Meta-analysis of clinical prediction models

Thomas P. A. Debray

Julius Center for Health Sciences and Primary Care

Utrecht University

Meta-analysis of clinical prediction models

Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht.
PhD Thesis. University Utrecht, Faculty of medicine, with a summary in Dutch.

ISBN	978-90-393-6015-6
Cover Design	Todor Krastev
Layout	Thomas Debray
Print	Ipskamp Drukkers BV

Meta-analysis of clinical prediction models

Meta-analyse van klinische prediciemodellen

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof. dr. G.J. van der Zwaan, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op dinsdag 29 oktober 2013 des middags te 2.30 uur

door

Thomas Paul Alfons Debray
geboren op 18 juni 1985 te Leuven, België

Promotoren:	Prof.dr. K.G.M. Moons
	Prof.dr. E.W. Steyerberg
Co-promotoren:	Dr. ir. H. Koffijberg
	Dr. Y. Vergouwe

The studies in this thesis were funded by the Netherlands Scientific Organization (ZonMw 9120.8004, 918.10.615 and 916.11.126).
Additional financial support from stichting BAZIS for the publication of this thesis is gratefully acknowledged.

Contents

Part I

General Introduction

"Whosoever desires constant success must change his conduct with the times."

– Niccolo Machiavelli

During the most recent decades, the impact of statistical modeling techniques on clinical decision making has increased profoundly. The development of clinical prediction models deserves particular attention as it aims to facilitate medical decision making through explicit modeling of uncertainty. Clinical prediction models are, for instance, commonly used to estimate a disease's or outcome's presence (diagnosis) or future occurrence (prognosis) in an individual before deciding further management. Typically, prediction models rely on several predictors. These may range from individual characteristics, signs and symptoms, to results of more invasive or costly measures such as imaging and biomarker test results. By relying on statistical methods and empirical data, prediction models enhance the objectivity and transparency of medical decision making. In addition, they may lead to improved patient care if their predictions are sufficiently accurate and effectively applied by professionals, and if adequate decision thresholds have been established.

The performance of prediction models increases as larger datasets are used during model development and relevant predictors are included in the model. However, in medical research practical constraints on time and cost typically do not permit all potential predictors in large numbers of individuals to be collected for each (model development) study. As a consequence, many prediction models are often developed from relatively small(er) datasets, relying on predictor selection strategies which may compromise valid probability estimations. Indeed, there are many examples of so developed prediction models that perform more poorly than anticipated when actually implemented in other datasets or in routine care. Researchers then frequently responded by developing a new prediction model in their own dataset, even when previous similar models are readily available [132, 171, 239]. For example, there are over 60 published models aiming to predict outcome after breast cancer [11], over 100 for predicting long-term outcome in neurotrauma patients [186], over 20 for identifying patients at risk of prolonged stay at the Intensive Care Unit [75] and over 20 for predicting the risk of cardiovascular disease in patients with type 2 diabetes [269]. This practice of continuously developing new duplicative models for the same outcome or target population based on new, often small, datasets ignores the previously collected scientific evidence.

Meta-analysis has become a powerful tool to combine the results from different studies on a similar topic [31, 70, 106, 109, 252]. This practice is now commonly applied in therapeutic intervention research where effect estimates for a particular intervention from different studies are combined and synthesized. The meta-analytical paradigm is also highly relevant to diagnostic and prognostic prediction research, where aggregate data (AD) such as predictor-outcome associations and previously published prediction models are often available from the literature. Sometimes, raw data with measurements on subject level can also be obtained from other research groups in the field, and is commonly denoted as individual participant data (IPD). Meta-analytical methods may therefore be considered to synthesize IPD and AD, and to derive more up-to-date and better generalizable prediction models. Because meta-analysis generally involves data aggregation, it increases the effective sample size during the prediction model's development phase [278]. This, in turn, may improve the identification of important predictors and the accuracy of estimated predictor-outcome associations in future or other subjects. In short, meta-analytical approaches can considerably increase the efficiency (and cost-effectiveness) of diagnostic and prognostic prediction modeling research by making more effective use of available evidence.

We distinguish the following types of data when considering a meta-analysis in clinical prediction research:

- IPD (individual participant data) is available from one or multiple studies.
- AD (aggregate data) is available from one or multiple studies and may consist of:
  - univariable predictor-outcome associations
  - previously published (multivariable) prediction models with the same predictors
  - previously published (multivariable) prediction models with different predictors
- A combination of IPD and AD is available from one or multiple studies. Because IPD can always be transformed into AD, this thesis does not evaluate how IPD and AD can directly be combined (although such methods may further improve the synthesis process) [204, 207, 254].

OUTLINE OF THIS THESIS

The studies presented in this thesis seek to investigate some of these key issues.

Chapters 1–4 present several approaches for IPD meta-analysis (Section II). **Chapter 1** assesses the effect of ignoring clustering of participants within studies in an IPD-MA. **Chapter 2** illustrates how clusters can be retained during risk-factor or predictor finding studies by using a one-stage or two-stage approach. One-stage methods use the IPD of each study and meta-analyze using the exact binomial distribution, whereas two-stage methods first reduce evidence to the aggregated level (e.g. odds ratios) and then meta-analyse assuming approximate normality. Subsequently, **Chapter 3** presents a framework for developing prediction models when IPD is available from multiple studies. Finally, **Chapter 4** describes how a prediction model can be developed in an IPD meta-analysis when some predictor variables have not been measured in each study. As a whole, these chapters explore IPD meta-analysis techniques and how they may be best implemented.

Chapters 5–7 propose several approaches for developing a prediction model when AD is available (Section III). **Chapter 5** considers the situation in which the AD consists of univariate predictor-outcome associations. **Chapter 6** considers the situation in which the AD consists of previously published prediction models with the same predictors. Finally, **Chapter 7** considers the situation in which the AD consists of previously published prediction models with different predictors. Together, these papers address AD meta-analysis techniques.

The **General Discussion** addresses some other issues in the performance of prediction models and the extent to which validation studies may provide insight into their generalizability.

