Research Article

Statistics in Medicine

Received 7 April 2011, Accepted 16 March 2012 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5412

Aggregating published prediction models with individual participant data: a comparison of different approaches

Thomas P. A. Debray,[a*†] Hendrik Koffijberg,[a] Yvonne Vergouwe,[b] Karel G. M. Moons[a‡] and Ewout W. Steyerberg[b‡]

During the recent decades, interest in prediction models has substantially increased, but approaches to synthesize evidence from previously developed models have failed to keep pace. This causes researchers to ignore potentially useful past evidence when developing a novel prediction model with individual participant data (IPD) from their population of interest. We aimed to evaluate approaches to aggregate previously published prediction models with new data. We consider the situation that models are reported in the literature with predictors similar to those available in an IPD dataset. We adopt a two-stage method and explore three approaches to calculate a synthesis model, hereby relying on the principles of multivariate meta-analysis. The former approach employs a naive pooling strategy, whereas the latter accounts for within-study and between-study covariance. These approaches are applied to a collection of 15 datasets of patients with traumatic brain injury, and to five previously published models for predicting deep venous thrombosis. Here, we illustrated how the generally unrealistic assumption of consistency in the availability of evidence across included studies can be relaxed. Results from the case studies demonstrate that aggregation yields prediction models with an improved discrimination and calibration in a vast majority of scenarios, and result in equivalent performance (compared with the standard approach) in a small minority of situations. The proposed aggregation approaches are particularly useful when few participant data are at hand. Assessing the degree of heterogeneity between IPD and literature findings remains crucial to determine the optimal approach in aggregating previous evidence into new prediction models. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: prediction research; prediction models; meta-analysis; logistic regression; multivariable; Bayesian inference

1. Introduction

It is well known that many prediction models do not generalize well across patient populations [1–6]. This quandary may occur, for example, when prediction models are developed from small data sets, when too many predictors were studied compared with the effective sample size, or when the population in which the model is validated or applied diverges (substantially) from the population where the model was developed. Although the use of larger datasets for model development covers a straightforward solution, in practice this option is frequently not possible owing to, for example, cost constraints, ethical considerations or inclusion problems.

It is remarkable that despite the scarcity of individual participant data (IPD), there is an abundance of prediction models in the medical literature, even for the same clinical problem. For example, there are over 60 published models aiming to predict outcome after breast cancer [7, 8], over 25 for predicting

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

[b] *Center for Medical Decision Sciences, Department of Public Health, Erasmus Medical Center Rotterdam, Rotterdam, The Netherlands*

[*] *Correspondence to: Thomas P. A. Debray, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Stratenum 6.131, P.O. Box 85500, 3508GA Utrecht, The Netherlands.*

[†] *E-mail: T.Debray@umcutrecht.nl*

[‡] *Equal contribution*

long-term outcome in neurotrauma patients [9], and about 10 to diagnose venous thromboembolism. This dispersion of information reduces the scientific and clinical utility of prognostic research overall. Prior knowledge from previous research goes unused and clinicians are left to pick from a cacophony of unreliable prognostic models with limited scope. This is undesirable for all parties involved.

Conceptually, combining prior knowledge from multiple studies is already widespread in etiologic and intervention research, in the form of meta-analyses [10]. More elaborate approaches, for example, for synthesizing the accuracy of diagnostic tests [11], have also recently emerged but remain largely lacking in prediction research, despite the fact that the potential gains are arguably even greater [12]. The closest existing equivalent techniques focus upon updating of existing prediction models that are being applied to a different setting [3, 5, 13–15]. Approaches for using prior knowledge in prediction research are underdeveloped [12]. Some published approaches rely on evidence that is typically not published, such as covariance matrices or regression coefficients, or lack a formal statistical foundation [16, 17].

We aimed to investigate how previously published prediction models or studies can be used in the development of a (new) prediction model when published models and the IPD incorporate similar predictors. We realize that published prediction models often differ in their composition through the inclusion of different covariates in the models, the transformations and coding applied, and adjustment for overfitting [18, 19]. We here assume, as a start, that identical model formulations are available for the published prediction models.

We adopt the two-stage method proposed by Riley *et al.* [20] and explore three approaches to aggregate the published prediction models (with similar predictors) with IPD. These approaches reduce the available IPD to aggregate data (AD), and combine this evidence with the AD from the literature (i.e., the published prediction models). The first two approaches calculate an overall synthesis model, whereas the third approach employs a Bayesian perspective to adapt the coefficients of previously published prediction models with the IPD at hand. The approaches are evaluated here through testing the predictive performance of prediction models for 6-month outcome in 15 traumatic brain injury (TBI) datasets [21, 22]. In addition, we illustrate their application in a genuine example involving the prediction of deep vein thrombosis (DVT).

2. Methods

We consider the situation in which an individual participant dataset (IPD) as well as a number of previously published multivariate logistic regression models are available. The IPD is described by $i = 1, \ldots, K$ independent predictors, a dichotomous outcome, and contains N_{IPD} subjects. The characteristics and observed outcome of subject $s = 1, \ldots, N_{\text{IPD}}$ in these data are denoted as $x_{s1}, \ldots x_{sK}$ and y_s, respectively. The AD from the literature studies are represented by the published prediction models, and can be obtained from individual study publications or directly from the study authors themselves. We assume that the literature models have a similar set of predictors as the IPD, and were developed with a similar prediction task in mind. Furthermore, we assume that for each of $j = 1, \ldots, M$ previously published prediction models, the estimated regression coefficients $\hat{\beta}_{0j}, \ldots, \hat{\beta}_{Kj}$ and their corresponding standard errors $\hat{\sigma}_{0j}, \ldots, \hat{\sigma}_{Kj}$ are available. The regression coefficients obtained from the IPD are denoted as $\hat{\beta}_{1,\text{IPD}}, \ldots, \hat{\beta}_{K,\text{IPD}}$ (with intercept $\hat{\beta}_{0,\text{IPD}}$) and their respective variance–covariance matrix as $\hat{\Sigma}_{\text{IPD}}$. Although we focus on the presence of one IPD, it is possible to add additional IPDs in a similar manner.

From this situation, we propose three approaches to then combine the literature models with the IPD and derive a novel, aggregated prediction model with coefficients $\beta_{0,\text{UPD}}, \ldots, \beta_{K,\text{UPD}}$ and variance–covariance matrix Σ_{UPD} (with variance elements $\sigma^2_{0,\text{UPD}}, \ldots, \sigma^2_{K,\text{UPD}}$ where UPD stands for “updated”). These approaches adopt the two-stage method described by Riley *et al.* [20], where the available IPD are reduced to AD, and then combined with existing AD using meta-analytical techniques. Specifically, the IPD is first reduced to $\hat{\beta}_{0,\text{IPD}}, \ldots, \hat{\beta}_{K,\text{IPD}}$ and $\hat{\Sigma}_{\text{IPD}}$, and then aggregated with $\hat{\beta}_{0j}, \ldots, \hat{\beta}_{Kj}$ and $\hat{\sigma}_{0j}, \ldots, \hat{\sigma}_{Kj}$ using meta-analysis techniques appropriate for multivariate synthesis. The first two approaches derive an average synthesis model across the included study populations, which may not be relevant to the population of interest. For this reason, the third approach assumes that the IPD reflects the clinically relevant population, and uses the synthesis model from the literature for updating the regression coefficients from the IPD. Finally, all aggregation approaches reestimate the model intercept in the IPD to ensure that updated models remain well calibrated. For all three approaches, this can be achieved by fitting a logistic regression model in the IPD, using an offset variable that is calculated from the updated regression coefficients:

$$\Pr(y_s = 1) = \text{logit}^{-1}(\beta_{0,\text{adj}} + \text{offset}) \tag{1}$$

$$\text{where offset} = \hat{\beta}_{1,\text{UPD}} x_{s1} + \ldots + \hat{\beta}_{K,\text{UPD}} x_{sK} \tag{2}$$

In this expression, $\beta_{0,\text{adj}}$ is the only free parameter that is used as new estimate for the intercept of the aggregated prediction model. The variance–covariance matrix $\hat{\Sigma}_{\text{UPD}}$ can be adjusted according to the variance-correlation decomposition:

$$\widehat{\text{cov}}\left(\hat{\beta}_{0,\text{adj}}, \hat{\beta}_{i,\text{UPD}}\right) = \frac{\hat{\sigma}_{0,\text{adj}}}{\hat{\sigma}_{0,\text{UPD}}} \widehat{\text{cov}}\left(\hat{\beta}_{0,\text{UPD}}, \hat{\beta}_{i,\text{UPD}}\right) \quad \text{where } i = 1, \ldots, K \tag{3}$$

All approaches were implemented in R 2.14.1 [23]. The corresponding source code is available on request.

