Received: 6 May 2021 | Revised: 26 February 2023 | Accepted: 19 May 2023

DOI: 10.1002/sim.9817

RESEARCH ARTICLE

Statistics in Medicine WILEY

Propensity-based standardization to enhance the validation and interpretation of prediction model discrimination for a target population

Valentijn M. T. de Jong[1,2] | **Jeroen Hoogland**[1,3] | **Karel G. M. Moons**[1] | **Richard D. Riley**[4] | **Tri-Long Nguyen**[5] | **Thomas P. A. Debray**[1,6]

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

[2]Data Analytics and Methods Task Force, European Medicines Agency, Amsterdam, The Netherlands

[3]Department of Epidemiology and Data Science, Amsterdam University Medical Centers, Amsterdam, The Netherlands

[4]Institute of Applied Health Research, College of Medical and Dental Sciences, University of Birmingham, Birmingham, UK

[5]Section of Epidemiology, Department of Public Health, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark

[6]Smart Data Analysis and Statistics, Utrecht, The Netherlands

Correspondence
Valentijn M. T. de Jong, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Str. 6.131, P.O. Box 85500, 3508GA Utrecht, The Netherlands.
Email: Valentijn.M.T.de.Jong@gmail.com

Funding information
European Commission, Grant/Award Number: 825746

External validation of the discriminative ability of prediction models is of key importance. However, the interpretation of such evaluations is challenging, as the ability to discriminate depends on both the sample characteristics (ie, case-mix) and the generalizability of predictor coefficients, but most discrimination indices do not provide any insight into their respective contributions. To disentangle differences in discriminative ability across external validation samples due to a lack of model generalizability from differences in sample characteristics, we propose propensity-weighted measures of discrimination. These weighted metrics, which are derived from propensity scores for sample membership, are standardized for case-mix differences between the model development and validation samples, allowing for a fair comparison of discriminative ability in terms of model characteristics in a target population of interest. We illustrate our methods with the validation of eight prediction models for deep vein thrombosis in 12 external validation data sets and assess our methods in a simulation study. In the illustrative example, propensity score standardization reduced between-study heterogeneity of discrimination, indicating that between-study variability was partially attributable to case-mix. The simulation study showed that only flexible propensity-score methods (allowing for non-linear effects) produced unbiased estimates of model discrimination in the target population, and only when the positivity assumption was met. Propensity score-based standardization may facilitate the interpretation of (heterogeneity in) discriminative ability of a prediction model as observed across multiple studies, and may guide model updating strategies for a particular target population. Careful propensity score modeling with attention for non-linear relations is recommended.

KEYWORDS
concordance, external validation, prediction model, propensity score, standardization

Valentijn M. T. de Jong and Jeroen Hoogland contributed equally to this study. Tri-Long Nguyen and Thomas P. A. Debray contributed equally to this study.

Disclaimer: The views expressed in this paper are the personal views of the authors and may not be understood or quoted as being made on behalf of or reflecting the position of the regulatory agency/agencies or organizations with which the authors are employed/affiliated.

1 | INTRODUCTION

Prediction models provide estimates of absolute risk that a particular health status is present (diagnosis) or will occur in the future (prognosis).[1] The development of prediction models has seen a rapid growth in medicine. Unfortunately, many prediction models perform worse when tested (ie, validated) in or applied to new individuals.[2-5] Common reasons for inaccurate predictions for new individuals are the violation of model assumptions, omission of important predictors, poor handling of missing data in the development or validation data and, in particular, overfitting of the developed model.[6-9] The performance of model predictions may also be affected by the (lack of) representativeness of samples in view of the target population. For instance, datasets for validation may have differences (compared to the target population) in patient characteristics (case-mix), in predictor and outcome measurements, and in (the presence of) measurement error, thereby affecting prediction model performance across validation samples.[10-13]

It is widely advocated that when researchers develop a new prediction model, they explore whether its predictions are sufficiently accurate across different validation samples from the target population.[14-16] The model's predictive performance is therefore assessed in new samples that have not been used during its development (so-called external validation).[10,11,17-20] To facilitate investigation of model performance, researchers developing a novel prediction model ideally collect individual participant data (IPD) from two or more settings, institutes or even predesigned (validation) studies.[16] One (or more) data set can then be used for model development and the remaining studies for external validation purposes. Results from the validation study are then used to confirm whether the model is adequate or to recommend certain revisions prior to its implementation in practice for (a) particular target population(s). Clearly, the choice of appropriate data for the purpose of validation is no trivial task. Important characteristics that affect such a decision include sample size, availability of predictors and outcome, and representativeness of the study population with respect to the target population in which the model will be used. If the validation sample does not fully represent the target population, estimates of model performance may be misleading. Although it is clearly preferable that a prediction model is validated in a large and representative sample from the target population, such data are not always available. Nonetheless, validation sets from populations sufficiently related to the target population may still provide the required information. Therefore, the question we here consider is how we can use validation data sets that are not fully representative of the target population of interest for the purpose of prediction model validation.

When pursuing an external validation study that is not fully representative of the target population, changes in model performance with respect to the development study should be interpreted with caution. Decline in a prediction model performance measure does not necessarily imply that the model coefficients (eg, predictor weights) are not generalizable to the target population. In particular, discriminative ability may drop if the validation sample is considerably more homogeneous than the development sample, even if the model generalizes well:[21] it is simply more difficult to discriminate amongst participants that are similar to each other. Likewise, adequate performance upon external validation does not necessarily imply that the model transports well to the target population, as this requires some degree of consistent model performance across multiple validation samples with different case-mix which may not have been fully reflected in the external validation data at hand.[10,11] The interpretation of prediction model performance becomes particularly challenging in IPD meta-analysis, where studies may differ in eligibility criteria, measurement methods and so forth. The presence of between-study heterogeneity is a common concern in meta-analysis of prediction model performance, and obfuscates to what extent the model is actually generalizable.

To disentangle the possible sources of variability in prediction model performance across multiple validation studies, it has been recommended to quantify the relatedness between the development and validation samples.[3] This allows for the isolation of changes in performance that can only be attributed to the use of regression coefficients that lack transportability, thereby guiding which type of model revisions may be necessary. An alternative approach that has been suggested is the model-based concordance measure, which can be used to quantify of the effect of case-mix on discriminative ability.[22,23]

Further, for cases when IPD from both development and validation samples are available, Debray et al proposed to develop a so-called membership model (not to be confused with the actual prediction model),[3] which calculates the probability that an individual belongs to the development or validation sample. The concordance index of this membership model reflects the degree of (dis)similarity of the development and validation samples and can be used to identify whether the evaluation of a particular model's performance is likely to be affected by case-mix differences.[3,4] In this article, we build on this membership model framework and consider the use of propensity score weighting methods to standardize prediction model concordance measures for a particular target population. These concordance measures may be

estimated in one or more validation samples from settings and populations that are different from but related to the intended target population, as well as at least one sample from the intended target population.

Our proposed weighting methods are derived from well-known epidemiological approaches to standardizing samples with respect to their covariate distributions.[24,25] Such standardization may help researchers to interpret differences in discriminative ability at external validation (as compared to the development sample or to other validation samples) and identify the usefulness of specific model updating or revision strategies. This usage of propensity-based standardization can also be conducted when IPD from multiple samples are readily available, which appears particularly useful in an individual participant data meta-analysis (IPDMA) or when using a large electronic healthcare database that includes multiple data sets (clusters) that can be used for model development validations.

IPDMA and routinely collected clustered healthcare data sets are used increasingly often to develop new prediction models and to assess their performance in external data.[26,27] Often, the studies from an IPDMA differ in design and participant characteristics, and may not always adequately represent the population where the model will eventually be implemented. It is, for instance, possible that some data sets were obtained from randomized trials, or comprise patients from earlier time periods. Even though these data sets can still be used to inform prediction model development, performance estimates derived from these data sets can be misleading if no effort is made to adjust for their poor representativeness of the target population.

We have recently suggested to use propensity scores to assist external validation of clinical prediction models.[28,29] This article explores the untapped value of propensity-based standardization in the validation of clinical prediction models using IPD from two or more sources (studies or another form of clusters). We aim to provide metrics of discriminative ability that are adjusted for case-mix variation, and to test statistical properties of this methodology. In Section 2, we present propensity-based standardization methods in the context of clinical prediction models as well as propensity-standardized validation measures, in Section 3 we describe a motivating example with illustrative data from multiple studies on the diagnosis of deep vein thrombosis (DVT), in Section 4 we provide a simulation study to assess the proposed methods, and finally Section 5 provides a discussion of our results.

2 | PROPENSITY SCORE STANDARDIZATION AND CLINICAL PREDICTION MODELS

Propensity score methods were initially proposed for the estimation of causal (eg, treatment) effects in non-randomized data.[24] Clinical prediction models typically do not aim to provide a causal explanation[30,31] and therefore do not (strictly) require the incorporation of treatment propensity scores.[32] Although it is possible to account for received treatments during the development and validation of prediction models,[33,34] we propose a different use of propensity score methods. In particular, when IPD from multiple settings or populations are available to the researcher, one can estimate the probability that a certain individual is a member of a certain sample. Provided that at least one sample is from the prediction model's target population, these propensity scores can then be used to standardize the available samples with respect to that specific target population. This target population could for instance be captured by a specific data source (eg, the development sample of the prediction model), or represent an amalgamation of multiple data sources. By adopting propensity-score methods to standardize the available validation samples, we can facilitate the interpretation of a particular model's discriminative ability in the intended target population, even when those validation samples do not fully reflect the target population of interest. In other words, we use propensity scores to make the validation sample(s) exchangeable with a sample (or multiple) from the target population.

2.1 | Standardization to membership propensity scores

For individual i, we define the membership propensity score, $m_{S_i}(j)$, as the conditional probability of being member of study sample $j, j = 1, \ldots, J$:

$$m_{S_i}(j) = \Pr(S_i = j|\boldsymbol{X}_i), \tag{1}$$

where S_i is a random variable denoting the study sample of individual i, and $\boldsymbol{X}_i = (X_1, \ldots, X_R)$ denotes a vector that contains the individual's values for predictor $r, r = 1, \ldots, R$. In an analysis of J study samples, we have $S_i \in (1, 2, \ldots, J)$.