Part II

Individual Participant Data Meta-Analysis

1

Individual participant data meta-analyses should not ignore clustering

Journal of Clinical Epidemiology 2013, **66**(8): 865–873.

Ghada M. A. Abo-Zaid
BiWei Guo
Jon J. Deeks
Thomas P. A. Debray
Ewout W. Steyerberg
Karel G. M. Moons
Richard D. Riley

Abstract

Individual participant data (IPD) meta-analyses often analyse their IPD as if from a single study. We compare this approach to analyses that rather account for clustering of patients within studies by applying logistic regression models in real and simulated examples. Results indicate that the estimated prognostic effect of age in patients with traumatic brain injury is similar regardless of whether clustering is accounted for. However, a family history of thrombophilia is found to be a diagnostic marker of deep vein thrombosis (odds ratio = 1.30, 95% CI: 1.00 to 1.70; p = 0.05) when clustering is accounted for, but not when it is ignored (odds ratio = 1.06, 95% 0.83 to 1.37; p = 0.64). Similarly, the treatment effect of nicotine gum on smoking cessation is severely attenuated when clustering is ignored (odds ratio = 1.40; 95% CI: 1.02 to 1.92) rather than accounted for (odds ratio = 1.80; 95% CI: 1.29 to 2.52). Simulations show models accounting for clustering perform consistently well, but downwardly biased effect estimates and low coverage can occur when ignoring clustering. In conclusion, researchers must routinely account for clustering in IPD meta-analyses, otherwise misleading effect estimates and conclusions may arise.

Chapter 1

"The aim of science is not to open the door to infinite wisdom, but to set a limit to infinite error."

– Bertolt Brecht, *Life of Galileo*

INDIVIDUAL Participant Data Meta-Analysis (IPD-MA) refers to when participant-level data are obtained from multiple studies and then synthesised [203]. This contrasts the usual meta-analysis approach, which obtains and then synthesises aggregate data (such as a treatment effect estimates) extracted from study publication or study authors [235]. IPD offers many potential advantages for the meta-analyst [203, 235, 236]; in particular it reduces reliance on the reporting quality of individual studies as, with the raw data at hand, the meta-analyst can be more flexible and consistent in their choice of analysis method, can estimate directly the effect estimates of interest, and better account for study heterogeneity and subgroup effects.

Methods for IPD-MA use either a one-step or two-step approach [226]. In the two-step approach, the Individual Participant Data (IPD) are first analysed separately in each study using an appropriate statistical method for the type of data being analysed. For example, to assess the association between a continuous factor (e.g. age) and the odds of a binary outcome (e.g. death) a logistic regression model might be fitted, to produce aggregate data for each study, such as the odds ratio and its associated standard error; these are then synthesised in the second step using a suitable model for meta-analysis of aggregate data, such as one weighting by the inverse of the variance whilst assuming fixed or random effects across studies. In the one-step approach, the IPD from all studies are modelled simultaneously; this again requires a model specific to the type of data being synthesised, alongside appropriate specification of the meta-analysis assumptions (e.g. fixed or random-effects across studies). Clustering of patients within studies can be accounted for by stratifying the analysis by study (i.e. by estimating a separate intercept for each study) or by assuming the study intercepts (baseline risk) are randomly drawn from some distribution.

Many existing articles discuss the implementation and merits of one-step and two-step IPD-MA methods [115, 135, 200, 204, 264, 266, 288], and the methods often give very similar results [156, 181, 204]. For example for time-to-event data, Tudur Smith and Williamson [263] show through simulation that when there is no heterogeneity in effect and the proportional hazards assumption holds, a one-step stratified Cox model produces similar effect estimates to the two-step (inverse variance weighted) approach. For continuous outcome data analysed using linear models, Olkin and Sampson [181] and subsequently Matthew and Nordström [156, 157], show that the one-step and two-step approaches provide identical results when estimating a treatment effect under certain theoretical conditions, although when covariates are added differences may occur. Jones *et al.* [135] consider longitudinal continuous outcome data and empirically show that the one-step and two-step approaches produce similar effect estimates, as long as correlations between time-points are incorporated. For binary outcome data, there may be some advantage of a one-step approach when the event risk or rate is low or the sample size is small; in contrast to the two-step approach, the one-step approach allows the exact binomial distribution to be used and does not require continuity corrections when zero events occur [98, 250].

However, potentially of more concern than the choice of one-step or two-step approach, is that there is growing evidence that researchers undertake the one-step approach but ignore the clustering of patients within studies, thereby treating the IPD as if it all came from one study. For example, Simmonds et al. examined IPD-MA of randomised trials, and found that 3 of 14 using a one-step approach ignored clustering [226]. Similarly, Abo-Zaid *et al.* examined IPD-MA of prognostic factor studies, and found that 5 of 11 using a one-step approach did not state they accounted for clustering [3].

Using real examples and through simulation, we therefore studied the potential impact of ignoring clustering on IPD-MA results, and report our findings in this article. We focus on IPD-MA aimed at quantifying whether a single (continuous or binary) factor or determinant of interest is associated with (the odds of) a binary outcome. For example, one may wish to summarise the outcome risk in a treatment group relative to the control group (i.e. estimate a treatment effect); estimate whether a certain prognostic marker is associated with future event risk (i.e. estimate a prognostic effect); or quantify whether the presence of a certain diagnostic test result increases or decreases the probability of having a particular disease. These are common situations in the (IPD) meta-analysis field. We first introduce three one-step and two-step models of interest, and subsequently apply them to three real applications. Finally, we evaluate the performance of the one-step methods is through simulation and conclude with discussion and recommendations.

ONE-STEP AND TWO-STEP IPD-MA APPROACHES

Consider there are $i = 1$ to m independent studies that each assess the binary outcome of interest for n_i participants. Let y_{ik} be the outcome (1 = event, 0 = no event) of participant k in study i, where $k = 1$ to n_i, and let x_{ik} be a participant-level factor (covariate), which could be continuous or binary. We term an "IPD study" one that provides y_{ik} and x_{ik} for the n_i participants in the study. Note that, for a binary factor, if the number of participants and events for each of the two categories are known, then IPD for these two variables can simply be reconstructed by creating a row for each participant and delegating them event responses and covariate status that collectively mirror the observed frequencies.

