 3 (2010): 1–48.

10. B. Van Calster, E. W. Steyerberg, L. Wynants, and M. Van Smeden, "There Is no Such Thing as a Valiyeard Prediction Model," *BMC Medicine* 21, no. 1 (2023): 70.

11. J. IntHout, J. P. Ioannidis, M. M. Rovers, and J. J. Goeman, "Plea for Routinely Presenting Prediction Intervals in Meta-Analysis," *BMJ Open* 6, no. 7 (2016): e010247.

12. A. A. H. De Hond, E. W. Steyerberg, and B. Van Calster, "Interpreting Area Under the Receiver Operating Characteristic Curve," *Lancet Digital Health* 4, no. 12 (2022): e853–e855.

13. F. Roques, P. Michel, A. R. Goldstone, and S. A. M. Nashef, "The Logistic Euroscore," *European Heart Journal* 24, no. 9 (2003): 882–883.

14. R. DerSimonian and N. Laird, "Meta-Analysis in Clinical Trials," *Controlled Clinical Trials* 7, no. 3 (1986): 177–188.

15. M. Borenstein, L. V. Hedges, J. P. Higgins, and H. R. Rothstein, "A Basic Introduction to Fixed-Effect and Random-Effects Models for Meta-Analysis," *Research Synthesis Methods* 1, no. 2 (2010): 97–111.

16. A. Whitehead and J. Whitehead, "A General Parametric Approach to the Meta-Analysis of Randomized Clinical Trials," *Statistics in Medicine* 10, no. 11 (1991): 1665–1677.

17. K. Sidik and J. N. Jonkman, "A Simple Confidence Interval for Meta-Analysis," *Statistics in Medicine* 21, no. 21 (2002): 3153–3159.

18. D. T. Stan, "RStan: the R Interface to Stan," (2023): *R package version 2.32.3*.

19. W. Wiecek and R. Meager, "BAGGR: Bayesian Aggregate Treatment Effects," (2022): *R package version 0.7.6*.

20. G. C. M. Siontis, I. Tzoulaki, P. J. Castaldi, and J. P. A. Ioannidis, "External Validation of New Risk Prediction Models Is Infrequent and Reveals Worse Prognostic Discrimination," *Journal of Clinical Epidemiology* 68, no. 1 (2015): 25–34.

21. T. P. A. Debray, Y. Vergouwe, H. Koffijberg, D. Nieboer, E. W. Steyerberg, and K. G. M. Moons, "A New Framework to Enhance the Interpretation of External Validation Studies of Clinical Prediction Models," *Journal of Clinical Epidemiology* 68, no. 3 (2015): 279–289.

22. S. G. Finlayson, A. Subbaswamy, K. Singh, et al., "The Clinician and Dataset Shift in Artificial Intelligence," *New England Journal of Medicine* 385, no. 3 (2021): 283–286.

23. G. Gulati, J. Upshaw, B. S. Wessler, et al., "Generalizability of Cardiovascular Disease Clinical Prediction Models: 158 Independent External Validations of 104 Unique Models," *Circulation. Cardiovascular Quality and Outcomes* 15, no. 4 (2022): 248–260.

24. S. A. Dijkland, K. A. Foks, S. Polinder, et al., "Prognosis in Moderate and Severe Traumatic Brain Injury: A Systematic Review of Contemporary Models and Validation Studies," *Journal of Neurotrauma* 37, no. 1 (2020): 1–13.

25. A. J. Vickers, B. Van Calster, and E. W. Steyerberg, "Net Benefit Approaches to the Evaluation of Prediction Models, Molecular Markers, and Diagnostic Tests," *British Medical Journal* 352 (2016): i6.

26. B. Van Calster, D. J. McLernon, M. Van Smeden, L. Wynants, and E. W. Steyerberg, "Calibration: The Achilles Heel of Predictive Analytics," *BMC Medicine* 17, no. 1 (2019): 1–7.

27. M. Sperrin, R. D. Riley, G. S. Collins, and G. P. Martin, "Targeted Validation: Validating Clinical Prediction Models in Their Intended Population and Setting," *Diagnostic and Prognostic Research* 6, no. 1 (2022): 24.

28. E. W. Steyerberg, G. J. Borsboom, H. C. Van Houwelingen, M. J. Eijkemans, and J. D. F. Habbema, "Validation and Updating of Predictive Logistic Regression Models: A Study on Sample Size and Shrinkage," *Statistics in Medicine* 23, no. 16 (2004): 2567–2586.

29. T. S. Genders, E. W. Steyerberg, H. Alkadhi, et al., "A Clinical Prediction Rule for the Diagnosis of Coronary Artery Disease: Validation, Updating, and Extension," *European Heart Journal* 32, no. 11 (2011): 1316–1330.

30. M. Binuya, E. Engelhardt, W. Schats, M. Schmidt, and E. Steyerberg, "Methodological Guidance for the Evaluation and Updating of Clinical Prediction Models: A Systematic Review," *BMC Medical Research Methodology* 22, no. 1 (2022): 316.

31. H. dAA, I. M. Kant, M. Fornasa, et al., "Predicting Readmission or Death After Discharge From the ICU: External Validation and Retraining of a Machine Learning Model," *Critical Care Medicine* 51, no. 2 (2023): 291–300.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.

Appendix A

The data selection is shown in the flowchart (Figure A1). At the start, there are a total of 2030 validations of 575 CPMs. After filtering we have 1603 validations from 469 CPMs.

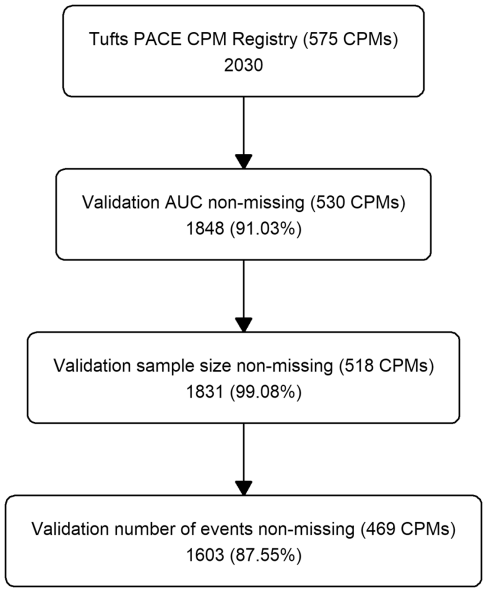

FIGURE A1 | Flowchart of data filtering.