Additionally, let s_i denote the study sample to which individual i actually belongs, such that the propensity score $m_{S_i}(S_i = s_i)$ quantifies the conditional probability of individual i being member of its originating sample. Further, in all cases, by definition, $\sum_j m_{S_i}(j) = 1$, for each i. As a result, taking sample J as the reference sample, the propensity score $m_{S_i}(j)$ can be estimated by a multinomial logistic regression model:

$$m_{S_i}(j) = \begin{cases} \dfrac{\exp(\phi_j + \boldsymbol{\gamma}_j' \boldsymbol{X}_i)}{1 + \sum_{h=1}^{J-1} \exp(\phi_h + \boldsymbol{\gamma}_h' \boldsymbol{X}_i)} & \text{for } j \neq J, \\ 1 / \left(1 + \sum_{h=1}^{J-1} \exp(\phi_h + \boldsymbol{\gamma}_h' \boldsymbol{X}_i)\right) & \text{for } j = J, \end{cases} \tag{2}$$

where $\boldsymbol{\gamma}_j = (\gamma_{j1}, \ldots, \gamma_{jR})'$ denotes the coefficients for $X_1, \ldots, X_R$ in the jth linear predictor. In the presence of only two study samples, that is, $J = 2$, Equation (2) reduces to the binary logistic regression model.

It may be common that the differences between study samples are non-linear (on the scale of the linear predictor) in nature. For instance, suppose we have multiple observational study samples of participants of all ages and one sample from a randomized trial that includes only middle-aged adults. A logistic model with a linear term for age would not be able to discriminate between participants from the trial vs observational studies, as both the high and low values for age would be indicative for membership of an observational study. To capture this association, a nonlinear term would be necessary.[35] As the true differences between the validation sample and the target population can be highly nonlinear, we advocate that a flexible modeling approach is used, for instance by the use of spline functions, piecewise polynomials or artificial neural networks.[36] Here, we model:

$$m_{S_i}(j) = \begin{cases} \dfrac{\exp(\phi_j + \sum_{r=1}^{R} \sum_{m=1}^{M} \gamma_{jrm} f_{rm}(X_{ir}))}{1 + \sum_{h=1}^{J-1} \exp\left(\phi_h + \sum_{r=1}^{R} \sum_{m=1}^{M} \gamma_{hrm} f_{rm}(X_{ir})\right)} & \text{for } j \neq J, \\ 1 / \left(1 + \sum_{h=1}^{J-1} \exp\left(\phi_h + \sum_{r=1}^{R} \sum_{m=1}^{M} \gamma_{hrm} f_{rm}(X_{ir})\right)\right) & \text{for } j = J, \end{cases} \tag{3}$$

where $f_{rm}(.)$ represents $m, m = 1, \ldots, M$, non-linear functions. In a way akin to the use of propensity scores for causal inference,[25] the membership propensity score $m_{s_i}(j)$ can subsequently be used to construct the standardization weight with respect to sample j:

$$w_i(j, s_i) = \frac{m_{S_i}(j)}{m_{S_i}(s_i)}. \tag{4}$$

For instance, consider that a prediction model was developed using data from an observational study (sample b) and is subsequently validated using EHR data from a certain hospital (sample a). Although both the development and validation sample may contain individuals from the model's target population, it may occur that the validation sample captures a population with a wider age range, or that the validation sample contains proportionally fewer children, or any other subgroup. Then, the (case-mix) diversity of the target population as reflected by the original development sample may not be well represented by the validation sample. Estimates of prediction model discriminative ability from validation sample a may therefore be misleading if no account is made for the case-mix differences with respect to sample b from the target population. For this reason, we can standardize the distribution of sample a with respect to sample b, such that the weighted validation sample a better represents the target population of sample b. Thereby, this enhances the interpretation of subsequent discriminative ability estimates from the validation sample. For individuals from sample a, we assign the following weights:

$$w_i(b, s_i = a) = \frac{m_{S_i}(b)}{m_{S_i}(a)}. \tag{5}$$

Conversely, for individuals from sample b, the weights are defined as:

$$w_i(b, s_i = b) = \frac{m_{S_i}(b)}{m_{S_i}(b)} = 1. \tag{6}$$

These weights are derived from standardization weighting methods described in the causal inference literature, commonly referred to as "inverse probability weighting" or "standardized mortality ratio" weighting.[37,38] (In inverse probability weighting, the numerator is the propensity of belonging to any of the studied samples; that is, the numerator equals $\sum_j m_{S_i}(j) = 1$ instead of $m_{S_i}(j)$.)

We note that this method requires that the sample that is to be standardized fully captures the domain of the distribution of participant characteristics in the target population with non-zero probability. Propensity score methods are used to weight observed samples. As such, they can down-weight overrepresented groups and up-weight underrepresented groups. However, if a certain subgroup is not represented in the validation sample (hence violating the positivity assumption as referred to in the propensity score literature), it cannot be up-weighted.

Further, the proposed method may also be useful when multiple validation samples are available, and one sample originates from the target population. The other validation samples can then be standardized with respect to the sample from the target population.

2.2 | Validation of prediction models in standardized samples

After the weights are defined, propensity score methods can be used to standardize the discriminative ability estimates in external validation samples with respect to the sample of the target population (eg, the original development sample). By removing the difference in case-mix, this approach may help to interpret estimates of prediction model discrimination in external validation studies with respect to the original development sample. In other words, prediction model discrimination is adjusted for differences in case-mix, which may help to identify causes of poor transportability that cannot directly be attributed to case-mix differences, such as model coefficients that do not generalize to different settings. For instance, when data from a randomized controlled trial (RCT) are available for validating an existing model that was developed in a data from an observational study, it may be more difficult to discriminate between trial patients with and without the outcome due to the stricter inclusion criteria and thus reduced case-mix variability.[21,39] The estimated discriminative ability in the RCT data would then be a biased estimate of the discriminative ability of the model in the model's intended target population. Propensity score methods may help to appreciate and even alleviate this issue by standardizing the validation samples according to the case-mix distribution of the targeted population. Then, any remaining differences (beyond chance) in discriminative ability estimates across the validation samples can be attributed to reasons other than case-mix variation. In the cases of our applied example and simulation study, the case-mix distribution of the target population is represented by the development sample, but does not need to be the case. For instance, in an IPDMA the target population may be captured by one or more validation study samples. If the samples (say b, c, and d) of the target population are truly different, this would require one to define a mixture distribution and to construct the sample for the target population by merging individuals from these samples b, c, and d with a weight that is a function of their propensity to belong to the target population.

Below, we describe how measures of prediction model concordance can be standardized with respect to differences in case-mix between samples. We use the original development sample as the target of standardization, such that any remaining discriminative ability differences between the development and validation sample can be interpreted as a consequence of model coefficients that do not generalize to the population from which the validation sample is drawn (and therefore a lack of model transportability).

2.2.1 | Standardized concordance statistic

Discrimination can be assessed with the concordance (c)-statistic (or area under the receiver operating curve, AUC). For a randomly selected patient $i_+, i_+ \in (1, \ldots, N_+)$, with the outcome and a randomly selected patient $i_-, i_- \in (1, \ldots, N_-)$, without the outcome, the c-statistic estimates the probability that patient i_+ has the highest predicted probability p_{i+} of the outcome. The c-statistic for binary outcomes can be described as:

$$c = \frac{1}{N_+ N_-} \sum_{i_+=1}^{N_+} \sum_{i_-=1}^{N_-} I(p_{i+} > p_{i-}), \tag{7}$$

where $I(p_{i+} > p_{i-})$ is an indicator function that takes the value 1 if $p_{i+} > p_{i-}$ and 0 in all other cases. Optionally, it may take the value of 0.5 if $p_{i+} = p_{i-}$, such that no excessive penalty is given to ties.

We propose to apply a weighting procedure to the c-statistic, similar to precedents.[40,41] We propose to define weights of concordant pairs according to the pairs' propensity scores for the target population. Assuming independence between members of a same pair, the propensity score of a pair is equal to the product of the propensity scores of the members of the pair. Accordingly, the weight of the pair is equal to the product of the weights of the members of the pair. Then, the standardized c-statistic is given by:

$$c_s = \frac{1}{W}\sum_{i_+=1}^{N_+}\sum_{i_-=1}^{N_-} I(p_{i+} > p_{i-})w_{i+}w_{i-}, \tag{8}$$

where $\sum_{i_+=1}^{N_+}\sum_{i_-=1}^{N_-} w_{i+}w_{i-} = W$ denotes the sum of all weights such that the standardized c-statistic is bounded between 0 and 1. Note that $\frac{1}{W} = \frac{1}{N_+N_-}$ when all weights equal 1, and Equation (8) reduces to Equation (7).

Alternatively, the standardized c-statistic may be obtained by the bootstrap. The individuals of the validation sample are then sampled with replacement with probability equal to their respective weights (rescaled to range from 0 to 1) and the (unstandardized) c-statistic is estimated on the resulting samples. The center and percentiles of the resulting propensity weighted distribution of c-statistics then estimate the standardized c-statistic and its confidence interval, respectively, similar to the percentile method for the bootstrap estimation of the unstandardized confidence interval.[42]

The R code for estimating the weighted concordance is available as an R package on Github (https://github.com/VMTdeJong/wAUC). In the next section, we present an applied example, in which we estimate the standardized concordance metric for existing models at multiple external validations.