Given such IPD, there are a number of ways researchers could estimate the summary risk or odds ratio across studies. We focus here on the use of a logistic regression framework, via either a one-step approach ignoring clustering, a one-step approach accounting for clustering, or a two-step approach, as now described.

Model (1): one-step ignoring clustering

With this method, the IPD from all studies are stacked and analysed together as if they were a single study, thus the clustering of patients within different studies is ignored. The standard logistic model can be written as:

$$\begin{aligned} y_{ik} &\sim \text{Bernoulli}\,(p_{ik}) \\ \text{logit}\,(p_{ik}) &= \alpha + \beta' x_{ik} \end{aligned} \tag{1.1}$$

The common α term for all studies shows that clustering is being ignored, and α can be interpreted as the log-odds of the event for patients with x_{ik} equal to zero. The term β provides the log odds ratio, comparing the odds of the event for two patients who differ in x_{ik} by one unit. Note that β is also assumed common to all studies, and so we have a fixed effect meta-analysis here. We consider a random effects approach and multivariable model extensions in our Discussion.

Model (2): one-step accounting for clustering

Here, the IPD from all studies are also stacked and analysed together, but the clustering of patients within different studies is accounted for. The logistic model can be written as:

$$\begin{aligned} y_{ik} &\sim \text{Bernoulli}\,(p_{ik}) \\ \text{logit}\,(p_{ik}) &= \alpha_i + \beta' x_{ik} \end{aligned} \tag{1.2}$$

Now the intercept term is not fixed, and α_i gives the log-odds of the event in study i for those participants with x_{ik} equal to zero. The separate α_i term for each study shows that clustering per study is being accounted for at the baseline level, i.e. each study is allowed to have their own baseline risk.

Model (3): two-step approach

Here, the IPD of each study is analysed separately, and the log odds ratio estimates from each study are then combined (averaged) in an inverse-variance weighted fixed effect meta-analysis, as follows.

STEP 1 (each study separately):

$$\begin{aligned} y_{ik} &\sim \text{Bernoulli}(p_{ik}) \\ \text{logit}\,(p_{ik}) &= \alpha_i + \beta_i' x_{ik} \end{aligned} \tag{1.3}$$

$$\tag{1.4}$$

STEP 2 (meta-analysis of aggregate data):

$$\begin{aligned} \hat{\beta}_i &= \beta + \epsilon_i \\ \epsilon_i &\sim \mathcal{N}\left(0, \text{var}(\hat{\beta}_i)\right) \end{aligned} \tag{1.5}$$

By first analysing each study separately, this approach automatically accounts for the clustering of patients within studies. In the second step, the var($\hat{\beta}_i$) estimates are assumed known, which is a common assumption in the meta-analysis field, and the pooled prognostic effect estimate ($\hat{\beta}$) will be a weighted average of the $\hat{\beta}_i$ s, with study weights equal to the inverse of var($\hat{\beta}_i$).

The parameters in equations 1.1 and 1.2, and those in both steps of Model (3), can be estimated using maximum likelihood estimation. Models (1), (2) and (3) Step 1 are using a logistic regression model framework, available in most statistical software packages, and Model 3 Step 2 can be fitted using standard meta-analysis modules, such as *metan* in STATA [231]. Note that, when x_{ik} is a binary factor and the event risk is low and/or the sample size is small, some studies may have zero events for one of the factor's groups. The one-step approach accommodates such studies automatically through their contribution to the likelihood. However, the two-step approach first requires a so-called continuity correction (e.g. 0.5) to be added to all cells in such studies, in order to estimate a sensible log odds ratios and its standard error. This is a clear limitation of the two-step method and this issue has been well discussed in the literature [35], and is not the focus of this article. We only consider examples without zero cells in this article.

EMPIRICAL IPD META-ANALYSIS EXAMPLES

We now introduce three motivating IPD-MA examples to illustrate the potential similarities and differences of the models in meta-analyses of diagnostic studies, prognostic studies, and (randomised) therapeutic trials.

Mortality after traumatic brain injury

Hukkelhoven *et al.* [120] performed a meta-analysis of 14 prospective studies to assess the 6-month mortality risk in patients with traumatic brain injury. Their key objective was to examine the association between age and 6-month mortality risk. Biologically this relationship is plausible, as the adult brain is hypothesised to have decreased capacity for repair as it ages [49], due to a decreasing number of functioning neurons and a greater exposure to minor repetitive insults to the brain as age increases. In their meta-analysis, IPD were available for four studies (totalling 2 659 patients), containing the 6-month mortality outcome (dead or alive) and age for each patient in each study. This IPD is summarised in our e-Appendix [2].

Of interest is the odds ratio comparing the odds of death by 6 month for two patients aged 10 years apart. Only a linear relationship with age was assumed. The results for each of models (1) to (3) are shown in Table 1.1, and there are only small, unimportant statistical and clinical differences between them. Age is identified to have a statistically significant ($p < 0.001$) association with the odds of 6-month mortality in all models, and the odds ratio is 1.41 in the one-step model ignoring clustering, and a slightly lower 1.37 in the two-step approach and one-step accounting for clustering. The standard error of the log odds ratio estimate is almost identical, 0.030 in the two-step and 0.029 in the others. There was no evidence of between-study heterogeneity in the odds ratio ($I^2 = 0$), suggesting the fixed effect modelling assumption was appropriate. Based on this application alone, the observed findings might lead researchers to decide that it does not matter whether clustering is accounted for.