2.1. *Univariate meta-analysis*

A straightforward strategy to combine the previously published prediction models with IPD is to summarize their corresponding multivariate coefficients and standard errors. We propose the weighted least squares approach as a first simple approach to combine the coefficients. Appropriate weights for the coefficients can be obtained from their corresponding standard errors or study sample size when these are not available. This approach corresponds to a typical meta-analysis involving fixed or random effects as commonly applied to univariate regression coefficients or effect estimates. Here, the coefficient $\hat{\beta}_{ij}$ is weighted according to $w_{ij} = 1/\left(\hat{\sigma}_{ij}^2 + \tau_j^2\right)$ with τ_j^2 the between-study variance of $\hat{\beta}_j$.

As the coefficients are pooled independently for each predictor, dependencies between regression coefficients are ignored. This simplification is not necessarily problematic when the previously published regression coefficients are homogeneous. However, when estimates for these coefficients are known to be correlated across studies, a more advanced approach that accounts for between-study covariance may be more appropriate. We will discuss such an approach next.

2.2. *Multivariate meta-analysis*

The concept of multivariate meta-analysis is relatively new to the medical literature and can be seen as a generalization of DerSimonian and Laird's methodology for summarizing effect estimates [10, 24]. In contrast to univariate meta-analysis, the multivariate approach accounts for within-study covariance (instead of within-study variance). Furthermore, multivariate meta-analysis estimates between-study covariance (rather than between-study variance) of regression coefficients, and may therefore better account for heterogeneity across studies. This explicit distinction of within-study and between-study (co)variance has become paramount in epidemiological research. For this reason, we do not pursue other potentially useful approaches where evidence is aggregated from a different perspective, such as the generalized least squares approach proposed by Becker *et al.* [16].

In this section, we present a generalized random effects model that accounts for within-study and between-study covariance of the regression coefficients when pooling them. A univariate [25] and bivariate random effects model [26] for this purpose can be generalized as follows:

$$(\beta_0, \beta_1, \ldots, \beta_k)_l^{\mathrm{T}} \sim \mathcal{N}^{K+1}\left(\mu_{\text{re}}, (\Sigma_{\text{re}})_l\right) \tag{4}$$

with

$$(\Sigma_{\text{re}})_l = \Sigma_{\text{bs}} + \Sigma_l \tag{5}$$

and

$$\Sigma_{\text{bs}} = \begin{pmatrix} \tau_0^2 & \tau_{01} & \ldots & \tau_{0K} \\ \tau_{01} & \tau_1^2 & \ldots & \tau_{1K} \\ \ldots & \ldots & \ldots & \ldots \\ \tau_{0K} & \tau_{1K} & \ldots & \tau_K^2 \end{pmatrix} \tag{6}$$

and

$$\Sigma_l = \begin{pmatrix} \sigma_0^2 & \mathrm{cov}(\beta_0, \beta_1) & \dots & \mathrm{cov}(\beta_0, \beta_K) \\ \mathrm{cov}(\beta_0, \beta_1) & \sigma_1^2 & \dots & \mathrm{cov}(\beta_1, \beta_K) \\ \dots & \dots & \dots & \dots \\ \mathrm{cov}(\beta_0, \beta_K) & \mathrm{cov}(\beta_1, \beta_K) & \dots & \sigma_K^2 \end{pmatrix}_l \tag{7}$$

In the expressions earlier, between-study estimates are denoted as *bs* and random-effects estimates as *re*. Here, l denotes each included set of predictors from literature and IPD, that is, $l = \{1, \dots, M, \mathrm{IPD}\}$.

We explicitly distinguish between the within-study and between-study covariance of the regression coefficients, denoted as Σ_l (for study l) and Σ_{bs}, respectively. Estimates for $(\beta_0, \beta_1, \dots, \beta_K)_l$ and Σ_l can be obtained from $(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_K)_l$ and $\hat{\Sigma}_l$, respectively. The unknown parameters in μ_{re} and Σ_{bs} can be estimated with maximum likelihood, and provide the pooled means $\mu_{\mathrm{UPD}} = \mu_{\mathrm{re}}$ and covariance matrix $\Sigma_{\mathrm{UPD}} = \left(\sum_{l=1}^{M+1} (\Sigma_{\mathrm{re}})_l^{-1}\right)^{-1}$. Their corresponding log-likelihood is given by $\ell(\mu_{\mathrm{re}}, \Sigma_{\mathrm{bs}}) = \sum \ell_l(\mu_{\mathrm{re}}, \Sigma_{\mathrm{bs}})$ where $\ell_l(\mu_{\mathrm{re}}, \Sigma_{\mathrm{bs}}) = \log(\Pr(\beta_{0l}, \dots, \beta_{Kl} | \mu_{\mathrm{re}}, (\Sigma_{\mathrm{re}})_l))$ and $\Pr(\beta_{0l}, \dots, \beta_{Kl} | \mu_{\mathrm{re}}, (\Sigma_{\mathrm{re}})_l) \sim \mathcal{N}^{K+1}(\mu_{\mathrm{re}}, (\Sigma_{\mathrm{re}})_l)$. To facilitate convergence of the maximum likelihood estimation procedure, we used the independently pooled estimates of the previously published regression coefficients as initial values for μ_{re}, and a zero-matrix as initial choice for Σ_{bs}. In addition, we used the Cholesky decomposition to ensure that Σ_{bs} is positive semidefinite.

Although Σ_l is fully defined for the IPD, its non-diagonal entries are usually unknown for previously published regression coefficients. For this reason, we propose to impute missing entries in $\hat{\Sigma}_l$ based on the observed correlations in $\hat{\Sigma}_{\mathrm{IPD}}$, according to

$$\hat{\Sigma}_{\phi\psi l} = \widehat{\mathrm{cov}}\left(\hat{\beta}_{\phi l}, \hat{\beta}_{\psi l}\right) = \frac{\widehat{\mathrm{cov}}\left(\hat{\beta}_{\phi,\mathrm{IPD}}, \hat{\beta}_{\psi,\mathrm{IPD}}\right) \hat{\sigma}_{\phi l}\, \hat{\sigma}_{\psi l}}{\hat{\sigma}_{\phi,\mathrm{IPD}}\, \hat{\sigma}_{\psi,\mathrm{IPD}}} \tag{8}$$

with $\phi, \psi = 0, \dots, K$. This imputation strategy assumes that the within-study covariance of regression coefficients is exchangeable across all studies. Alternatively, it is possible to restrict non-diagonal entries in $\hat{\Sigma}_l$ to zero, according to $\hat{\Sigma}_l = \mathrm{diag}\left(\hat{\sigma}_{0l}^2, \hat{\sigma}_{1l}^2, \dots, \hat{\sigma}_{Kl}^2\right)$. The former approach may be more appropriate in more homogeneous sets of studies, as then the correlations from the IPD are likely to be closer to the underlying correlations in the included AD. Furthermore, it is possible to assume a common correlation value among all slopes (e.g., $\hat{\Sigma}_{\phi\psi l} = 0.2\, \hat{\sigma}_{\phi l}\, \hat{\sigma}_{\psi l}$), or to introduce uncertainty in the correlation parameter(s) by adopting a Bayesian perspective [16, 27]. Finally, simulation studies have revealed that multivariate meta-analysis models appear to be fairly robust to errors made in approximating within-study covariances when only summary effect estimates (here represented by the regression coefficients) are of interest [27].

The complexity of the meta-analysis is mostly defined by Σ_{bs}. If each element in this matrix is modeled as an unknown parameter, a full random effects meta-analysis is performed. Conversely, if all (non-diagonal) entries in Σ_{bs} and Σ_l are restricted to zero, the regression coefficients are pooled independently as described in Section 2.1. Furthermore, it is possible to perform a reduced random effects meta-analysis by restricting a selection of Σ_{bs}-elements to zero. For instance, we can assume fixed effects for β_1 by choosing $\tau_1^2 = \tau_{0,1} = \tau_{1,2} = \dots = \tau_{1,K} = 0$. Additional fixed effects can be introduced in a similar manner. We argue that by restricting the amount of unknown parameters in Σ_{bs}, estimates for their corresponding values may become more robust. The stability of μ_{re} and Σ_{bs} may further be improved by introducing (weakly) informative prior distributions. Unfortunately, such approach ultimately requires the use of highly advanced distributional families, which may not have a straightforward interpretation or implementation. Implementing these is beyond the scope of this article.

Finally, the described approach can easily be extended to scenarios in which multiple IPDs are available. In these scenarios, Σ_l is fully defined for multiple studies and hence allows an improved estimation of the unknown parameters. Alternatively, it is possible to adopt a one-stage approach that does not reduce the IPD to AD, but instead accounts for the fact that some studies provide IPD, and some studies provide only AD [28]. Similarly, when no IPDs are available, the non-diagonal entries of Σ_l are (probably) undefined for all studies, and making reasonable assumptions about these entries becomes more important to obtaining valid results.