3 | APPLIED EXAMPLE: DIAGNOSIS OF DEEP VEIN THROMBOSIS

DVT increases a patient's risk of post-thrombotic syndrome and pulmonary embolism, which can be fatal.[43] In DVT suspected patients, often no DVT is present on advanced reference testing.[44] We here consider for illustrative purposes the development and validation of eight different prediction models that could help in the diagnosis of DVT in patients that are suspected of having DVT and use the IPD of 10 002 patients, of whom 1864 had DVT, from thirteen different cross-sectional diagnostic studies across multiple countries.[45] In this example, the prediction model was developed on a sample from the intended target population, as is generally recommended.[14] Conversely, samples used for external validation are more often convenience samples: samples that have much in common with the target population but do not fully reflect it and happen to be available. In this example, the development and validation samples had similar participant characteristics on average (Table 1). For instance, in the development data 22% of the participants had DVT, whereas 18% of the participants did in the pooled validation samples. However, there was major heterogeneity across the validation

TABLE 1 Clinical characteristics of development and validation data for a model for diagnosing DVT.

	Development		Validation			
Variable	**No**	**Yes**	**No**	**Yes**	**Min**	**Max**
DVT	1006	289 (22%)	7132	1575 (18%)	5%	39%
D-dimer abnormal	398	897 (69%)	3629	5078 (58%)	39%	82%
Calf difference > = 3 cm	739	556 (43%)	6248	2459 (28%)	15%	43%
Oral contraceptive (OC)	1167	128 (10%)	8166	541 (6%)	0%	23%
Male	828	467 (36%)	5327	3380(39%)	33%	48%
Absence of leg trauma	197	1098 (85%)	1587	7120 (82%)	67%	95%
Vein distension	1038	257 (20%)	7748	959 (11%)	0%	20%
Presence of malignancy	1214	81 (6%)	7954	753 (9%)	4%	18%
Recent surgery or bedridden	1114	181 (14%)	7777	930 (11%)	5%	18%

Note: Validation, left: Count (%) for all validation data combined. Validation, right: lowest and highest frequencies of the presence of each respective predictor in the different validation data sets.

TABLE 2 Coefficients of eight prediction models for diagnosing DVT.

Model	Intercept	D-dimer	Calf diff > 3	OC	Male	No trauma	Vein	Malignancy	Surgery
1	−3.39	2.58							
2	−3.84	2.42	1.11						
3	−3.90	2.44	1.13	0.40					
4	−4.25	2.46	1.15	0.72	0.72				
5	−4.87	2.49	1.17	0.72	0.73	0.68			
6	−4.95	2.47	1.16	0.70	0.72	0.66	0.52		
7	−4.93	2.44	1.14	0.72	0.70	0.64	0.52	0.53	
8	−5.02	2.43	1.15	0.76	0.71	0.67	0.53	0.50	0.42

Note: Empty cells indicate the coefficient for the respective predictor is assumed zero. For predictor definitions see Table 1.

data sets: in one study only 5% had DVT and in another as many as 39% did. Similarly, somewhat more participants in the development data (69%) had an abnormal D-dimer value than in the pooled validation data (58%), but in the individual validation data sets this ranged from 39% to 82%. Clearly, some of the validation data sets are not representative of the target population from which the development data are sampled, which may lead to worse discriminative ability in these validation data sets than had they all been representative of the target population. As the validation data as a whole might represent the target population quite well, one may not expect poor discriminative ability on average but heterogeneous results across the different validation sets.

3.1 | Methods

The data from one study in the target population (referred to as "development study") were used to develop eight logistic regression prediction models for the probability that DVT is present. We considered that these models were available from the published literature, and used the 12 remaining data sets to externally validate them. This was used to mimic a situation where existing models are already available, and the IPD from their development sample can be obtained. Coefficients for eight prespecified predictors were estimated: positive d-dimer test, calf difference >3 cm (not available as continuous), oral contraceptive usage, male sex, no presence of leg trauma, vein distension, active malignancy, and recent surgery (Table 2). Prediction model 1 consisted of only the first predictor and prediction model 2 consisted of the first two and so forth. Our aim was to investigate to what extent these 8 models generalized well across the 12 external validation samples and to what extent variability in their concordance could be attributed to lack of transportability or rather to case-mix heterogeneity.

We externally validated the discriminative ability of the 8 prediction models in each of the 12 external validation studies, by estimating for each prediction model the traditional unstandardized c-statistic and our proposed standardized c-statistic described in Section 2.2. The propensity model for the standardized c-statistic was a multinomial logistic regression model with linear terms for DVT and all predictors, where the endpoint was the study identifier. The development sample was taken as the target population. This reflects the situation where the inclusion and exclusion criteria for the development sample are defined such that this sample reflects the population in which the prediction model is to be used, as is generally recommended.[14]

To summarize these models' estimated discrimination across the 12 validation studies and to investigate their generalizability across the different settings and populations,[46,47] we then applied random-effects meta-analysis:

$$\begin{aligned} \hat{\theta}_j &\sim \mathcal{N}\left(\theta_j, V(\hat{\theta}_j)\right), \\ \theta_j &\sim \mathcal{N}\left(\theta_{\mathrm{RE}}, \tau^2\right), \end{aligned} \tag{9}$$

where $\hat{\theta}_j$ is the logit concordance estimated in validation sample j, with estimated variance $V(\hat{\theta}_j)$, θ_{RE} is the summary parameter of mean logit discrimination and τ^2 the heterogeneity of true logit concordance values across validation samples. The concordance estimates were first converted to the logit scale to satisfy the normality assumption of the random

effects meta-analysis, as recommended.[48] The meta-analyses were estimated with REML and 95% confidence intervals were estimated by the Hartung-Knapp-Sidik-Jonkman method.[49,50] Approximate 95% prediction intervals for the discriminative ability in a new study were constructed using the approach of Higgins et al, using the t-distribution with 10 degrees of freedom.[51] The confidence intervals for the propensity-weighted c-statistic were obtained with 5000 resamples with replacement, with sampling probabilities equal to the weights as defined in Section 2.

3.2 | Results

The meta-analysis summary estimates indicate that, as expected, discriminative ability was greater for those prediction models that included more predictors. In particular, the pooled c-statistic for the prediction model that only included D-dimer as predictor was 0.70 (95% CI from 0.66 to 0.74), whereas the prediction model with 8 predictors yielded a pooled c-statistic of 0.82 (95% CI from 0.80 to 0.84). Further, we found that summary estimates for the c-statistic that were obtained via propensity standardization did not differ much from the crude (ie, non-standardized) summary estimates (Table S1 in the online Supporting Information, Figure 1). This similarity in the estimated center of the distribution of discriminative ability measures implies that in this applied example, the propensity score method did not reveal any bias due to sampling from populations different from the target population.

However, we found that the discriminative ability estimates in the external validation samples were prone to substantial between-study heterogeneity. The approximate 95% prediction interval for the pooled c-statistic for model 1 ranged from 0.55 to 0.82, indicating that the model does not transport well to some of the non-target populations. However, when standardizing the validation studies, the heterogeneity estimate $\hat{\tau}$ for model 1 decreased from 0.29 to 0.08 and the 95% prediction interval became much more narrow (0.64-0.72). These additional results reveal that between-study heterogeneity in the discriminative ability of prediction model 1 can partially be attributed to differences in case-mix, rather than the use of non-generalizable model coefficients.

The inclusion of additional predictors reduced the heterogeneity of the unstandardized c-statistic. For instance, for prediction model 8, we found $\hat{\tau} = 0.17$ for the unstandardized c-statistic. On the other hand, the heterogeneity of the standardized c-statistic was greater for models that included more predictors. For instance, for prediction model 8 we found $\hat{\tau} = 0.28$ for the standardized c-statistic. Hence, in the unstandardized validation samples the heterogeneity of the c-statistic was lower for models that included more predictors, indicating that the additional predictors corrected for

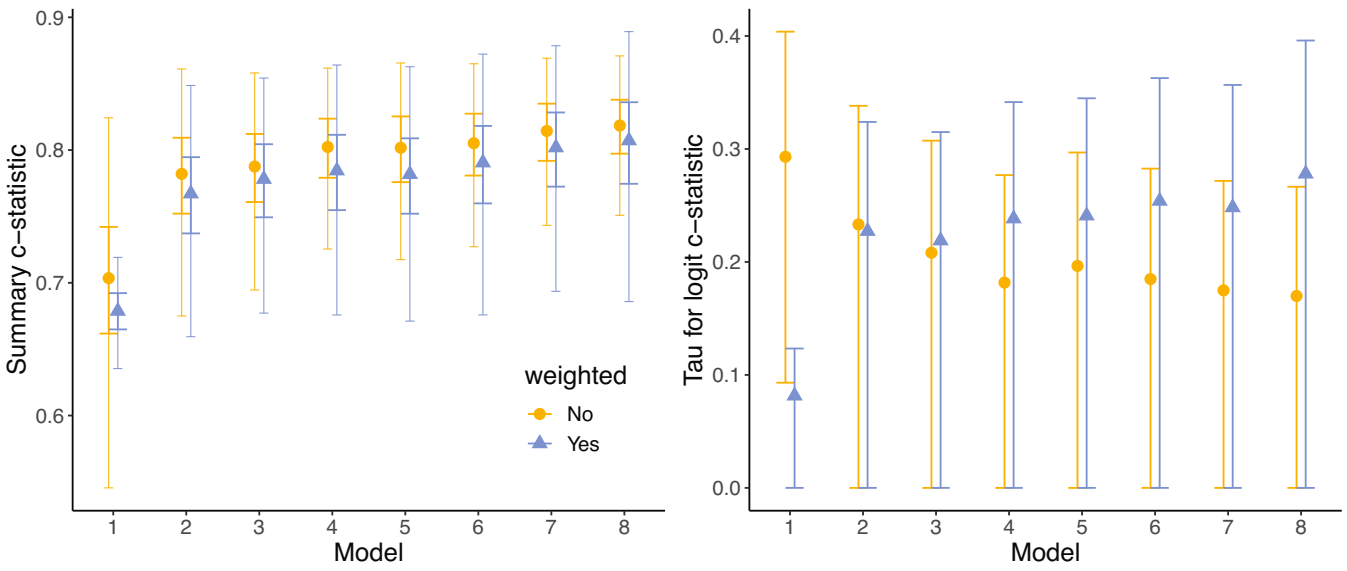

FIGURE 1 Summary (95% CI and 95% approximate PI), and heterogeneity estimates (95% CI) of the concordance in 12 external validation samples before and after weighting. Summary and heterogeneity values are respectively estimated as the inverse logit of summary parameter θ_{RE} and as the standard deviation parameter τ in the random effects meta-analysis of the logit c-statistics estimated in the external validation samples. CI, confidence interval; Approximate PI, approximate prediction interval.

differences in case-mix. In the standardized validation samples the heterogeneity of the c-statistic was greater for models that included more predictors, indicating that for the patients in the validation samples who had predictor variables similar to those in the development data there was greater variability in the discriminative performance.