Diagnosis of deep vein thrombosis

IPD are available from six studies of patients with suspected deep vein thrombosis (DVT) [17, 140, 232, 258, 261, 282], and of interest is whether a family history of thrombophilia (defined as yes or no) is associated with the risk of truly having DVT. One might expect patients with a family history of thrombophilia to be more likely to have a genuine DVT than those without. The studies are summarised in Table A.1 in the Appendix, and contained a total of 4 599 patients of which

Table 1.1: Results from the empirical examples

Example	Method	$\hat{\beta}$ (SE)	OR	95% CI	p-value
1	Two-step	0.316 (0.030)	1.372	1.295 to 1.454	< 0.001
	One-step ignoring clustering	0.341 (0.029)	1.407	1.329 to 1.488	< 0.001
	One-step accounting for clustering	0.317 (0.029)	1.373	1.296 to 1.455	< 0.001
2	Two-step	0.280 (0.135)	1.323	1.015 to 1.725	0.038
	One-step ignoring clustering	0.060 (0.128)	1.062	0.825 to 1.365	0.642
	One-step accounting for clustering	0.263 (0.136)	1.301	0.996 to 1.697	0.053
3	Two-step	0.570 (0.174)	1.769	1.257 to 2.488	0.001
	One-step ignoring clustering	0.355 (0.161)	1.400	1.020 to 1.916	0.037
	One-step accounting for clustering	0.589 (0.170)	1.802	1.290 to 2.517	0.001

SE = standard error; OR = odds ratio; CI = confidence interval for the odds ratio

[1] Traumatic Brain Injury results for the association between age/10 and the odds of 6-month mortality, for each of the three IPD models.

[2] Results for the effect of a family history of thrombophilia on the odds of truly having deep vein thrombosis, for each of the three IPD models.

[3] Results for the effect of nicotine gum on the odds of giving up smoking.

909 (19.8%) truly have DVT [2]. The proportion of patients in each study with a family history of thrombophilia ranged from 0.03 to 0.26.

As in the TBI example, there is no heterogeneity ($I^2 = 0\%$) and the two-step approach and the one-step approach accounting for clustering obtain similar estimates, standard errors and confidence intervals (Table 1.1); they estimate that the odds of DVT are about 1.3 times higher for patients with a family history of thrombophilia, and the findings are (close to) statistically significant at the 5% level ($p = 0.038$ or 0.053). However, the one-step approach ignoring clustering estimates a much smaller odds ratio of 1.06, and there is now no statistically significant evidence that family history is an important risk factor (p = 0.64); the standard error of is also smaller compared to the other models. Thus, in this example the one-step approach ignoring clustering provides different statistical and clinical conclusions than the other approaches.

Smoking cessation and use of nicotine gum

Rice and Stead [194] perform a meta-analysis of 51 randomised trials to examine whether the use of nicotine gum increases the chances of stopping smoking. Altman and Deeks [13] used these trials to show the impact on the estimated number needed to treat when clustering of studies was ignored. We now extend this to consider the impact on the odds ratio. Specifically, for illustrative purposes we consider a meta-analysis of just two of the trials (the same two used by Altman and Deeks), which are summarised in our e-Appendix [2] and the results shown in Table 1.1 ($I^2 = 14.3\%$). As in the DVT example, the one-step method ignoring clustering produces a smaller summary odds ratio (1.48) that is much closer to 1 than the other methods, which rather give estimates around 1.8 with wider confidence intervals.

SIMULATION METHODS

The above examples illustrate that the decision to account for clustering in IPD meta-analysis is potentially important. To look more generally at how ignoring clustering affects the statistical properties of estimates, we now present a simulation study of models (1) and (2).

Simulation procedure

Full details of our simulation are provided in our e-Appendix [2]. Briefly, for multiple scenarios we simulated IPD (i.e. patient outcomes and prognostic factor values) for meta-analyses based on $m =$ 5 or 10 studies; small (30 to 100 patients) or larger study sizes (upto 1000 patients); a continuous or binary factor (x_{ik}); a binary outcome y_{ik} (1 = event, 0 = alive) where $y_{ik} \sim \text{Bernoulli}(p_{ik})$ and $\text{logit}(p_{ik}) = \alpha_i + \beta x_{ik}$; the chosen parameters of $\alpha_i \sim \mathcal{N}(\alpha, \sigma_\alpha^2)$; and for binary factors a β of either 0, 0.1, or 0.9 (relating to an odds ratio of 1, 1.1, and 2.45 respectively) and for continuous factors a β of either 0 (no effect), 0.1 (small effect) or 0.3 (large effect).

All scenarios considered are listed in e-Appendix [2]. In each scenario, we generated 1 000 IPD-MA datasets and then fitted Models (1) and (2) to each, and recorded $\hat{\beta}$ and its standard error. Each model's performance was then examined by calculating the bias, mean-square error (MSE), mean standard error and coverage for $\hat{\beta}$.

Simulation results

The simulation results for scenarios with five studies and small samples sizes are summarised in Table 1.2, Table 1.3 and e-Appendix [2]. The findings were very similar when the number of studies was changed to 10, or when a larger sample size was allowed.

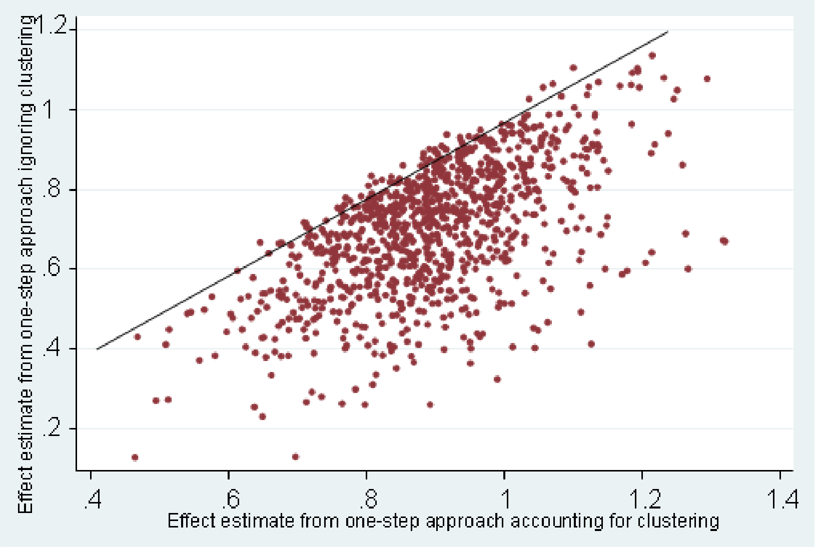
For both binary and continuous factors, when there was zero or small variation in baseline risk (α_i) the performance of the models was very similar. The bias in $\hat{\beta}$ was close to zero, the MSE was approximately the same, and the coverage was always close to 95%. When the variation in α_i was large (scenarios 13 to 18, 22 to 24), the one-step approach accounting for clustering continues to perform consistently well with suitable bias and coverage. However, the one-step approach ignoring clustering often performs poorly, with downward bias and low coverage especially when the true effect size was large. For example, in scenario 13 (where the true β was 0.9), the one-step model ignoring clustering has a large downward bias of -0.21 and a low coverage of 87.6%, reflecting a small mean standard error. This scenario is illustrated in Figure 1.1, which shows the one-step approach ignoring clustering produces smaller standard errors in each meta-analysis and generally (though not always) smaller effect estimates than the one-step approach accounting for clustering.