2.3. *Bayesian inference*

The approaches described in Sections 2.1 and 2.2 estimate a "pooled" prediction model whenever a number of previously published prediction models as well as IPD are available. It may be clear that an average synthesis model across the included study populations may not always reflect the population of interest. Here, we assume that the IPD represents the clinically relevant population. Good prediction in these particular subjects is hence of primary interest. Therefore, we consider an alternative approach where the evidence from existing prediction models is used to update the regression coefficients from the IPD. To this purpose, we apply a Bayesian framework where a summary of the previously published regression coefficients serves as prior for the regression coefficients in the IPD. This summary of literature evidence can be obtained through the approach described in Section 2.2:

$$\mu_{\text{PRIOR}} = \mu_{\text{re}} \tag{9}$$

$$\Sigma_{\text{PRIOR}} = \left(\sum_{j=1}^{M} (\Sigma_{\text{re}})_j^{-1} \right)^{-1} \tag{10}$$

Note that this prior distribution does not include estimates from the IPD. Instead, we assume that the estimated coefficients from the IPD follow a multivariate normal distribution with mean μ_{IPD} and covariance matrix Σ_{IPD}. This distribution represents the likelihood and can be formulated as $\Pr(\beta_{0,\text{IPD}}, \ldots, \beta_{K,\text{IPD}} | \mu_{\text{IPD}}, \Sigma_{\text{IPD}}) \sim \mathcal{N}^{K+1}(\mu_{\text{IPD}}, \Sigma_{\text{IPD}})$. We propose to construct a conjugate prior distribution for μ_{IPD} with $\Pr(\mu_{\text{IPD}}) \sim \mathcal{N}^{K+1}(\mu_{\text{PRIOR}}, \Sigma_{\text{PRIOR}})$ such that the posterior density $\Pr(\mu_{\text{IPD}} | \beta_{0,\text{IPD}}, \ldots, \beta_{k,\text{IPD}}, \Sigma_{\text{IPD}}) \sim \mathcal{N}^{K+1}(\mu_{\text{POST}}, \Sigma_{\text{POST}})$ can be determined analytically:

$$\mu_{\text{UPD}} = \left(\Sigma_{\text{PRIOR}}^{-1} + \Sigma_{\text{IPD}}^{-1} \right)^{-1} \left(\Sigma_{\text{PRIOR}}^{-1} \, \mu_{\text{PRIOR}} + \Sigma_{\text{IPD}}^{-1} \, \mu_{\text{IPD}} \right) \tag{11}$$

$$\Sigma_{\text{UPD}} = \left(\Sigma_{\text{PRIOR}}^{-1} + \Sigma_{\text{IPD}}^{-1} \right)^{-1} \tag{12}$$

Here, the parameters μ_{IPD} and Σ_{IPD} can be substituted by $(\hat{\beta}_{0,\text{IPD}}, \ldots, \hat{\beta}_{K,\text{IPD}})$ and $\hat{\Sigma}_{\text{IPD}}$, respectively. Consequently, the vector μ_{UPD} represents the expected (posterior) value of the multivariate regression coefficients $\beta_{0,\text{UPD}}, \ldots, \beta_{K,\text{UPD}}$, and Σ_{UPD} represents the expected (posterior) value of the corresponding variance–covariance matrix. When multiple IPDs are available, it is possible to subsequently add each IPD using Bayesian inference.

3. Application: traumatic brain injury

We tested univariate meta-analysis, multivariate meta-analysis, Bayesian inference, and standard logistic regression (SLR) modeling (i.e., analysis using the IPD only) on 15 empirical datasets of TBI patients. TBI is a leading cause of death and disability worldwide with a substantial economic burden [29, 30]. It is difficult to establish a reliable prognosis on admission [31]. This requires the consideration of multiple and easily accessible risk factors in multivariable prognostic models [5, 22, 32, 33]. Many prognostic models with admission data are readily available from the literature [32]. However, most models were developed on relatively small sample sizes originating from a single center or region and lack external validation [9, 32]. Therefore, their aggregation might improve the generalization of novel prognostic models.

3.1. *Application setup*

To test the potential value of our approaches, we used 15 series of IPD collected in the International Mission for Prognosis and Analysis of Clinical Trials in TBI (IMPACT) project [21]. The outcome used in each of these trials was the Glasgow Outcome Scale score (GOS) at 6 months after injury, dichotomized between severe and moderate disability.

We fitted a logistic regression model to each of the available datasets and considered a core set of conventional TBI prognostic factors (age, motor score, and pupil response to light) (Table I) [22, 32]. In this manner, we aimed to simulate scenarios in which a common set of core predictors is available and can be aggregated with IPD. We realize that, for many genuine examples, the assumption of literature models sharing the same set of parameters is unrealistic. This problem also arises in our application, where some of the previously published regression coefficients are unknown because some studies did

Table I. Estimated regression coefficients (and standard error) from the IMPACT data.

Characteristics	Coding		Logistic regression coefficients for favorable versus unfavorable outcome after 6 months TBI.							
			TINT	**TIUS**	**SLIN**	**SAPHIR**	**PEGSOD**	**HIT I**	**UK4**	**TCDB**
Patients			1,118	1,041	409	919	1,510	350	791	603
Study type			RCT	RCT	RCT	RCT	RCT	RCT	Obs.	Obs.
Intercept		$\hat{\beta}_0$	−2.48 (0.21)	−3.06 (0.25)	−2.06 (0.34)	−2.43 (0.24)	−2.76 (0.21)	−2.66 (0.47)	−2.13 (0.26)	−2.24 (0.33)
Age, years		$\hat{\beta}_1$	0.03 (0.00)	0.04 (0.01)	0.03 (0.01)	0.04 (0.00)	0.04 (0.00)	0.03 (0.01)	0.04 (0.01)	0.05 (0.01)
Motor score	None	$\hat{\beta}_2$	1.49 (0.95)	1.42 (0.76)	NA	0.69 (0.23)	1.52 (0.17)	1.36 (0.37)	1.33 (0.35)	2.05 (0.39)
	Extension	$\hat{\beta}_3$	1.84 (0.24)	1.93 (0.25)	1.69 (0.38)	1.50 (0.25)	2.58 (0.25)	2.53 (0.54)	1.67 (0.42)	2.14 (0.37)
	Abnormal flexion	$\hat{\beta}_4$	1.10 (0.19)	1.61 (0.23)	0.63 (0.29)	0.47 (0.23)	1.47 (0.21)	1.95 (0.47)	1.18 (0.49)	0.74 (0.33)
	Normal flexion	$\hat{\beta}_5$	0.51 (0.17)	0.81 (0.18)	0.28 (0.27)	0.19 (0.20)	0.82 (0.18)	0.80 (0.42)	0.44 (0.25)	0.48 (0.28)
	Localizes/obeys		Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
	Untestable/missing	$\hat{\beta}_6$	NA	NA	NA	0.34 (1.23)	NA	1.08 (0.77)	0.94 (0.24)	−0.14 (0.47)
Pupillary reactivity	Both pupils reacted		Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
	One pupil reacted	$\hat{\beta}_7$	0.82 (0.19)	0.28 (0.23)	1.08 (0.28)	1.22 (0.17)	0.48 (0.19)	0.42 (0.35)	0.80 (0.26)	0.70 (0.35)
	No pupil reacted	$\hat{\beta}_8$	1.28 (0.22)	1.29 (0.19)	2.08 (0.80)	NA	1.05 (0.13)	2.15 (0.42)	2.04 (0.28)	1.54 (0.26)

			SKB	**EBIC**	**HIT II**	**NABIS**	**CSTAT**	**PHARMOS**	**APOE**
Patients			126	822	819	385	517	856	756
Study type			RCT	Obs.	RCT	RCT	RCT	RCT	Obs.
Intercept		$\hat{\beta}_0$	−1.77 (0.68)	−3.12 (0.28)	−2.70 (0.28)	−2.14 (0.41)	−2.46 (0.35)	−1.50 (0.24)	−3.15 (0.27)
Age, years		$\hat{\beta}_1$	0.04 (0.02)	0.04 (0.00)	0.03 (0.01)	0.04 (0.01)	0.03 (0.01)	0.02 (0.01)	0.04 (0.00)
Motor score	None	$\hat{\beta}_2$	0.56 (0.63)	1.61 (0.28)	1.07 (0.24)	0.97 (0.33)	0.88 (0.41)	0.54 (0.34)	1.31 (1.15)
	Extension	$\hat{\beta}_3$	0.63 (0.71)	1.90 (0.35)	2.07 (0.35)	1.69 (0.40)	1.49 (0.33)	1.31 (0.27)	NA
	Abnormal flexion	$\hat{\beta}_4$	1.30 (0.76)	1.53 (0.36)	1.63 (0.29)	1.76 (0.40)	1.14 (0.32)	1.03 (0.23)	NA
	Normal flexion	$\hat{\beta}_5$	−0.18 (0.74)	1.33 (0.27)	0.48 (0.25)	0.75 (0.33)	0.07 (0.29)	0.64 (0.19)	1.28 (0.56)
	Localizes/obeys		Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
	Untestable/missing	$\hat{\beta}_6$	−0.65 (0.74)	1.12 (0.25)	0.97 (0.34)	0.77 (0.72)	NA	0.51 (0.23)	1.17 (0.19)
Pupillary reactivity	Both pupils reacted		Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
	One pupil reacted	$\hat{\beta}_7$	1.09 (0.46)	1.01 (0.29)	0.37 (0.24)	1.03 (0.37)	1.53 (0.28)	0.52 (0.19)	0.87 (0.37)
	No pupil reacted	$\hat{\beta}_8$	NA	1.44 (0.23)	1.26 (0.23)	1.18 (0.29)	1.87 (0.32)	0.47 (0.37)	2.04 (0.36)

Note : NA, not available.

not contain all categories of the motor score or pupil response. Instead of discarding the corresponding predictors from the aggregated model, we propose using uninformative regression coefficients when they cannot be estimated from the data. We argue that this strategy can also be applied in other examples where the literature models do not share the same set of parameters. Finally, we measured the area under the receiver operator characteristic curve (AUC) and the Brier score (BS) of the aggregated models as indication of performance. Whereas the former quantifies the model's ability to distinguish high-risk from low-risk patients, the latter assesses the accuracy of its predictions [34, 35].