In conclusion, the standardized prediction model concordance measures disentangled the effect of differences in case-mix and the generalizability of predictor-outcome associations between the development and validation samples. In the following section, we assess the statistical properties of the proposed methods through a simulation study.

4 | SIMULATIONS

In this simulation, we aim to study the validity of different methods to estimate the discriminative ability of a prediction model in a certain target population, when validation data from a population with a (somewhat) different case-mix is available. The simulation is split into two sections. In Section 4.1, we aim to show the impact of different deviations in case-mix, whereas in Section 4.2 we investigate which variables should be included in the propensity model. In both sets of simulations we investigate the requirement to include non-linear effects.

4.1 | Methods simulations set A

We investigated the effect of changes in the predictor distribution (or case-mix) on the estimated concordance in an external validation set. We changed the predictor distribution of the validation sample with respect to the development sample, which resembled a sample from the target population. The simulation was performed in four stages:

1. Draw a development sample from the target population and develop one prediction model on this sample.
2. Draw a validation sample and develop the propensity models on both samples combined.
3. Estimate the propensity-weighted and unweighted concordances in the validation sample.
4. Draw a reference sample from the target population and estimate concordance for the prediction model on this sample.

We drew samples as follows:

$$\begin{aligned} X_b &\sim Bernoulli(\pi), \\ X_c &\sim N(\mu, \sigma^2), \\ Y &\sim Bernoulli(g(-2 + \beta_b X_b + \beta_c X_c)), \end{aligned} \tag{10}$$

where $g(z) = 1/(1 + e^{-z})$). For the development and reference samples we fixed the following parameters: $\pi = 0.2$, $\mu = 0$, $\sigma = 1$, $\beta_b = 1$, and $\beta_c = 1$. This reference sample could thus be used to estimate the out-of-sample concordance in the target population. The validation samples were drawn following the same procedure, but we introduced differences in case-mix to simulate between-study heterogeneity. We investigated six scenarios in which the validation sample predictor distribution was different from the target population and one null scenario in which the validation sample originated from the target population. The predictor distributions for the validation sample were as follows as compared to the target population:

1. No difference; the development and validation samples were sampled from the same distribution.
2. The standard deviation σ of the continuous predictor decreased to 0.8.
3. The standard deviation σ of the continuous predictor increased to 1.2.
4. The prevalence π of the binary predictor was decreased to 0.1.
5. The prevalence π of the binary predictor was increased to 0.4.
6. The mean μ of the continuous predictor decreased to -0.4.
7. The mean μ of the continuous predictor increased to 0.4.
8. The continuous predictor was truncated at $[-1, 1]$.
9. The continuous predictor in the validation data still followed Equation (10), but now the continuous predictor in the development sample was truncated at $[1, 1]$.

We applied two binary logistic propensity models with different parameter specifications. The first estimated the probability of study membership by a linear combination of the predictors (Equation 2), whereas the second included natural

cubic splines for the continuous predictor (Equation 3). Hence, we estimated two standardized estimates and one unstandardized estimate of the concordance in each simulation iteration. We evaluated these estimates by comparing them with the observed concordance in the reference sample and computing bias and mean square error (MSE). Each scenario was repeated for 2000 iterations, for two sample sizes of the development sample: $n = 500$ and $n = 2000$. The validation sample always contained 2000 observations and the reference sample 10 000.

4.2 | Methods simulations set B

In additional simulations, in an extension of scenario 1 (decreased σ), we investigated whether an additional variable, W, that is not part of the prediction model but that has been measured in each data set should be included in the propensity model, when trying to standardize the validation set towards a target population. We consider three sampling mechanisms for W and X_c:

$$\begin{aligned} W &\sim N(0,1) \\ X_c &\sim N(\mu,1), \end{aligned} \tag{11}$$

$$\begin{aligned} W &\sim N(0,1) \\ X_c &\sim N(W+\mu,1), \end{aligned} \tag{12}$$

$$\begin{aligned} X_c &\sim N(\mu,1), \\ W &\sim N(X_c,1). \end{aligned} \tag{13}$$

For the association between Y and W, we now allow Y to be dependent on W as well:

$$Y \sim Bernoulli(g(-2+\beta_b X_b+\beta_c X_c+\beta_w W)), \tag{14}$$

This gives six combinations of associations for X_c, W, and Y, which we investigated in the following simulation scenarios:

1. $\beta_w = 0$ (ie, no association), sampling of W and X_c according to Equation (11).
2. $\beta_w = 0$ (ie, no association), sampling of W and X_c according to Equation (12).
3. $\beta_w = 0$ (ie, no association), sampling of W and X_c according to Equation (13).
4. $\beta_w = 1$, sampling of W and X_c according to Equation (11).
5. $\beta_w = 1$, sampling of W and X_c according to Equation (12).
6. $\beta_w = 1$, sampling of W and X_c according to Equation (13).

We applied four binary logistic propensity models with different parameter specifications. The first estimated the probability of study membership by a linear combination of the predictors in the prediction model (Equation 2), whereas the second the variables in the prediction model but also the covariate W that was not part of the prediction model. The third included natural cubic splines for the continuous predictor in the prediction model (Equation 3), and the fourth included natural cubic spline for the covariate W that was not part of the prediction model as well as the continuous variable in the prediction model. Hence, we estimated four standardized estimates and one unstandardized estimate of the concordance in each simulation iteration. The sample sizes were the same as in simulation set A. The R code for the simulation study is available on Github (https://github.com/VMTdeJong/wAUC-sim).

4.3 | Results simulation set A

Figure 2 highlights the predictor distributions in the first repetition for three scenarios where the distribution of one of the predictors was altered in the validation sample. Consequently, the unweighted distribution of the corresponding predictor in the validation data did not match the target distribution. In scenario #4 (top), where the prevalence of the binary predictor was increased in the validation data, both propensity methods appropriately weighted the distribution of the validation sample, leading to a weighted distribution of the binary predictor to be similar to the target distribution. Also in scenario #6 (middle), where the mean of the continuous predictor was increased in the validation data, both propensity

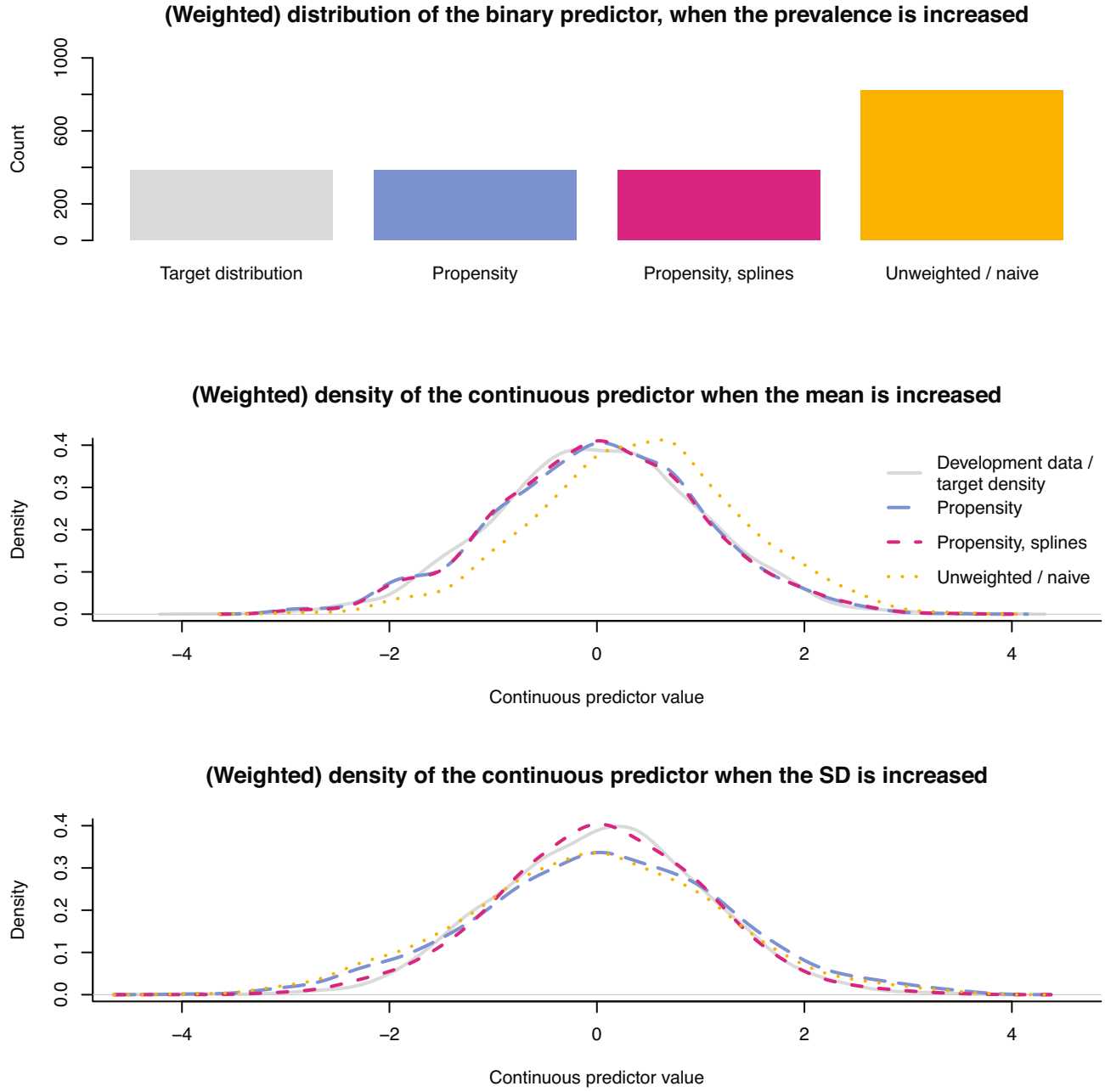

FIGURE 2 Distributions of predictors before and after weighting. Top: Scenario #4, where the prevalence of the binary predictor was increased in the validation data. Middle: Scenario #6, where the mean of the continuous predictor was increased in the validation data. Bottom: Scenario #2, where the standard deviation of the continuous predictor was increased in the validation data. The development sample size was 2000 in each of these plots.

methods appropriately weighted the density of the continuous predictor. However, in scenario #2 (bottom), where the standard deviation of the continuous predictor was increased in the validation data, only the propensity method that used splines appropriately weighted the density of the continuous predictor. The density as weighted by the propensity method with linear terms was nearly identical to the density of the original (unweighted) validation sample.