Link to the applied examples

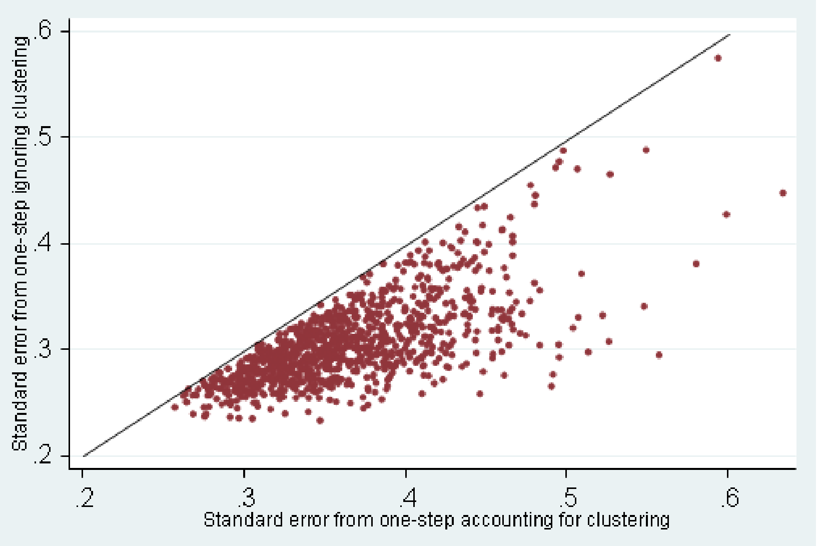
When the two-step approach was fitted to the TBI data, step one produced separate alpha estimates in each study. The weighted average of these alphas was -2.1, and their between-study standard deviation was 0.20. Thus the TBI data mirrors closely simulation scenario 19 (Table 1.2), where alpha was -2.1, the standard deviation of alpha was 0.2, and the true effect was 0.3. In this scenario there was no difference between models (1) and (2) in terms of bias, MSE and coverage, and so it is unsurprising that the TBI application shows very similar model (1) and (2) results.

Figure 1.1: Simulation results for scenario 13.

(a) Effect Estimates



(b) Standard error of effect estimates



Comparison of the 1 000 simulation results from the one-step accounting clustering versus the one-step ignoring clustering for scenario 13 with 5 studies, small study sample sizes and a binary factor, where the standard deviation of alpha was 1.5 (i.e. $\sigma_\alpha = 1.5$), the true beta was 0.9 (i.e. $\beta = 0.9$), and the prevalence was 0.2 (i.e. $\pi = 0.2$).

In contrast to the TBI example, the DVT and smoking applications showed that ignoring clustering produced a substantially smaller odds ratio estimate and a smaller standard error of $\hat{\beta}$ than other methods (Table 1.1). Variability in baseline risk with only a small number of studies is a potential cause of these differences, and - in accordance with some of the simulation results in this situation (e.g. Figure 1.1) - ignoring clustering appears to be producing estimates with a downward bias and low coverage in these examples. Other mechanisms may also be causing differences to occur in these examples, beyond those identified by our simulations, such as between-study variation in the proportion of patients who are factor positive [13].

DISCUSSION

IPD-MA are increasingly used. Riley *et al.* [203] found 383 published in the medical literature before March 2009, with an average of 49 published per year since 2005. In this article we have examined the impact of ignoring clustering of patients within studies when analysing IPD of multiple studies with binary outcomes, where an odds ratio is of interest. In some situations statistical inferences do not alter whether clustering is accounted for, as seen in the TBI application. However, there are situations when the approaches can differ substantially in their performance and this can impact upon statistical and clinical inferences. This was seen in the DVT and smoking examples, and in our simulations with large between-study variability in baseline risk.

There are two key recommendations from our work. The first is that it is inappropriate to simply ignore the clustering of patients within studies and to analyse the IPD as if coming from a single study. When there is large variability in baseline risk, the simulations show that this nave approach leads to a downward bias, with small standard errors that produce a low coverage substantially less than 95%; this problem appears to become worse as the true effect size increases. The DVT example shows that ignoring clustering would lead to a potentially important diagnostic marker for DVT being missed, whilst in the smoking example the effect of nicotine gum on smoking cessation would have been severely underestimated. Other articles in non meta-analysis settings have also identified the danger of ignoring clustering, such as in cluster randomised trials [29, 187] and multicentre randomised trials [143]. Steyerberg *et al.* [240] show that in a logistic regression analysis of a clinical trial with multiple strata, the odds ratio of 0.853 when ignoring clustering is reduced to 0.820 when adjusting for strata, an increase of 25% on the logistic scale. Similarly, Hernandez *et al.* [110] and Turner *et al.* [265] show that adjustment for prognostic covariates in logistic regression increases power to detect a genuine effect. Statistically speaking, by ignoring clustering one specifies a marginal model which assumes all studies have the same baseline risk, but by accounting for clustering one specifies a conditional model that correctly conditions each

Table 1.2: Simulation results for some of the scenarios.