3.2. *Practical example*

As an illustration, we used the HIT I study [36] as IPD, the HIT II study [37] as validation data, and the prediction models of the remaining studies as previously published evidence (Table II). We calculated the I^2 index of heterogeneity for each separate (and known) regression coefficient of the previously published prediction models by performing a univariate meta-analysis [38]. These coefficients were found to be moderately to strongly heterogeneous with $I^2(\hat{\beta}_0) = 0.71$, $I^2(\hat{\beta}_1) = 0.15$, $I^2(\hat{\beta}_2) = 0.49$, $I^2(\hat{\beta}_3) = 0.40$, $I^2(\hat{\beta}_4) = 0.52$, $I^2(\hat{\beta}_5) = 0.48$, $I^2(\hat{\beta}_6) = 0.54$, $I^2(\hat{\beta}_7) = 0.53$ and $I^2(\hat{\beta}_8) = 0.61$. These estimates should however be interpreted with caution, as much discrepancy between the previously published regression coefficients is caused by small standard errors. Next, we imputed previously published regression coefficients that could not be estimated from the data and performed a sensitivity analysis to assess two different imputation approaches.

To this effect, we evaluated $\hat{\beta}_\phi = 0$ with $\hat{\sigma}^2_\phi = 100$ and compared it with a mean imputation with $\hat{\sigma}^2_\phi = \sum_{j=1}^{M} \hat{\sigma}^2_{\phi j}$. Finally, we aggregated the previously published prediction models with the IPD. The considered approaches are: SLR modeling ignoring the literature studies, univariate meta-analysis, multivariate meta-analysis, and Bayesian inference. We also performed a logistic regression analysis using all available IPD datasets (except for the validation study), and used the resulting model as "gold standard" for comparing the aggregated models. Because the multivariate meta-analysis approach requires the within-study covariance of the previously published prediction models to be fully specified, we evaluated two strategies for imputing missing (i.e., non-diagonal) entries in Σ_l. As explained earlier, we compared a strategy that involved imputing missing covariance entries based on observed correlation in the IPD with a strategy based on restricted non-diagonal entries in Σ_l to zero.

Results (Table II) from this example illustrate that particular choices for imputing missing regression coefficients and unknown within-study covariance do not have a large impact on the resulting prediction model. Although each strategy yields somewhat different estimated regression coefficients, most variation seems to arise from the uncertainty in the available regression coefficients. The example also illustrates that regression coefficients of aggregated prediction models are more similar to the coefficients from the reference "gold standard" model (compared with SLR modeling). Furthermore, we noticed that prediction models incorporating prior evidence achieved slightly improved AUC and Brier scores. It is possible that improvements in this particular example are relatively small owing to the strong relation between the IPD and validation data (the HIT II study is a follow-up study of the HIT I study). Finally, we noticed a considerable decrease in the standard errors of estimated regression coefficients when prior evidence was incorporated. Although these errors are not of primary concern in prediction research, they reflect an improved stability of the derived prediction models.

3.3. *Performance study*

In order to evaluate the overall performance of aggregation models, we performed a split-sample procedure where IPD and validation data were sampled (without replacement) from a common dataset. The prediction models generated from the remaining datasets were used as prior evidence for the aggregation methods. This procedure was repeated 100 times for each scenario to ensure stable estimates of model performance. We evaluated $N_{\text{IPD}} = 500$ and $N_{\text{IPD}} = 200$, and imputed unknown regression coefficients according to $\hat{\beta}_\phi = 0$ with $\hat{\sigma}^2_\phi = 100$.

Results indicate that all aggregation approaches perform similarly and yield prediction models with an improved AUC and Brier score (Table III). These improvements particularly occur in small datasets ($N_{\text{IPD}} = 200$) but do not necessarily disappear when more IPD is at hand ($N_{\text{IPD}} = 500$). Furthermore, we noticed that aggregated prediction models perform similarly compared with models derived with the IPD from all original studies (*Full IPD modeling*). Finally, we noticed that standard errors of aggregated regression coefficients tend to be smaller when estimated with multivariate meta-analysis (compared with univariate meta-analysis).

Table II. An illustration of the proposed approaches in the TBI application: updated regression coefficients (and standard error) when the HIT I study ($N = 350$) is used as individual participant dataset, the HIT II study ($N = 819$) as validation dataset and the remaining studies as evidence from the literature.

	(Intercept)	**Age, years**	**Motor score ***					**Pupillary reactivity ****			
	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$	AUC	BS
SLR modeling Analysis ignoring literature studies	−2.66 (0.47)	0.03 (0.01)	1.36 (0.37)	2.53 (0.54)	1.95 (0.47)	0.80 (0.42)	1.08 (0.77)	0.42 (0.35)	2.15 (0.42)	0.745 (0.017)	0.206 (0.008)
Full IPD modeling Analysis with IPD of all original studies stacked	−2.52 (0.07)	0.04 (0.00)	1.22 (0.07)	1.88 (0.08)	1.21 (0.07)	0.60 (0.06)	0.98 (0.08)	0.80 (0.06)	1.48 (0.06)	0.749 (0.017)	0.207 (0.007)
Uninformative regression coefficients for missing estimates in the literature models $\left(\hat{\beta}_\phi = 0 \text{ with } \hat{\sigma}^2_\phi = 100\right)$											
Univariate meta-analysis	−2.67 (0.12)	0.04 (0.00)	1.20 (0.13)	1.81 (0.12)	1.17 (0.12)	0.60 (0.09)	0.82 (0.13)	0.83 (0.10)	1.46 (0.12)	0.749 (0.017)	0.203 (0.007)
Multivariate meta-analysis missing within-study covariance restricted to zero	−2.67 (0.12)	0.04 (0.00)	1.21 (0.10)	1.81 (0.09)	1.17 (0.10)	0.60 (0.07)	0.81 (0.11)	0.83 (0.07)	1.44 (0.12)	0.749 (0.017)	0.203 (0.007)
Multivariate meta-analysis missing within-study covariance imputed from IPD	−2.67 (0.12)	0.04 (0.00)	1.20 (0.08)	1.81 (0.08)	1.17 (0.07)	0.60 (0.06)	0.82 (0.10)	0.83 (0.07)	1.46 (0.07)	0.749 (0.017)	0.203 (0.007)
Bayesian inference missing within-study covariance restricted to zero	−2.65 (0.12)	0.04 (0.00)	1.19 (0.11)	1.83 (0.09)	1.19 (0.09)	0.59 (0.07)	0.81 (0.11)	0.81 (0.07)	1.51 (0.12)	0.749 (0.017)	0.203 (0.007)
Mean imputation for missing estimates in the literature models $\left(\text{with } \hat{\sigma}^2_\phi = \sum_{j=1}^{M} \hat{\sigma}^2_{\phi j}\right)$											
Univariate meta-analysis	−2.67 (0.12)	0.04 (0.00)	1.20 (0.13)	1.81 (0.12)	1.17 (0.12)	0.60 (0.09)	0.81 (0.13)	0.83 (0.10)	1.46 (0.12)	0.749 (0.017)	0.203 (0.007)
Multivariate meta-analysis missing within-study covariance restricted to zero	−2.67 (0.12)	0.04 (0.00)	1.21 (0.10)	1.81 (0.09)	1.17 (0.10)	0.60 (0.07)	0.81 (0.11)	0.83 (0.07)	1.44 (0.12)	0.749 (0.017)	0.203 (0.007)
Multivariate meta-analysis missing within-study covariance imputed from IPD	−2.67 (0.12)	0.04 (0.00)	1.20 (0.08)	1.81 (0.08)	1.17 (0.07)	0.60 (0.06)	0.81 (0.10)	0.83 (0.07)	1.46 (0.07)	0.749 (0.017)	0.203 (0.007)
Bayesian inference missing within-study covariance restricted to zero	−2.65 (0.12)	0.04 (0.00)	1.19 (0.11)	1.83 (0.09)	1.21 (0.08)	0.59 (0.07)	0.81 (0.11)	0.79 (0.10)	1.47 (0.08)	0.749 (0.017)	0.202 (0.007)

Note : The area under the receiver operator characteristic curve (AUC) and the Brier score (BS) of the aggregated models are presented as measure of performance in HIT II. Standard errors for the AUC were obtained through the standard error of the Somer's D statistic. Standard errors for the Brier score were estimated according to $\text{sd}[(p_s - o_s)^2]/\sqrt{N}$. The categorical variables Motor score (*) and Pupillary reactivity (**) were coded as factors (cfr. Table I).