4.3.1 | Bias

Bias was either positive, negative or (near-)zero, depending on the scenario and method (Figure 3). In scenario 0, where the validation sample originated from the target population, all methods had near-zero bias. But in scenario's 1 and 2,

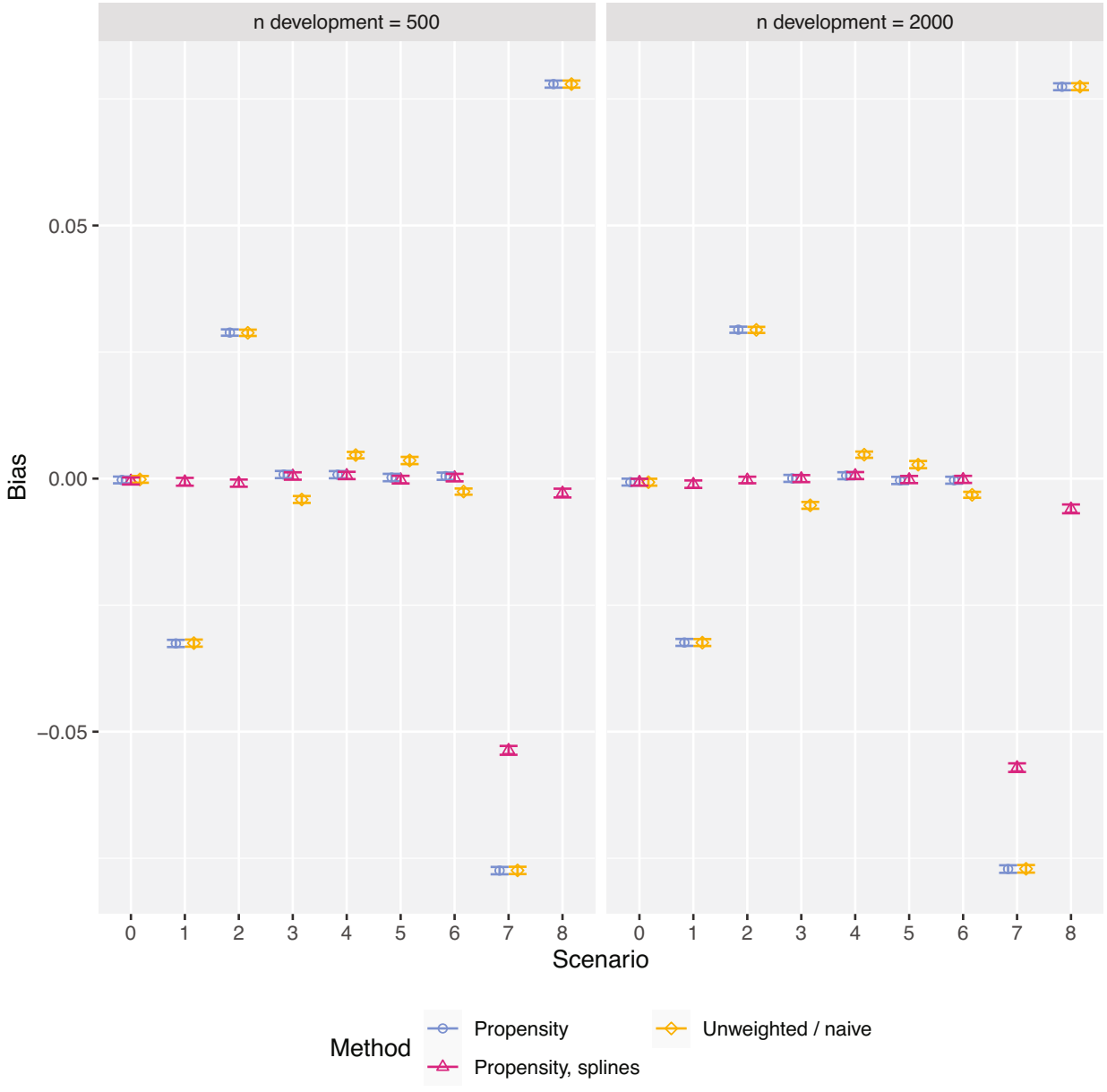

FIGURE 3 Bias and 95% CI of the estimated concordance before and after weighting. Concordance was estimated in the (weighted) external validation data for all methods. Bias was subsequently estimated by subtracting the estimate in the reference sample. 95% confidence intervals (CI) were estimated as 2.5% and 97.5% percentiles of 5000 bootstrap samples.

where the standard deviation of the continuous predictor was changed, both the propensity method with linear terms and the unweighted method were negatively and positively biased respectively, whereas the spline-based propensity method had near-zero bias. In scenario 3 and 4, where the prevalence of the binary predictor was changed, both propensity methods were nearly unbiased, whereas the unweighted method showed negative and positive bias, respectively. In scenario 5 and 6, where the mean of the continuous predictor was changed, both propensity methods were nearly unbiased, whereas the unweighted method showed positive and negative bias, respectively.

In scenario 7, where continuous predictor was truncated in the validation sample, all methods showed negative bias, but less so for the propensity method with splines. In scenario 8, the unweighted method and the propensity method with linear terms were positively biased, whereas the propensity method with splines was nearly unbiased when the development sample size was 500. Sample size had a minimal impact on the results; only in scenario 8 we see a notable difference. There the propensity method with splines also showed some bias, but this was of a far smaller order of magnitude than the propensity method with linear terms and the unweighted method.

4.3.2 | MSE

In scenario 0, where the validation sample originated from the target population, all methods had low RMSE (Figure 4). In scenario 1 and 2, where the standard deviation of the continuous predictor was changed, the RMSE of the propensity method with linear terms and the unweighted method were far greater than that of the propensity method with splines. In scenario 3 and 4, where the prevalence of the binary predictor was changed, the methods had similar RMSE, though in scenario 4 the unweighted method had slightly lower RMSE when the sample size was small. In scenario 5, where the standard deviation of the continuous predictor was lower in the validation sample, all methods had similar RMSE. In scenario 6, where the standard deviation of the continuous predictor was higher in the validation sample, the unweighted method had slightly lower RMSE.

In scenario 7, where the continuous predictor was truncated in the validation sample, all methods had much higher RMSE than in the other scenarios, though the propensity method with splines had a lower RMSE than the other methods. In scenario 8, where the continuous predictor was truncated in the development sample, the propensity method with

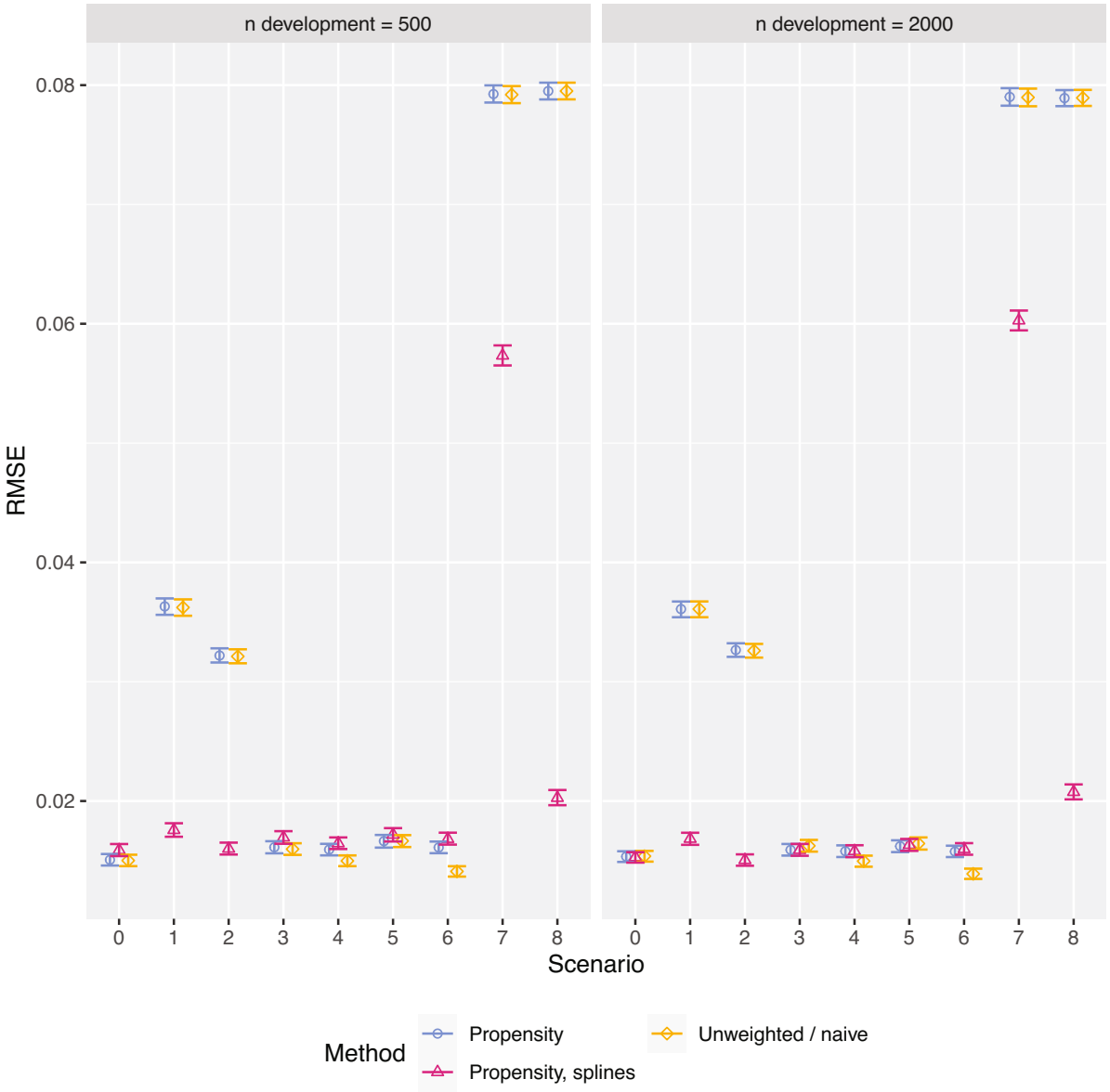

FIGURE 4 Root mean square error and 95% CI of the estimated concordance before and after weighting. Concordance is estimated in the (weighted) external validation data for all methods. RMSE was subsequently estimated using the estimate in the reference sample. 95% confidence intervals (CI) were estimated as 2.5% and 97.5% percentiles of 5000 bootstrap samples.