Sc.	Meta-analysis model	Parameters				Estimates for β				
		α	σ_α	π	β	Mean	Bias	MSE	Coverage	SE
Binary factor										
1	One-step ignoring clustering	-1.27	0	0.5	0.9	0.91	0.01	0.03	94.90 %	0.16
	One-step accounting for clustering	-1.27	0	0.5	0.9	0.92	0.02	0.03	94.70 %	0.16
3	One-step ignoring clustering	-1.27	0	0.5	0	0.00	0.00	0.02	94.90 %	0.16
	One-step accounting for clustering	-1.27	0	0.5	0	0.00	0.00	0.02	94.90 %	0.16
13	One-step ignoring clustering	-1.27	1.5	0.2	0.9	0.69	-0.21	0.15	87.60 %	0.31
	One-step accounting for clustering	-1.27	1.5	0.2	0.9	0.92	0.02	0.14	94.80 %	0.36
15	One-step ignoring clustering	-1.27	1.5	0.2	0	-0.02	-0.02	0.22	94.00 %	0.33
	One-step accounting for clustering	-1.27	1.5	0.2	0	0.00	0.00	0.26	94.00 %	0.38
16	One-step ignoring clustering	-1.27	1.5	0.5	0.9	0.70	-0.20	0.04	46.20 %	0.09
	One-step accounting for clustering	-1.27	1.5	0.5	0.9	0.90	0.00	0.05	94.90 %	0.11
18	One-step ignoring clustering	-1.27	1.5	0.5	0	0.00	0.00	0.04	94.90 %	0.09
	One-step accounting for clustering	-1.27	1.5	0.5	0	0.00	0.00	0.05	94.70 %	0.11

Sc=Scenario

We used $m = 5$ studies in the meta-analysis, with each a sample size of $n_i \sim \mathcal{U}(30, 100)$. For each patient in each trial, x_{ik} was randomly sampled according to $x_{ik} \sim$ Bernoulli(π) (binary factors) or $x_{ik} \sim \mathcal{N}(4, 1.5^2)$ (continuous factors). The outcome y_{ik} (1 = event, 0 = alive) was randomly sampled according to $y_{ik} \sim$ Bernoulli(p_{ik}) where $\text{logit}(p_{ik}) = \alpha_i + \beta x_{ik}$ and $\alpha_i \sim \mathcal{N}(\alpha, \sigma_\alpha^2)$. Each scenario was repeated 1 000 times. The means of the resulting estimates for β and their respective standard errors are presented here as 'Mean' and 'SE'. We also calculated the bias, Mean Squared Error (MSE) and coverage of all the estimates.

Table 1.3: Simulation results for some of the scenarios.

Sc.	Meta-analysis model	Parameters				Estimates for β				
		α	σ_α	π	β	Mean	Bias	MSE	Coverage	SE
Continuous factor										
19	One-step ignoring clustering	-2.10	0.2		0.30	0.30	0.00	0.01	96.29 %	0.09
	One-step accounting for clustering	-2.10	0.2		0.30	0.31	0.01	0.01	96.36 %	0.09
21	One-step ignoring clustering	-2.10	0.2		0	0.00	0.00	0.00	95.10 %	0.12
	One-step accounting for clustering	-2.10	0.2		0	0.00	0.00	0.00	94.90 %	0.12
22	One-step ignoring clustering	-2.10	1.5		0.30	0.23	-0.07	0.01	84.10 %	0.09
	One-step accounting for clustering	-2.10	1.5		0.30	0.31	0.01	0.01	94.80 %	0.10
24	One-step ignoring clustering	-2.10	1.5		0	0.00	0.00	0.01	95.40 %	0.11
	One-step accounting for clustering	-2.10	1.5		0	0.00	0.00	0.02	95.60 %	0.12

Sc=Scenario
We used $m = 5$ studies in the meta-analysis, with each a sample size of $n_i \sim \mathcal{U}(30, 100)$. For each patient in each trial, x_{ik} was randomly sampled according to $x_{ik} \sim \text{Bernoulli}(\pi)$ (binary factors) or $x_{ik} \sim \mathcal{N}(4, 1.5^2)$ (continuous factors). The outcome y_{ik} ($1 =$ event, $0 =$ alive) was randomly sampled according to $y_{ik} \sim \text{Bernoulli}(p_{ik})$ where $\text{logit}(p_{ik}) = \alpha_i + \beta x_{ik}$ and $\alpha_i \sim \mathcal{N}(\alpha, \sigma_\alpha^2)$. Each scenario was repeated 1 000 times. The means of the resulting estimates for β and their respective standard errors are presented here as 'Mean' and 'SE'. We also calculated the bias, Mean Squared Error (MSE) and coverage of all the estimates.

patients response on the study there are in. For logistic models, Robinson and Jewell [209] have shown that marginal models give potentially attenuated (biased) effect estimates and have lower power to detect genuine effects than conditional models. For logistic regression, this phenomenon is also known as non-collapsibility of the odds ratio , as conditional odds ratios are typically larger than marginal odds ratios after conditioning on important covariates, with the increase becoming higher as the true odds ratio increases and the number of included important covariates increases. Gail et al. showed analytically and through simulation that Cox and exponential regression models for survival data with censoring also produce downwardly biased treatment effect estimates when important covariates are omitted, unless the true treatment effect is zero or close to zero. For linear regression or generalised linear models with a log link (e.g. Poisson regression) the asymptotic bias from omitting covariates is zero regardless of the true effect size ; yet, even for such models the precision of effect estimates can still be severely affected by ignoring important covariates (clustering) [209]. Statisticians thus may not be surprised by our findings, but we hope our findings raise awareness to the IPD-MA community, many of whom currently ignore clustering [3, 226]. We thus recommend researchers always account for clustering in their IPD-MA, and report how they did so in any subsequent publication.