Table III. Performance of aggregated prediction models, expressed by means of the area under the receiver operator characteristic curve (AUC) and the Brier score (BS).

	UK4				**EBIC**			
	$N_{\mathrm{IPD}} = 500$ ($N_{\mathrm{VAL}} = 291$)		$N_{\mathrm{IPD}} = 200$ ($N_{\mathrm{VAL}} = 591$)		$N_{\mathrm{IPD}} = 500$ ($N_{\mathrm{VAL}} = 322$)		$N_{\mathrm{IPD}} = 200$ ($N_{\mathrm{VAL}} = 622$)	
	AUC (SE)	BS (SE)	AUC (SE)	BS (SE)	AUC (SE)	BS (SE)	AUC (SE)	BS (SE)
SLR modeling Analysis ignoring literature studies	0.813 (0.022)	0.165 (0.010)	0.801 (0.011)	0.172 (0.006)	0.810 (0.019)	0.179 (0.010)	0.801 (0.013)	0.185 (0.007)
Full IPD modeling Analysis with IPD of all original studies stacked	0.822 (0.020)	0.174 (0.009)	0.821 (0.008)	0.176 (0.003)	0.814 (0.019)	0.176 (0.009)	0.814 (0.010)	0.176 (0.004)
Univariate meta-analysis	0.821 (0.020)	0.162 (0.009)	0.820 (0.008)	0.164 (0.005)	0.815 (0.019)	0.176 (0.009)	0.814 (0.010)	0.176 (0.004)
Multivariate meta-analysis Missing within-study covariance restricted to zero	0.820 (0.020)	0.162 (0.009)	0.820 (0.008)	0.164 (0.005)	0.815 (0.019)	0.176 (0.009)	0.814 (0.010)	0.177 (0.005)
Bayesian inference Missing within-study covariance restricted to zero	0.820 (0.020)	0.162 (0.009)	0.820 (0.008)	0.164 (0.005)	0.814 (0.019)	0.176 (0.009)	0.814 (0.010)	0.177 (0.005)
	HIT II				**PHARMOS**			
	$N_{\mathrm{IPD}} = 500$ ($N_{\mathrm{VAL}} = 319$)		$N_{\mathrm{IPD}} = 200$ ($N_{\mathrm{VAL}} = 619$)		$N_{\mathrm{IPD}} = 500$ ($N_{\mathrm{VAL}} = 356$)		$N_{\mathrm{IPD}} = 200$ ($N_{\mathrm{VAL}} = 656$)	
	AUC (SE)	BS (SE)	AUC (SE)	BS (SE)	AUC (SE)	BS (SE)	AUC (SE)	BS (SE)
SLR modeling Analysis ignoring literature studies	0.739 (0.021)	0.201 (0.008)	0.728 (0.013)	0.207 (0.007)	0.642 (0.024)	0.237 (0.007)	0.627 (0.017)	0.243 (0.007)
Full IPD modeling Analysis with IPD of all original studies stacked	0.744 (0.020)	0.205 (0.007)	0.742 (0.010)	0.207 (0.004)	0.653 (0.022)	0.242 (0.008)	0.656 (0.009)	0.242 (0.004)
Univariate meta-analysis	0.744 (0.020)	0.199 (0.008)	0.743 (0.010)	0.199 (0.005)	0.654 (0.023)	0.236 (0.008)	0.657 (0.009)	0.236 (0.004)
Multivariate meta-analysis Missing within-study covariance restricted to zero	0.745 (0.020)	0.198 (0.008)	0.743 (0.010)	0.199 (0.005)	0.654 (0.024)	0.236 (0.008)	0.657 (0.009)	0.236 (0.004)
Bayesian inference Missing within-study covariance restricted to zero	0.745 (0.019)	0.198 (0.008)	0.743 (0.010)	0.199 (0.005)	0.654 (0.024)	0.236 (0.008)	0.657 (0.009)	0.236 (0.004)

Note : For multivariate meta-analysis and Bayesian inference, we used uninformative regression coefficients when missing.

4. Application: deep venous thrombosis

To confirm the potential value of the proposed approaches, we describe a genuine clinical example involving the prediction of deep venous thrombosis (DVT). In this example, we aggregated five previously published prediction models [39–44] with one IPD set, and evaluated different strategies for coping with missing predictor values and within-study covariance. We used an IPD ($N = 1028$) from the Amsterdam–Maastricht–Utrecht Study on thromboEmbolism (AMUSE-1) [45] and aggregated these data with the prediction models described next. A detailed description of the predictors can be found in the Appendix. After aggregation, we validated the original and aggregated models in an independent dataset of 791 participants [46].

Unfortunately, we encountered some difficulties during incorporation of the previously published prediction models. For instance, some articles did not report the original regression coefficients and standard errors of the prediction model and reported a scoring rule with weights instead, with score $=$ weight$_1 x_1 + \ldots +$ weight$_\mathrm{K} x_K$ (e.g., Wells rule, modified Wells rule, and Hamilton rule). We attempted to reconstruct the original regression coefficients and standard errors by deriving a prediction model in the IPD with the scoring rule as single variable according to:

$$\Pr(\text{DVT presence}) = \text{logit}^{-1}(\beta_{\text{adj}0} + \beta_{\text{adj}1}\text{score}) \tag{13}$$

The resulting slope $\hat{\beta}_{\text{adj}1}$ is then multiplied with the reported weights to obtain an estimate for the original regression coefficients, and $\hat{\beta}_{\text{adj}0}$ is used as estimate for the model intercept. Conservative estimates for the corresponding standard errors can be obtained by assuming

$$\sigma_{\text{adj}1} = \left(\sum_{j=1}^{M} \sigma_j^{-2} \right)^{-1/2} \tag{14}$$

This assumption implies that the standard errors σ_j are equal for all regression coefficients of the model under consideration. The standard error for the model intercept can be directly obtained from $\hat{\sigma}_{\text{adj}0}$. Alternatively, reported p-values of regression coefficients can be converted into standard errors by assuming normality. An advantage of this approach is that the AUC of reconstructed models remains equal to the performance of the original models, as the linear predictors are proportionally identical.

We illustrate this approach using the Wells rule. This rule consists of nine clinical items where WellsScore $=$ 1 malign $+$ 1 par $+$ 1 surg $+$ 1 tend $+$ 1 leg $+$ 1 calfdif3 $+$ 1 pit $+$ 1 vein $-$ 2 altdiagn. We attempted to reconstruct the original regression coefficients and standard errors by deriving a prediction model in the IPD with the Wells score as single variable. This approach yielded the following model: $\Pr(\text{DVT presence}) = \text{logit}^{-1}(-2.66 + 0.52\ \text{WellsScore})$. Consequently, we may reconstruct the original regression coefficients as follows: $\hat{\beta}_0 = -2.66$, $\hat{\beta}_{\text{malign}} = 0.52$, $\hat{\beta}_{\text{par}} = 0.52$, $\hat{\beta}_{\text{surg}} = 0.52$, $\hat{\beta}_{\text{tend}} = 0.52$, $\hat{\beta}_{\text{leg}} = 0.52$, $\hat{\beta}_{\text{calfdif3}} = 0.52$, $\hat{\beta}_{\text{pit}} = 0.52$, $\hat{\beta}_{\text{vein}} = 0.52$ and $\hat{\beta}_{\text{altdiagn}} = -1.04$. We found $\hat{\sigma}_{\text{adj}0} = 0.15$ and $\hat{\sigma}_{\text{adj}1} = 0.05$, such that $\hat{\sigma}_0 = 0.15$ and $\hat{\sigma}_{\text{malign}}, \ldots, \hat{\sigma}_{\text{altdiagn}} = 0.16$.

We applied the previously published models in the validation data and observed an AUC < 0.634, and a Brier score > 0.133 for most models, with exception of the Oudega model (AUC $= 0.767$ and Brier score $= 0.125$).