linear terms and the unweighted method had far greater RMSE than the propensity method that used splines. Sample size of the development set was of very limited importance for RMSE.

4.4 | Results simulation set B

4.4.1 | Bias

In scenario 9 to 12 and 14, where an additional covariate was present in the samples from the target population and the validation sample, the propensity methods with splines were (nearly) unbiased regardless of whether the additional covariate was included in the propensity model, whereas the unweighted approach showed varying amounts of bias (Figure 5). Only in scenario 13, where the predictor X_c was causally affected by the covariate W and β_W was nonzero,

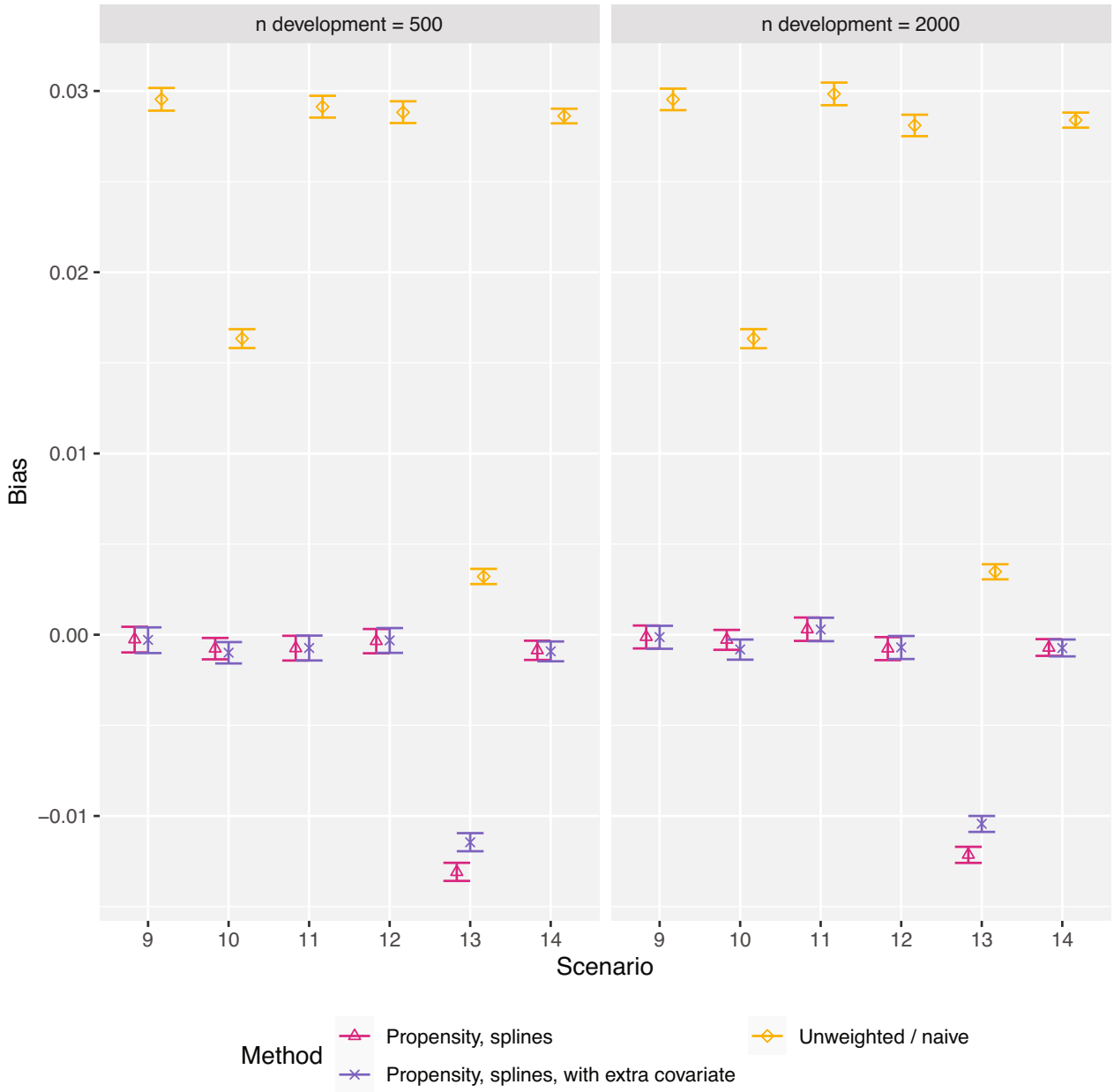

FIGURE 5 Bias and 95% CI of the estimated concordance before and after weighting possibly using an additional variable. Concordance was estimated in the (weighted) external validation data for all methods. Bias was subsequently estimated by subtracting the estimate in the reference sample. 95% confidence intervals (CI) were estimated as 2.5% and 97.5% percentiles of 5000 bootstrap samples.

both propensity methods were biased and adding the extra covariate only slightly reduced this, whereas the unweighted approach showed less bias.

4.4.2 | RMSE

In scenario 9 to 12 and 14, where an additional covariate was present in the samples from the target population and the validation sample, the propensity methods had less RMSE than the unweighted approach, regardless of whether the additional covariate was included in the propensity score model (Figure 6). Only in scenario 13, where the predictor

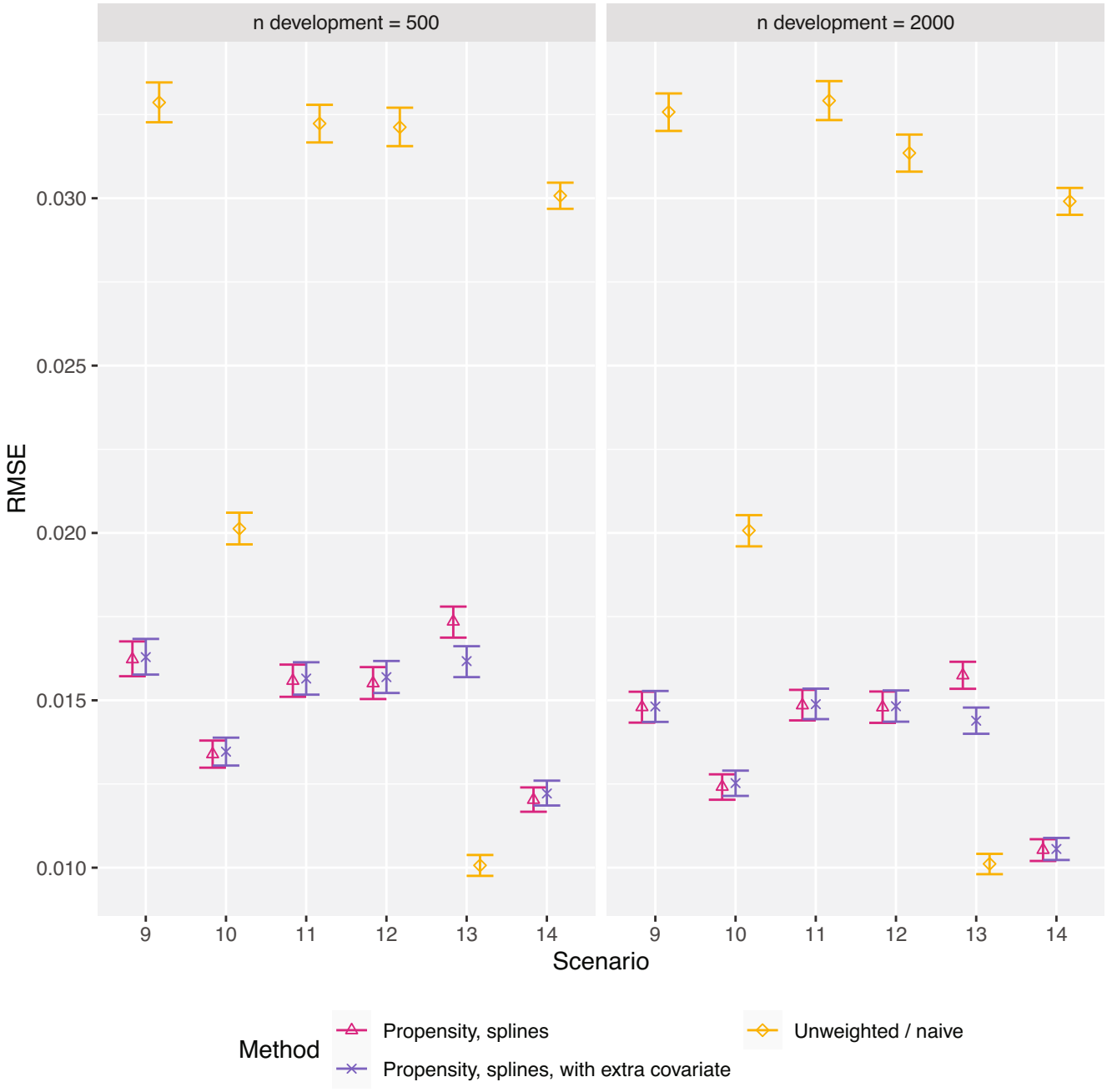

FIGURE 6 Root mean square error and 95% CI of the estimated concordance before and after weighting possibly using an additional variable. Concordance is estimated in the (weighted) external validation data for all methods. RMSE was subsequently estimated using the estimate in the reference sample. 95% confidence intervals (CI) were estimated as 2.5% and 97.5% percentiles of 5000 bootstrap samples.