The second important finding is that the one-step model accounting for clustering performs consistently well in all simulations considered, with bias close to zero and suitable coverage. Based on this, we recommend this method be routinely chosen to analyse IPD with binary outcomes. The two-step method will often give very similar results, as seen in the examples of previous sections. However, the one-step approach models the exact binomial nature of the data directly [98, 250], whilst the two-step approach produces log odds ratio estimates in the first step, which are then assumed normally distributed in the second step. This additional normality assumption may be inappropriate when the number of patients in studies is small and/or when the number of events is small. For this reason the exact one-stage approach of model (1) is generally more suitable for synthesising two by two tables. The Mantel-Haenszel method or Peto method have also been suggested to overcome this issue [151, 295], but model (1) can more easily be extended to include multiple factors and continuous variables, so is our preferred method. It can also be easily extended to allow between-study heterogeneity in the effect of interest [250]. One could also allow a random-effects distribution on the baseline risk, rather than estimating a separate α_i for each study. This requires an sadditional distributional assumption to be made for α_i s, and for this reason we prefer model (1) as described above. A distribution on the baseline risk is perhaps useful if the baseline risk is itself of interest, but in our examples the focus was only on the effect of the included factor.

Note that it is not possible to predict the direction of bias induced by ignoring clustering in any single example. For example, our simulations with large variability in baseline risk show that ignoring clustering leads to a downward bias *on average*, but Figure 1.1 highlights that in a sole application the actual estimates when ignoring clustering may occasionally be larger than when accounting for clustering. Indeed, the TBI application had a slightly higher odds ratio when ignoring clustering. Our simulations are also limited to particular choices of parameter values and, like all simulation studies, other permutations of values and alternative scenarios In particular, between-study variation in prevalence of the binary factor and/or between-study heterogeneity in effect may reveal different findings.

None of our binary factor examples or simulations contained studies with zero events in a particular group, as this issue has been examined before [35] and been shown to induce bias in the two-step approach as, unlike the one-step approach, it requires a continuity correction to be added. Our simulations and examples also did not consider between-study heterogeneity in effects, but our recommendations are likely to generalise to this setting also [51, 98]. We also recognise that IPD meta-analyses are not without limitations. Some covariates may not be available for all IPD studies [78], and IPD may not be available from all studies requested [7]. In this situation novel methods may be required to synthesise the IPD effectively [207].

In conclusion, we have shown that researchers synthesising IPD from multiple studies should account for the clustering of patients within different studies. Lumping the IPD into a single dataset and naively analysing as if from a single study can produce misleading effects estimates and clinical conclusions, and the correct approach is a one-step or a two-step IPD-MA that correctly accounts for clustering.

ACKNOWLEDGEMENTS

We thank those researchers who agreed to share their individual participant data from the International Mission for Prognosis and Analysis of Clinical Trials (IMPACT) project to facilitate this article. We also gratefully acknowledge the following authors for sharing of individual participant data from the deep vein thrombosis (DVT) studies: R.A. Kraaijenhagen, D.B. Toll, R. Oudega, K.G.M. Moons, D.R. Anderson, P.S. Wells, S.M. Stevens and P.S. Wells.

2

Individual participant data meta-analysis for a binary outcome: one-stage or two-stage?

PLoS ONE 2013, **8**(4): e60650.

Thomas P. A. Debray
Karel G. M. Moons
Ghada M. A. Abo-Zaid
Hendrik Koffijberg
Richard D. Riley

Abstract

A fundamental aspect of epidemiological studies concerns the estimation of factor-outcome associations to identify risk factors, prognostic factors and potential causal factors. Because reliable estimates for these associations are important, there is a growing interest in methods for combining the results from multiple studies in individual participant data meta-analyses (IPD-MA). When there is substantial heterogeneity across studies, various random-effects meta-analysis models are possible that employ a one-stage or two-stage method. These are generally thought to produce similar results, but empirical comparisons are few. We describe and compare several one- and two-stage random-effects IPD-MA methods for estimating factor-outcome associations from multiple risk-factor or predictor finding studies with a binary outcome. One-stage methods use the IPD of each study and meta-analyse using the exact binomial distribution, whereas two-stage methods reduce evidence to the aggregated level (e.g. odds ratios) and then meta-analyse assuming approximate normality. We compare the methods in an empirical dataset for unadjusted and adjusted risk-factor estimates. Results indicate that though often similar, on occasion the one stage and two-stage methods provide different parameter estimates and different conclusions. For example, the effect of erythema and its statistical significance was different for a one-stage (OR = 1.35, $p = 0.03$) and univariate two-stage (OR = 1.55, $p = 0.12$) method. Estimation issues can also arise: two-stage models suffer unstable estimates when zero cell counts occur and one-stage models do not always converge. We conclude that when planning an IPD-MA, the choice and implementation (e.g. univariate or multivariate) of a one-stage or two-stage method should be prespecified in the protocol as occasionally they lead to different conclusions about which factors are associated with outcome. Though both approaches can suffer from estimation challenges, we recommend employing the one-stage method, as it uses a more exact statistical approach and accounts for parameter correlation.

"A process cannot be understood by stopping it. Understanding must move with the flow of the process, must join it and flow with it."

– Frank Herbert, *Dune*

A fundamental aspect of epidemiological studies concerns the estimation of associations between independent variables (factors) and dependent variables (outcomes). Outcomes may include such as disease onset, disease presence (diagnosis), disease progression (prognosis), and death. Independent variables may include potential causal factors to unravel the pathophysiology or causal pathway of the outcome under study, but also non-causal predictors or risk-indicators of the outcome to enhance timely detection or prediction of the outcome, perhaps as part of a risk prediction model [97, 171, 219]. Studies that aim to explore which causal factors or predictors – often out of a number of candidate factors – are independently associated with a particular outcome have been referred to as risk factor or predictor finding studies [34, 36, 42, 108, 171, 201]. Reliable estimates of such factor-outcome associations are essential, certainly when they are meant to be causal, to properly guide public health initiatives and clinical practice for informing diagnosis and prognosis. As such, primary studies to identify causal factors or predictors are abundant in the medical literature. For example, in patients with neuroblastoma, a review identified 260 primary studies evaluating one or more novel tumour markers for their association with outcome [142, 199, 201]. When reviewing such evidence across multiple studies, the estimated factor-outcome associations across studies may be inconsistent and even contradictory [196, 225, 227]. This emphasizes the need for appropriate methods for meta-analysis and evidence synthesis in this area, in order to summarise the factor-outcome associations in the current evidence-base [3, 10, 14, 61, 109, 205, 218, 219], as commonly applied in intervention research [70, 111, 165, 179, 226]. However, due to numerous problems of published primary studies investigating factor-outcome associations, especially publication bias and selective reporting, meta-analyses based on published results are notoriously prone to bias [142, 201]. Problems with such aggregate data also arise in clinical research when differential treatment effects by patient characteristics are of concern [117]. Thus there is increasing interest in obtaining individual participant data (IPD) from these studies to facilitate a more reliable meta-analysis.