4.1. *Evidence aggregation*

Consequently, we aggregated the previously published prediction models with the IPD. The approaches considered are: standard logistic regression (ignoring the evidence from the literature), univariate meta-analysis, multivariate meta-analysis, and Bayesian inference. Because a relatively large number of predictors were considered, including all of them would preclude multivariate meta-analysis that would lead to clinically viable prediction models (15 predictors $+$ intercept). Hence, we focused on a subset of four important predictors: *malign*, *surg*, *calfdif3*, and *ddimdich*. A summary of the evidence from each of the literature sources and from the IPD is presented in Table IV. These were then pooled. In order to appraise the quality of the derived model (which only included four core predictors), we also fitted a more complex prediction model where we considered the eight predictors from the Oudega model. The AUC of the resulting model however decreased from 0.72 to 0.70, indicating that the simplified model is more generalizable and presents a better reference for comparing the aggregated prediction

Table IV. Overview of reconstructed regression coefficients (and standard errors) of the previously published prediction models in the DVT application.

Characteristics	**Logistic regression coefficients for DVT outcome**						
Prediction model	Wells	Modified Wells	Gagne	Hamilton	Oudega	IPD (4)	IPD (8)
Patients	593	530	276	309	1,295	1,028	1,028
(Intercept)	−2.66 (0.15)	−2.77 (0.15)	−1.69 (0.10)	−2.72 (0.17)	−5.47 (NA)	−3.95 (0.28)	−4.67 (0.37)
altdiagn	−1.05 (0.16)	−1.06 (0.17)	−1.77 (0.19)				
calfdif3	0.52 (0.16)	0.53 (0.17)	0.70 (0.19)	0.43 (0.18)	1.13 (0.34)	0.86 (0.20)	0.87 (0.21)
ddimdich					3.01 (0.91)	2.39 (0.29)	2.40 (0.30)
eryt				0.43 (0.18)			
histdvt		0.53 (0.17)	0.63 (0.19)	0.87 (0.18)			
leg	0.52 (0.16)	0.53 (0.17)					
malign	0.52 (0.16)	0.53 (0.17)	1.69 (0.19)	0.87 (0.18)	0.42 (0.24)	0.77 (0.36)	0.68 (0.36)
notraum					0.60 (0.19)		0.55 (0.25)
oachst			1.17 (0.19)		0.75 (0.24)		−12.44 (535)
par	0.52 (0.16)	0.53 (0.17)		0.87 (0.18)			
pit	0.52 (0.16)	0.53 (0.17)					
sex				0.43 (0.18)	0.59 (0.18)		0.60 (0.21)
surg	0.52 (0.16)	0.53 (0.17)	0.53 (0.19)	0.43 (0.18)	0.38 (0.19)	−0.13 (0.37)	−0.04 (0.38)
tend	0.52 (0.16)	0.53 (0.17)					
vein	0.52 (0.16)	0.53 (0.17)			0.48 (0.16)		0.22 (0.26)

Note : IPD (4) and IPD (8) represent the models derived from the AMUSE-1 study, with four and eight core predictors, respectively.

models. Finally, we compared the simplified aggregated models with a more extensive model derived with univariate meta-analysis using the eight predictors from the Oudega model. This model yielded the following regression coefficients (and standard error): $\hat{\beta}_0 = -4.70$ (0.10), $\hat{\beta}_{\text{calfdif3}} = 0.63$ (0.08), $\hat{\beta}_{\text{ddimdich}} = 2.45$ (0.28) $\hat{\beta}_{\text{malign}} = 0.79$ (0.20), $\hat{\beta}_{\text{notraum}} = 0.58$ (0.15), $\hat{\beta}_{\text{oachst}} = 1.01$ (0.15), $\hat{\beta}_{\text{sex}} = 0.54$ (0.11), $\hat{\beta}_{\text{surg}} = 0.46$ (0.08), and $\hat{\beta}_{\text{vein}} = 0.48$ (0.09).

4.2. *Results in the DVT case study*

Results in Table V indicate that the aggregated prediction models, despite including few(er) predictors, are superior to models that do not incorporate evidence from the literature. However, we also noticed that the Oudega model outperforms the aggregated models in terms of AUC (but achieves a similar Brier score). This discrepancy decreases when an extended model with eight predictors using univariate meta-analysis is derived (AUC = 0.759 and Brier Score = 0.124). These results possibly indicate that the Oudega model considerably contributes to the discriminative ability of the aggregated models. Particularly, it is the only literature model with a regression coefficient for *ddimdich*, a relatively strong predictor in DVT. We noticed that $\hat{\beta}_{\text{ddimdich}}$ was considerably smaller in the IPD and aggregated models, and much larger in the Oudega model and validation data ($\hat{\beta}_{\text{ddimdich}} = 3.95$, adjusted for the four core predictors), which may partially explain the decrease in discriminative ability. Furthermore, results indicate that different implementations for multivariate meta-analysis perform similarly. Estimated regression coefficients and standard errors, on the other hand, may considerably differ according to the implemented approach. For instance, we noticed that uninformative imputation yielded relatively large standard errors for $\hat{\beta}_{\text{ddimdich}}$. Possibly, these errors are inflated in multivariate meta-analysis because some of the estimated between-study correlations take extreme values: $\rho(\hat{\beta}_{\text{ddimdich}}, \hat{\beta}_0) = -0.79$ and $\rho(\hat{\beta}_{\text{ddimdich}}, \hat{\beta}_{\text{malign}}) = -0.97$ [47]. Finally, we noticed that standard errors of aggregated regression coefficients tend to be smallest when estimated with Bayesian inference.

5. Discussion

In line with previous research, we found that the aggregation and incorporation of previously published prediction models can indeed improve the performance of a novel prediction model [3, 13, 26, 48]. The case studies demonstrate that the proposed methods are particularly useful when a few participant data are at hand. Although the aggregation methods perform similarly in most scenarios, multivariate

Table V. Multivariate regression coefficients (and standard error) of the aggregated prediction models in the DVT application.

	$\hat{\beta}_0$	$\hat{\beta}_{\text{malign}}$	$\hat{\beta}_{\text{surg}}$	$\hat{\beta}_{\text{calfdif3}}$	$\hat{\beta}_{\text{ddimdich}}$	AUC	BS
SLR modeling	−3.95	0.77	−0.13	0.86	2.39	0.723	0.123
Analysis ignoring literature studies	(0.28)	(0.36)	(0.37)	(0.20)	(0.29)	(0.021)	(0.007)
Uninformative regression coefficients for missing estimates in the literature models $\left(\hat{\beta}_\phi = 0 \text{ with } \hat{\sigma}^2_\phi = 100\right)$							
Univariate meta-analysis	−3.94	0.80	0.46	0.63	2.44	0.730	0.123
	(0.10)	(0.20)	(0.08)	(0.08)	(0.28)	(0.019)	(0.007)
Multivariate meta-analysis	−3.52	0.75	0.40	0.64	1.95	0.730	0.122
missing within-study covariance restricted to zero	(0.10)	(0.17)	(0.11)	(0.10)	(1.02)	(0.019)	(0.007)
Bayesian inference	−3.28	0.49	0.45	0.68	1.64	0.738	0.122
missing within-study covariance restricted to zero	(0.10)	(0.14)	(0.08)	(0.10)	(0.20)	(0.020)	(0.007)
Mean imputation for missing estimates in the literature models $\left(\text{with } \hat{\sigma}^2_\phi = \sum_{j=1}^{M} \hat{\sigma}^2_{\phi j}\right)$							
Univariate meta-analysis	−4.08	0.80	0.46	0.63	2.60	0.730	0.123
	(0.10)	(0.20)	(0.08)	(0.08)	(0.24)	(0.019)	(0.007)
Multivariate meta-analysis	−3.96	0.72	0.40	0.74	2.43	0.738	0.123
missing within-study covariance restricted to zero	(0.10)	(0.18)	(0.09)	(0.12)	(0.45)	(0.020)	(0.007)
Bayesian inference	−3.88	0.72	0.38	0.80	2.30	0.738	0.123
missing within-study covariance restricted to zero	(0.10)	(0.16)	(0.08)	(0.10)	(0.21)	(0.020)	(0.007)

Note : The area under the receiver operator characteristic curve (AUC) and the Brier score (BS) of the aggregated models are presented together with their standard error as measure of performance in the validation dataset.

meta-analysis and Bayesian inference tend to yield smaller confidence intervals for the regression coefficients. According to previous research, this may be related to the fact that these approaches take more evidence into account [49] and allow more flexibility. The inclusion of additional evidence (i.e., within-study covariance) may, however, also introduce additional uncertainty and cause estimation difficulties, resulting in an inflation of standard errors [27, 47]. Finally, results indicate that the proposed aggregation approaches may considerably reduce model complexity without comprising their predictive accuracy. Particularly, by focusing on a set of core predictors, the model can be pruned effectively.

In this article, we evaluated and compared three evidence aggregation approaches in two case studies using real clinical data. The two case studies demonstrate that aggregation yields prediction models with an improved discrimination and calibration in a vast majority of scenarios, and result in equivalent performance (compared with the standard approach) in a small minority of situations. The exact preconditions for this occurrence could not be definitively established here. Possibly, data aggregation is little added value in scenarios where derivation and validation populations are highly similar and the AD from the literature is relatively different. The exact causes need to be further explored.