X_c was causally affected by the covariate W and β_W was nonzero, the propensity methods had greater RMSE than the unweighted approach.

4.5 | Summary

In some scenarios, notably when the standard deviation of a predictor changed or when a predictor was truncated, the unweighted validation approach produced concordance estimates that had large bias and RMSE. The propensity method with linear terms could not mitigate this issue when it occurred, but the spline-based propensity model did in most scenarios. The exception was when the continuous predictor was truncated in the validation sample. Then, the spline-based propensity model only slightly reduced the bias and RMSE.

Also, when a covariate W was present in the development and validation set that was not part of the prediction model, adding this covariate to the propensity model increased bias and RMSE unless the predictor X_c was causally affected by the covariate W and β_W was nonzero. Figures that show concordance as estimated on the development data are provided in the Supporting Information.

5 | DISCUSSION

We proposed a method for standardizing samples in which a prediction model can be validated for a target population. When validating a prediction model in an external validation study, there may be important differences between the case-mix distribution from this validation study and the intended target population and setting. When such differences occur, estimates of model discrimination in the validation study may not be representative of the discrimination in the intended target population and setting. For instance, it may occur that the estimated coefficients of a previously developed prediction model are generalizable to the validation sample, but that a difference in the case-mix distribution with respect to the development sample affects discriminative ability. Any deterioration in estimated discriminative ability, as compared to previous studies, should then not be attributed to the prediction model but to the sample of included participants.

Standardization methods, as shown in this study, facilitate the interpretation of prediction model discrimination differences found in a validation sample with respect to the target population and possibly other validation samples. In particular, by standardizing validation samples with respect to the target population, it becomes possible to remove or reduce the impact of case-mix effects on prediction model discrimination estimates found in the validation samples. In other words, standardization allows one to interpret validation study results as if the case-mix distributions would remain unchanged as compared to the original target population. In our study, we defined the development sample as originating from the target population, and standardized the validation samples towards this sample. In practice, multiple target populations of interest might exist for the same model. Indeed, IPD may even contain representative samples from multiple possible target populations. In such cases, our proposed methods provide the means to estimate discriminative ability in each one at a time, avoiding the mistake of averaging discriminative ability over all samples.

As case-mix differences can be found with regard to many variables (predictors and/or outcome), we propose a multivariable standardization approach, which has originally been described in the causal inference literature to balance covariate distributions across patient settings under different "exposures".[37,38] Transposing this framework to clinical prediction model development and validation research, one can consider the memberships to the development or external validation settings as "exposures". Similar approaches have been suggested to anticipate the external validity of results from RCTs[52] and to use a larger sample size by including propensity weighted external data to assess the intervention effect in a (single) trial.[53]

The results of the simulation study confirm that estimating an unweighted concordance in a sample that does not originate from the intended target population produces estimates of discrimination that are inaccurate and biased for the actual discrimination in a sample from the target setting in most of the scenarios that we investigated. In the scenarios of the simulation study where the standard deviation of the continuous predictor was altered in the validation data, only the propensity method that used splines appropriately weighted the density of the continuous predictor. This is because, in contrast to the scenarios where the prevalence of the binary predictor was changed, the change in the predictor distribution was nonlinear. When the standard deviation was decreased, both very large and very small values of the

continuous predictor indicated that the observation was not typical for the target population and should therefore be down-weighted, and vice versa when the standard deviation was increased. A propensity model with linear terms could not capture such a change in distribution, but one that utilizes splines could. Accordingly, the spline-based propensity model was the only method that produced estimates of the discrimination in the intended target population that were both unbiased and had small error in this scenario.

The proposed method appears particularly useful when the standard deviation of a continuous predictor is different in the validation sample. This may, for instance, be the case when the development data are taken from an RCT in a hospital setting, where strict inclusion criteria are used for participant selection, which leads to narrower distributions of predictors. Often participants with extreme values are excluded. For instance, it is common that both children and the elderly are excluded. If the model is then validated in observational data of the target population, which also includes children and the elderly, then the standard deviation of age will be greater in the validation data than in the development data. This will lead to higher unweighted estimates of discrimination, which would not be representative of the discrimination in the RCT setting. Propensity weighting of the observational data could then be used to adjust for this difference in predictor distributions, and thereby facilitate the interpretation of model discrimination in the RCT setting.

The opposite may be more difficult, however. If a prediction model is validated in a sample with restricted variability (ie, in terms of range) as compared to the development sample, then propensity standardization methods may not always be able to fully mitigate the effects of differing predictor distributions. This is because re-weighting can only alter the frequency distribution on the sampled domain, but it cannot extend the domain. For example, when applying a prediction model that was developed in adults to adolescents or children, differences in age distribution cannot be fully recovered using re-weighting due to a lack of representation in the validation data. We observed this problem in the simulation study when the continuous predictor was truncated in the validation sample (scenario 7). There, neither propensity method could fully correct the concordance estimates for the reduction in heterogeneity of the predictor distribution (though, the propensity method with nonlinear terms removed some of the bias).

We also investigated whether a baseline covariate that was not part of the prediction model but that had been measured in both the sample from the target population and in the validation sample should be included in the propensity model. The results show that such a covariate should not necessarily be included in the propensity model unless this covariate is a cause of both the outcome and a predictor included in the model. In this case, there exists a backdoor linking the predictor and the outcome via that third variable (W). If this variable W also affects the sample selection, along with the predictors and the outcome, then there is likely an additional relationship that is created via collider-stratification when working (restricting) on the study sample, which could explain the bias observed in the scenario where W affects X, and W is predictive of the outcome. This additional relationship could be theoretically removed if one conditioned on W. However, because the predictor and the outcome lie on the path between W and the sample selection, a simple propensity score including both W and the predictor is unlikely to be able to account for this complex structure, thereby not completely eliminating bias. Further studies are needed to explore whether a more thorough structural consideration of the propensity score construction could help eliminate bias.

Although, we observed considerable differences in heterogeneity between the standardized and unstandardized measures of prediction model discriminative ability in the motivating example of 12 external validation studies, the absolute differences in the summary estimates were minor. Overall, the validation samples were not very distinct from the development sample.

In terms of heterogeneity of the prediction models' discriminative ability, we did observe considerable differences after standardization. There was less heterogeneity of discrimination after standardization for the model with one predictor, though, as the number of predictors increased this difference disappeared or even reversed, because the additional predictors' coefficients were not generalizable to patients from all studies when these studies were weighted towards the development data.

5.1 | Limitations and future directions

Standardization using propensity score weighting methods can be performed in different ways.[24,25,37,38] Application of the propensity methods requires access to the validation data and an approximation of the joint density of

the target population. Here we have used a sample of the target population to fulfil the latter requirement, but other approximations might be conceived, such as a multivariate normal approximation or simpler methods such as raking.[54] Such methods may prove useful when development IPD is not available and only summary statistics can be accessed, but would require further research. An alternative approach is possible if the model-based concordance (mbc) has been estimated for the development data. Then, subtracting the mbc at validation would provide an estimate of the difference in discriminative ability resulting from differences in case-mix between the two samples.[23]

In our motivating example, we chose weights that allowed the validation samples to resemble the (single) development sample in terms of case-mix. Another strategy could be to define weights such that the sample to be standardized approximates all available studies or settings taken together (ie, "entire population"), akin to the "inverse probability weighting" described in the causal inference literature.[25] In fact, the choice between standardization weights should be made according to the target population. That is, it should depend on whether the prediction model aims for a specific homogeneous setting or a larger scale population. Further studies are needed to compare these weighting methods. Future research should also take into account other issues that may compromise model transportability, such as measurement error.[5]

Further, propensity scores might also be used to standardize samples for a specific target population during model development on a data set that consists of multiple, combined, data sets. In contrast to the here studied standard dichotomy between development and validation data sets, re-weighting the development data to match a specific target population increases the sample size available for model development in the target population. For instance, in an IPDMA with the aim to develop a prediction model, data from RTCs may be included. Due to strict eligibility criteria, data from these RCTs might not fully match the intended target population. Simply stacking every such available data set for model development purposes would then bias model parameters and deteriorate its predictive discriminative ability. Standardization may then help to estimate model parameters with respect to the target population and to assess its reproducibility in the targeted population. On the other hand, if a certain subgroup is not represented in the validation sample, it cannot be up-weighted, implying that data from an RCT with restrictive inclusion criteria cannot be standardized towards a wider population.

Although, we investigated a range of differences in case-mix between the target population and the validation sample, and we investigated several (possibly causal) relations between a predictor variable and a covariate that was not part of the prediction model in our simulation study, there are (combinations of) scenarios that we did not investigate. For instance, in our simulations the additional covariate was a baseline covariate, but we did not consider covariates or secondary outcomes measured after the measurement of the outcome of interest. Further, if the regression coefficients in the validation sample differ from those in the target population after weighting the validation sample, the predictor samples are not truly exchangeable or the predictor coefficients are different. The propensity score is then not sufficient to standardize the samples.

5.2 | Conclusion

Propensity score-based standardization may be applied to estimate the discriminative ability in the target population, when (some of the) validation samples do not fully reflect the target population. This helps to facilitate the interpretation of (heterogeneity in) prediction model discriminative ability observed in (multiple) validation studies and to guide the need for prediction model updating strategies (in particular the need for model re-estimation) or to accept that the validation sample does not reflect the target population of the developed model. Further research may focus on the use of propensity score weighting during model development on heterogeneous data sets, such as IPDMA or large clustered routine care data sets, to enhance the reproducibility of prediction models.

ACKNOWLEDGEMENTS

We thank all the reviewers for very constructive feedback that has improved the article. This project has received funding from the European Union's Horizon 2020 research and innovation programme under ReCoDID Grant agreement No. 825746.

DATA AVAILABILITY STATEMENT

We gratefully acknowledge the following authors for sharing of individual participant data from the deep vein thrombosis (DVT) studies: G. J. Geersing, N. P. A. Zuithoff, C. Kearon, D. R. Anderson, A. J. ten Cate-Hoek, J. L. Elf,

S. M. Bates, A. W. Hoes, R. A. Kraaijenhagen, R. Oudega, R. E. G. Schutgens, S. M. Stevens, S. C. Woller, P. S. Wells, and K. G. M. Moons. The DVT data that support the findings of this study are not publicly available, according to the conditions determined by the authors of the DVT studies, but are available on request from G. J. Geersing, by e-mailing G.J.Geersing@umcutrecht.nl.

The analyses in the applied example can be reproduced with an R script provided as supplementary online material. As the data are not publicly available, we used the DVTipd dataset in the metamisc R package. As this is a synthetic dataset, the results do not match the results presented in this manuscript.

ORCID

Valentijn M. T. de Jong https://orcid.org/0000-0001-9921-3468
Jeroen Hoogland https://orcid.org/0000-0002-2397-6052
Richard D. Riley https://orcid.org/0000-0001-8699-0735
Tri-Long Nguyen https://orcid.org/0000-0002-6376-7212
Thomas P. A. Debray https://orcid.org/0000-0002-1790-2719

REFERENCES

1. Riley RD, van der Windt D, Croft P, Moons KGM. *Prognosis Research in Healthcare: Concepts, Methods, and Impact*. Oxford: Oxford University Press; 2019.
2. Siontis GCM, Tzoulaki I, Castaldi PJ, Ioannidis JPA. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *J Clin Epidemiol*. 2015;68(1):25-34.
3. 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.
4. Vergouwe Y, Moons KGM, Steyerberg EW. External validity of risk models: use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *Am J Epidemiol*. 2010;172(8):971-980.
5. Pajouheshnia R, van Smeden M, Peelen LM, Groenwold RHH. How variation in predictor measurement affects the discriminative ability and transportability of a prediction model. *J Clin Epidemiol*. 2019;105:136-141.
6. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*. 1996;15(4):361-387.
7. van Smeden M, Moons KGM, de Groot JAH, et al. Sample size for binary logistic prediction models: beyond events per variable criteria. *Stat Methods Med Res*. 2018;28(8):2455-2474.
8. 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.
9. de Jong VMT, Eijkemans MJC, van Calster B, et al. Sample size considerations and predictive performance of multinomial logistic prediction models. *Stat Med*. 2019;38(9):1601-1619.
10. Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med*. 2000;19(4):453-473.
11. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med*. 1999;130(6):515-524.
12. Luijken K, Groenwold RHH, Calster BV, Steyerberg EW, van Smeden M. Impact of predictor measurement heterogeneity across settings on the performance of prediction models: a measurement error perspective. *Stat Med*. 2019;38(18):3444-3459.
13. Charlson ME, Ales KL, Simon R, MacKenzie CR. Why predictive indexes perform less well in validation studies: is it magic or methods? *Arch Intern Med*. 1987;147(12):2155-2161.
14. Moons KGM, 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.
15. Wynants L, Calster BV, Collins GS, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ*. 2020;369:m1328.
16. Riley RD, Tierney JF, Stewart LA. *Individual Participant Data Meta-Analysis: A Handbook for Healthcare Research*. Chichester: Wiley; 2021.
17. Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. *BMJ*. 2009;338:b605.
18. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. New York: Springer Science & Business Media; 2009.
19. Steyerberg EW, Moons KGM, van der Windt DA, et al. Prognosis research strategy (PROGRESS) 3: prognostic model research. *PLoS Med*. 2013;10(2):e1001381.
20. Moons KGM, Kengne AP, Grobbee DE, et al. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart*. 2012;98(9):691-698.
21. Austin PC, Steyerberg EW. Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. *BMC Med Res Methodol*. 2012;12(1):82.

22. van Klaveren D, Gönen M, Steyerberg EW, Vergouwe Y. A new concordance measure for risk prediction models in external validation settings. *Stat Med*. 2016;35(23):4136-4152.
23. van Klaveren D, Steyerberg EW, Gönen M, Vergouwe Y. The calibrated model-based concordance improved assessment of discriminative ability in patient clusters of limited sample size. *Diagn Progn Res*. 2019;3(1):11.
24. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1): 41-55.
25. Rosenbaum PR. Model-based direct adjustment. *J Am Stat Assoc*. 1987;82(398):387-394.
26. 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:i3140.
27. de Jong VMT, Moons KGM, Eijkemans MJC, Riley RD, Debray TPA. Developing more generalizable prediction models from pooled studies and large clustered data sets. *Stat Med*. 2021;40(15):3533-3559.
28. de Jong VMT, Hoogland J, Debray TPA, Nguyen TL. Propensity-based standardization to enhance the interpretation of predictive performance in external validation studies. In: de Jong VMT, ed. *Methods for Individual Participant Data Meta-Analysis in Prediction Research*. Utrecht: Utrecht University; 2020:73-86. doi:10.33540/469
29. Riley RD, Snell KIE, Wynants L, de Jong VMT, Moons KGM, Debray TPA. IPD meta-analysis for clinical prediction model research. In: Riley RD, Tierney JF, Stewart LA, eds. *Individual Participant Data Meta-Analysis: A Handbook for Healthcare Research*. Hoboken, NJ: Wiley Online Library; 2021:447-497.
30. Shmueli G. To explain or to predict? *Stat Sci*. 2010;25(3):289-310.
31. Hernán MA. The C-word: scientific euphemisms do not improve causal inference from observational data. *Am J Public Health*. 2018;108(5):616-619.
32. Nowacki AS, Wells BJ, Yu C, Kattan MW. Adding propensity scores to pure prediction models fails to improve predictive performance. *PeerJ*. 2013;1:e123.
33. Groenwold RHH, Moons KGM, Pajouheshnia R, et al. Explicit inclusion of treatment in prognostic modeling was recommended in observational and randomized settings. *J Clin Epidemiol*. 2016;78:90-100.
34. Pajouheshnia R, Peelen LM, Moons KGM, Reitsma JB, Groenwold RHH. Accounting for treatment use when validating a prognostic model: a simulation study. *BMC Med Res Methodol*. 2017;17(1):103.
35. Drake C. Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics*. 1993;49:1231-1236.
36. Blakely T, Lynch J, Simons K, Bentley R, Rose S. Reflection on modern methods: when worlds collide—prediction, machine learning and causal inference. *Int J Epidemiol*. 2020;49(6):2058-2064.
37. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000;11(5): 550-560.
38. Sato T, Matsuyama Y. Marginal structural models as a tool for standardization. *Epidemiology*. 2003;14(6):680-686.
39. Pajouheshnia R, Groenwold RH, Peelen LM, Reitsma JB, Moons KG. When and how to use data from randomised trials to develop or validate prognostic models. *BMJ*. 2019;365:l2154.
40. Wu S, Flach P. A scored AUC metric for classifier evaluation and selection. Proceedings of the ICML 2005 Workshop on ROC Analysis in Machine Learning; 2005.
41. Li J, Fine JP. Weighted area under the receiver operating characteristic curve and its application to gene selection. *J R Stat Soc Ser C Appl Stat*. 2010;59(4):673-692.
42. Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Stat Med*. 2000;19(9):1141-1164.
43. Qaseem A, Snow V, Barry P, et al. Current diagnosis of venous thromboembolism in primary care: a clinical practice guideline from the American Academy of family physicians and the American College of Physicians. *Ann Intern Med*. 2007;146(6):454-458.
44. Oudega R, Moons KGM, Hoes AW. Ruling out deep venous thrombosis in primary care. *Thromb Haemost*. 2005;94(01):200-205.
45. Geersing GJ, Zuithoff NPA, Kearon C, et al. Exclusion of deep vein thrombosis using the Wells rule in clinically important subgroups: individual patient data meta-analysis. *BMJ*. 2014;348:g1340.
46. Debray TPA, Damen JAAG, Snell KIE, et al. A guide to systematic review and meta-analysis of prediction model performance. *BMJ*. 2017;356:i6460.
47. Debray TPA, Damen JAAG, Riley RD, et al. A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes. *Stat Methods Med Res*. 2019;28(9):2768-2786.
48. 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.
49. Knapp G, Hartung J. Improved tests for a random effects meta-regression with a single covariate. *Stat Med*. 2003;22(17): 2693-2710.
50. Sidik K, Jonkman JN. Robust variance estimation for random effects meta-analysis. *Comput Stat Data Anal*. 2006;50(12): 3681-3701.
51. Higgins JPT, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *J R Stat Soc A Stat Soc*. 2009;172(1): 137-159.
52. Stuart EA, Cole SR, Bradshaw CP, Leaf PJ. The use of propensity scores to assess the generalizability of results from randomized trials. *J R Stat Soc A Stat Soc*. 2011;174(2):369-386.

53. Vo T, Porcher R, Chaimani A, Vansteelandt S. A novel approach for identifying and addressing case-mix heterogeneity in individual participant data meta-analysis. *Res Synth Methods*. 2019;10(4):582-596.
54. Deville JC, Särndal CE, Sautory O. Generalized raking procedures in survey sampling. *J Am Stat Assoc*. 1993;88(423):1013-1020. doi:10.1080/01621459.1993.10476369

SUPPORTING INFORMATION

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

How to cite this article: de Jong VMT, Hoogland J, Moons KGM, Riley RD, Nguyen T-L, Debray TPA. Propensity-based standardization to enhance the validation and interpretation of prediction model discrimination for a target population. *Statistics in Medicine*. 2023;42(19):3508-3528. doi: 10.1002/sim.9817