Finally, we have illustrated how the generally unrealistic assumption of consistency in the availability of evidence across included studies can be relaxed for real-life scenarios. Specifically, we have demonstrated how these methods can be applied when predictor values, covariance data, and even original regression coefficients are unknown. The fact that aggregation of such evidence succeeds in improving the performance of novel prediction models underscores the value and versatility of this methodology, as illustrated in the DVT example.

Based on these results from our empirical studies, the following tentative guidelines can be proposed. First, when there are relatively many IPD at hand and evidence from the literature is strongly heterogeneous with these data, the standard approach, by fitting a new model (from scratch) from that dataset without incorporating or synthesizing the published evidence, is acceptable. Secondly, when the evidence from the literature is moderately heterogeneous, or the IPD is relatively small, Bayesian inference (and multivariate meta-analysis) may improve calibration and discrimination of the newly developed prediction model. Even when the actual degree of heterogeneity is unknown, these approaches may still be preferred to the standard approach of fitting an entirely new model from scratch, and is relatively easy to implement. Finally, when the evidence from the literature is (relatively) homogeneous, univariate meta-analysis represents a superior approach for improving or updating the newly developed prediction model. Heterogeneity may be quantified using the I^2-statistic, where published criteria suggest adjectives of low, moderate, and high to I^2 values of 25 50 and 75% [38].

5.1. Limitations

Although we addressed important aspects of aggregating data in the two case studies, we did not assess or address the potential impact of selection bias. Conceivably, pooled regression coefficients may be overestimated or underestimated when important predictors are excluded. This problem may arise when literature models are derived using data-driven selection with stepwise methods, and particularly in small samples [50]. Furthermore, the selection of a core set of predictors may introduce additional bias when the excluded regression coefficients are strongly influential or correlated with the included predictors. This is known as confounding of pooled effects, and usually results in underestimation of pooled regression coefficients (as predictors are typically positive in clinical prediction research). It is therefore important to select a reasonable set of core predictors when pooling differently specified prediction models.

Another potential limitation of this article is the fact that only two clinical examples were examined. Conceivably, these may not be representative of the majority of clinical prediction research, and our evaluation of the evidence aggregation methods are not reproducible in different scenarios. We feel that this is unlikely because the examples used, TBI and DVT, are two typical areas of clinical prediction research for which we included numerous articles (15 and 5, respectively). We welcome the evaluation of these approaches in other case studies by other authors.

Finally, our DVT application illustrates that aggregated prediction models generally improve the predictive accuracy of novel prediction models but do not always outperform previously published prediction models in terms of discriminative ability. We demonstrated that this situation may occur when a strong predictor is poorly available from the literature and not well estimated in the IPD. Moreover, it is well known that the AUC is not the most sensitive measure to assess incremental value of predictors [51, 52]. For this reason, we also considered model accuracy in terms of the Brier score.

5.2. Conclusion

The incorporation of previously published prediction models into the development of a novel prediction model with a similar set of predictors is both feasible and beneficial when IPD are available. Particularly in small datasets, we noticed that the inclusion of such aggregate evidence may provide considerable leverage to improve the regression coefficients and discriminative ability of the new prediction model. However, it remains paramount that researchers identify to what extent the previously published prediction models are comparable with those in the available IPD, as the justification of the considered approaches depends on the clinical relevance of the aggregated model. Future research may therefore focus on the quantification of heterogeneity across prediction models. In conclusion, aggregation is better or at least equivalent. Real-life clinical examples support these conclusions.

Appendix A. Overview of the variables in the AMUSE-1 dataset.

sex	Gender
	0 = female
	1 = male
age	Age
side	Side of legpain
	0 = left side
	1 = right side
	2 = both sides
durat	Duration of symptoms
malign	Active malignancy
	0 = no active malignancy
	1 = active malignancy
par	Paresis
	0 = no paresis
	1 = paresis
surg	Recent surgery (or bedridden)
	0 = no recent surgery (or bedridden)
	1 = recent surgery (or bedridden)

tend	Tenderness venous system 0 = no localised tenderness deep venous system 1 = localised tenderness deep venous system
leg	Entire leg swollen 0 = entire leg not swollen 1 = entire leg swollen
calfdif	Calf difference
calfdif3	Calf difference >= 3 cm 0 = calf difference < 3 cm 1 = calf difference >= 3 cm
pit	Pitting edema 0 = no pitting edema 1 = pitting edema
vein	Vein distension 0 = no vein distension 1 = vein distension
altdiagn	Alternative diagnosis present 0 = no alternative diagnosis present 1 = alternative diagnosis present
oachst	Oral contraceptives or hst 0 = no oac or hst 1 = oac or hst used
notraum	Absence of leg trauma 0 = leg trauma present 1 = no leg trauma present
eryt	Erythema 0 = no erythema 1 = erythema
histdvt	History of previous DVT 0 = no history of previous DVT 1 = history of previous DVT
histpe	History of previous PE 0 = no history of previous PE 1 = history of previous PE
coag	Family history of thrombofilia 0 = no family history of thrombofilia 1 = family history of thrombofilia
trav	Prolonged traveling 0 = no prolonged traveling 1 = prolonged traveling
pregn	Pregnancy 0 = not pregnant 1 = pregnant
ddim	D-dimer value
ddimdich	Dichotimized d-dimer value 0 = D-dimer negative 1 = D-dimer positive
dvt	Final diagnosis of DVT 0 = no DVT 1 = DVT

Acknowledgements

We would like to thank Stan Buckens for his input and comments during the review process, and Netherlands Organization for Scientific Research (9120.8004, 918.10.615 and 916.11.126).

References

1. Bleeker SE, Moll HA, Steyerberg EW, Donders ART, Derksen-Lubsen G, Grobbee DE, Moons KGM. External validation is necessary in prediction research: a clinical example. *Journal of Clinical Epidemiology* 2003; **56**(9):826–832. DOI: 10.1016/S0895-4356(03)00207-5.
2. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Annals of Internal Medicine* 1999; **130**(6):515–524.
3. Moons KGM, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *British Medical Journal* 2009; **338**:b606. DOI: 10.1136/bmj.b606.
4. Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KGM. Internal and external validation of predictive models: a simulation study of bias and precision in small samples. *Journal of Clinical Epidemiology* 2003; **56**(5):441–447. DOI: 10.1016/S0895-4356(03)00047-7.
5. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer: New York, 2009.
6. Toll DB, Janssen KJM, Vergouwe Y, Moons KGM. Validation, updating and impact of clinical prediction rules: a review. *Journal of Clinical Epidemiology* 2008; **61**(11):1085–1094. DOI: 10.1016/j.jclinepi.2008.04.008.
7. Altman DG. Prognostic models: a methodological framework and review of models for breast cancer. *Cancer Investigation* 2009; **27**(3):235–243. DOI: 10.1080/07357900802572110.
8. Meads C, Ahmed I, Riley RD. A systematic review of breast cancer incidence risk prediction models with meta-analysis of their performance. *Breast Cancer Research and Treatment* 2011:1–13. DOI: 10.1007/s10549-011-1818-2.
9. Perel P, Edwards P, Wentz R, Roberts I. Systematic review of prognostic models in traumatic brain injury. *BMC Medical Informatics and Decision Making* 2006; **6**:38. DOI: 10.1186/1472-6947-6-38.
10. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**(3):177–188.
11. Reitsma JB, Glas AS, Rutjes AWS, Scholten RJPM, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology* 2005; **58**(10):982–990. DOI: 10.1016/j.jclinepi.2005.02.022.
12. Hemingway H, Riley RD, Altman DG. Ten steps towards improving prognosis research. *British Medical Journal* 2009; **339**:b4184. DOI: 10.1136/bmj.b4184.
13. Steyerberg EW, Eijkemans MJ, Van Houwelingen JC, Lee KL, Habbema JD. Prognostic models based on literature and individual patient data in logistic regression analysis. *Statistics in Medicine* 2000; **19**(2):141–160. DOI: 10.1002/(SICI)1097-0258(20000130)19:2⟨141::AID-SIM334⟩3.0.CO;2-O.
14. Steyerberg EW, Borsboom GJJM, van Houwelingen HC, Eijkemans MJC, Habbema JDF. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Statistics in Medicine* 2004; **23**(16):2567–2586. DOI: 10.1002/sim.1844.
15. Van Houwelingen HC, Thorogood J. Construction, validation and updating of a prognostic model for kidney graft survival. *Statistics in Medicine* 1995; **14**(18):1999–2008. DOI: 10.1002/sim.4780141806.
16. Becker BJ, Wu M. The synthesis of regression slopes in meta-analysis. *Statistical Science* 2007; **22**:414–429. DOI: 10.1214/07-STS243.
17. de Leeuw CA, Klugkist I. Augmenting data with published results in Bayesian linear regression. *Master's Thesis*, Faculty of Social and Behavioural Sciences, Utrecht University, Utrecht, The Netherlands, 2010.
18. Higgins J, Thompson S, Deeks J, Altman D. Statistical heterogeneity in systematic reviews of clinical trials: a critical appraisal of guidelines and practice. *Journal of Health Services Research and Policy* 2002; **7**(1):51–61. DOI: 10.1258/1355819021927674.
19. Balázs K, Hidegkuti I, De Boeck P. Detecting heterogeneity in logistic regression models. *Applied Psychological Measurement* 2006; **30**(4):322–344. DOI: 10.1177/0146621605286315.
20. Riley RD, Simmonds MC, Look MP. Evidence synthesis combining individual patient data and aggregate data: a systematic review identified current practice and possible methods. *Journal of Clinical Epidemiology* 2007; **60**(5):431–439. DOI: 10.1016/j.jclinepi.2006.09.009.
21. Marmarou A, Lu J, Butcher I, McHugh GS, Mushkudiani NA, Murray GD, Steyerberg EW, Maas AIR. IMPACT database of traumatic brain injury: design and description. *Journal of Neurotrauma* 2007; **24**(2):239–250. DOI: 10.1089/neu.2006.0036.
22. Steyerberg EW, Mushkudiani N, Perel P, Butcher I, Lu J, McHugh GS, Murray GD, Marmarou A, Roberts I, Habbema JDF, *et al.* Predicting outcome after traumatic brain injury: development and international validation of prognostic scores based on admission characteristics. *PLoS Medicine* 2008; **5**(8):e165. DOI: 10.1371/journal.pmed.0050165.
23. R Development Core Team. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing Vienna Austria*, 2011. URL http://www.r-project.org, ISBN 3-900051-07-0.
24. Jackson D, White IR, Thompson SG. Extending DerSimonian and Laird's methodology to perform multivariate random effects meta-analyses. *Statistics in Medicine* 2010; **29**(12):1282–1297. DOI: 10.1002/sim.3602.
25. Sutton AJ, Kendrick D, Coupland CA. Meta-analysis of individual- and aggregate-level data. *Statistics in Medicine* 2008; **27**(5):651–669. DOI: 10.1002/sim.2916.
26. van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine* 2002; **21**(4):589–624. DOI: 10.1002/sim.1040.
27. Ishak KJ, Platt RW, Joseph L, Hanley JA. Impact of approximating or ignoring within-study covariances in multivariate meta-analyses. *Statistics in Medicine* 2008; **27**(5):670–686. DOI: 10.1002/sim.2913.
28. Riley RD, Lambert PC, Staessen JA, Wang J, Gueyffier F, Thijs L, Boutitie F. Meta-analysis of continuous outcomes combining individual patient data and aggregate data. *Statistics in Medicine* 2008; **27**(11):1870–1893. DOI: 10.1002/sim.3165.

29. Hyder AA, Wunderlich CA, Puvanachandra P, Gururaj G, Kobusingye OC. The impact of traumatic brain injuries: a global perspective. *NeuroRehabilitation* 2007; **22**(5):341–353.
30. Levack WMM, Kayes NM, Fadyl JK. Experience of recovery and outcome following traumatic brain injury: a metasynthesis of qualitative research. *Disability and Rehabilitation* 2010; **32**(12):986–999. DOI: 10.3109/09638281003775394.
31. Jennett B, Teasdale G, Braakman R, Minderhoud J, Knill-Jones R. Predicting outcome in individual patients after severe head injury. *Lancet* 1976; **1**(7968):1031–1034. DOI: 10.1016/S0140-6736(76)92215-7.
32. Mushkudiani NA, Hukkelhoven CWPM, Hernández AV, Murray GD, Choi SC, Maas AIR, Steyerberg EW. A systematic review finds methodological improvements necessary for prognostic models in determining traumatic brain injury outcomes. *Journal of Clinical Epidemiology* 2008; **61**(4):331–343. DOI: 10.1016/j.jclinepi.2007.06.011.
33. Abu-Hanna A, Lucas PJ. Prognostic models in medicine. AI and statistical approaches. *Methods of Information in Medicine* 2001; **40**(1):1–5.
34. Brier GW. Verification of forecasts expressed in terms of probability. *Monthly Weather Review* 1950; **78**(1):1–3. DOI: 10.1175/1520-0493(1950)078⟨0001:VOFEIT⟩2.0.CO;2.
35. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; **143**(1):29–36.
36. Bailey I, Bell A, Gray J, Gullan R, Heiskanan O, Marks PV, Marsh H, Mendelow DA, Murray G, Ohman J, *et al.* A trial of the effect of nimodipine on outcome after head injury. *Acta Neurochir (Wien)* 1991; **110**(3–4):97–105. DOI: 10.1007/BF01400674.
37. The European study group on nimodipine in severe head injury. a multicenter trial of the efficacy of nimodipine on outcome after severe head injury. *Journal of Neurosurgery* 1994; **80**(5):797–804.
38. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *British Medical Journal* 2003; **327**(7414):557–560. DOI: 10.1136/bmj.327.7414.557.
39. Wells PS, Anderson DR, Bormanis J, Guy F, Mitchell M, Gray L, Clement C, Robinson KS, Lewandowski B. Value of assessment of pretest probability of deep-vein thrombosis in clinical management. *Lancet* 1997; **350**(9094): 1795–1798. DOI: 10.1016/S0140-6736(97)08140-3.
40. Wells PS, Anderson DR, Rodger M, Forgie M, Kearon C, Dreyer J, Kovacs G, Mitchell M, Lewandowski B, Kovacs MJ. Evaluation of D-dimer in the diagnosis of suspected deep-vein thrombosis. *The New England Journal of Medicine* 2003; **349**(13):1227–1235. DOI: 10.1056/NEJMoa023153.
41. Gagne P, Simon L, Le Pape F, Bressollette L, Mottier D, Le Gal G. [Clinical prediction rule for diagnosing deep vein thrombosis in primary care]. *La Presse Médicale* 2009; **38**(4):525–533. DOI: 10.1016/j.lpm.2008.09.022.
42. Subramaniam RM, Snyder B, Heath R, Tawse F, Sleigh J. Diagnosis of lower limb deep venous thrombosis in emergency department patients: performance of Hamilton and modified Wells scores. *Annals of Emergency Medicine* 2006; **48**(6):678–685. DOI: 10.1016/j.annemergmed.2006.04.010.
43. Geersing G-J, Janssen KJ, Oudega R, van Weert H, Stoffers H, Hoes A, Moons K, on behalf of the AMUSE Study. Diagnostic classification in patients with suspected deep venous thrombosis: physicians' judgement or a decision rule? *Journal of the Royal College of General Practitioners* 2010; **60**(579):742–748. DOI: 10.3399/bjgp10X532387.
44. Oudega R, Moons KGM, Hoes AW. Ruling out deep venous thrombosis in primary care. a simple diagnostic algorithm including D-dimer testing. *Thrombosis and Haemostasis* 2005; **94**(1):200–205. DOI: 10.1160/TH04-12-0829.
45. Büller HR, Ten Cate-Hoek AJ, Hoes AW, Joore MA, Moons KGM, Oudega R, Prins MH, Stoffers HEJH, Toll DB, van der Velde EF, *et al.* Safely ruling out deep venous thrombosis in primary care. *Annals of Internal Medicine* 2009; **150**(4):229–235. DOI: 10.1059/0003-4819-150-4-200902170-00003.
46. Toll DB, Oudega R, Vergouwe Y, Moons KGM, Hoes AW. A new diagnostic rule for deep vein thrombosis: safety and efficiency in clinically relevant subgroups. *Family Practice* 2008; **25**(1):3–8. DOI: 10.1093/fampra/cmm075.
47. Riley RD, Abrams KR, Sutton AJ, Lambert PC, Thompson JR. Bivariate random-effects meta-analysis and the estimation of between-study correlation. *BMC Medical Research Methodology* 2007; **7**:3. DOI: 10.1186/1471-2288-7-3.
48. Janssen KJM, Vergouwe Y, Kalkman CJ, Grobbee DE, Moons KGM. A simple method to adjust clinical prediction models to local circumstances. *Canadian Journal of Anaesthesia* 2009; **56**(3):194–201. DOI: 10.1007/s12630-009-9041-x.
49. Jackson D, Riley R, White IR. Multivariate meta-analysis: Potential and promise. *Statistics in Medicine* 2011; **30**(20):2481–2498. DOI: 10.1002/sim.4172.
50. Steyerberg EW, Eijkemans MJ, Habbema JD. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. *Journal of Clinical Epidemiology* 1999; **52**(10):935–942. DOI: 10.1016/S0895-4356(99)00103-1.
51. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* 2007; **115**(7): 928–935. DOI: 10.1161/CIRCULATIONAHA.106.672402.
52. Pencina MJ, D'Agostino RBS, D'Agostino RBJ, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in Medicine* 2008; **27**(2):157–172. DOI: 10.1002/sim.2929.