When IPD are available, meta-analysis is usually performed using a two-stage approach [226]. Each study is summarized by its factor-outcome association estimate and variance in the first stage, and these aggregate data (AD) are then appropriately combined across studies in the second stage. In this manner, a summary effect size, such as the odds or hazard ratio, is produced for each factor-outcome association of interest [69] whilst potentially accounting for between-study heterogeneity (e.g. due to different participant characteristics, methods of measurements, and undergone treatments) [2, 3, 33, 40, 95, 106, 114, 121, 196, 202]. An alternative method for IPD meta-analysis (IPD-MA) is a one-stage approach which synthesises the IPD from all studies in a single step, whilst accounting for clustering of patients within studies [68, 157, 233]. Assuming the sufficient AD are obtained from each study for the two-stage method, it is widely believed that one-stage and two-stage methods lead to similar conclusions [135, 156, 181]; however, empirical comparisons are relatively few. Indeed, because the design and implementation of one-stage and two-stage random-effects models may substantially differ, it is important to ascertain whether the choice of method can influence the final conclusions about whether a factor has a (statistically) significant association with the outcome.

In a recent empirical evaluation using a meta-analysis of 24 randomised trials of antiplatelets to prevent preeclampsia, Stewart *et al.* [233] conclude that 'two-stage and one-stage approaches to analysis produce similar results' and 'where an IPD review evaluates effectiveness based on sufficient data from randomised controlled trials, one-stage statistical analyses may not add much value to simpler two-stage approaches'. It is important to consider if this recommendation is valid in other empirical examples, and if it translates to epidemiological studies. In particular, epidemiological studies of factor-outcome associations may be affected by several covariates, namely confounders (in causal factor studies) or other predictors (in predictor finding studies) [33, 129, 279]. This situation may also arise in clinical trials when interactions occur between treatment effects and covariates, or when adjustment is needed for prognostic factors that are unbalanced between groups. Thus the random-effects framework needs to accommodate these covariates during modeling in order to estimate factor-outcome associations after adjusting for other factors. Factors that are strongly associated with the outcome might retain their association even when adjusting for other variables. However, there has again been little comparison of one-stage and two-stage IPD-MA methods when adjustment is required [78, 243].

The aim of this article is to describe and empirically evaluate possible one-stage and two-stage IPD-MA models for synthesizing (causal or predictive) factor-outcome association estimates across multiple studies where a continuous or binary factor is of interest in relation to a binary outcome. It is therefore similar in spirit to a recent description of methods for meta-analysis of time-to-event outcomes [257]. The methods are compared using an empirical example, to illustrate their advantages, differences and accessibility. Our methods all assume that between-study heterogeneity in baseline risk and factor-outcome associations exists, as it likely in practice, and so we only consider random-effects IPD-MA models. We examine different assumptions concerning the random effects, and consider how the models can be extended to adjust for other factors. Hereto, we describe two two-stage and three one-stage models for estimating unadjusted and adjusted factors. We finish by depicting some estimation procedures and approximations, and conclude with discussion and recommendations.

MOTIVATING EXAMPLE

Deep Vein Thrombosis (DVT) is a blood clot that forms in a vein in the body (usually in the lower leg or thigh). A (part of such) clot can break off and be carried through the bloodstream to the lungs and there cause a blockage (pulmonary embolism), preventing oxygenation of the blood and potentially causing death. The diagnosis DVT presence or absence can (ultimately) be made using repeated leg ultrasound, which requires patient referral and is to some extent burdening and time and money consuming. Hence, it is desirable to predict the presence or absence of DVT without having to refer patients for more cumbersome testing, by rather using easy to obtain predictors from their patient history, physical examination and simple blood assays. For this reason, in patients with a suspected DVT various studies aimed at estimating which factors – out of a range of candidate factors – are indeed associated with the presence or absence of DVT; in other words, which factors are useful diagnostic predictors of the probability that a patient truly has DVT.

A previous systematic review collected the IPD of patients with a suspected DVT from 13 studies ($n = 10\,002$), and this IPD contains information about the patients' history, physical examination and results from a biomarker test (Table 2.1, Table 2.2 and Table A.1 in the Appendix) [67, 84]. In this article, we use these data to illustrate the described meta-analysis methods for identifying important risk factors. We assume random effects for factor-outcome associations as the presence of heterogeneity between studies is expected due to differences in locale, setting and time. Detailed information about the included studies and predictors is available in the e-Appendix (Table S1 and Table S2).

METHODS

This section describes the framework for random-effects IPD-MA modeling of risk factor (predictor finding) studies with a binary outcome. Hereto, it identifies two sources of data: IPD and AD. IPD is represented by patient-level factor values (covariates) and outcomes, whereas AD consists of study-level summaries such as the estimated log odds ratios and corresponding standard errors for the factor-outcome associations reported [203]. We describe two-stage and then one-stage IPD-MA approaches [204] and describe how to account for differences in baseline risk across studies (clustering). Further, we show how to extend these methods to adjust for known risk factors, and evaluate some important estimation difficulties that arise when relatively few data are available. The DVT data is used to illustrate the methods and to identify some important differences.

Table 2.1: Overview of the DVT data

Study	N	ddimd=1	notraum=1	coag=1	eryt=1	sex=1	malign=1	par=1
1	1 028 (131)	472 (117)	743 (104)	19 (3)	382 (52)	376 (66)	54 (15)	13 (2)
2	814 (318)	598 (313)				307 (146)	86 (43)	35 (17)
3	153 (26)		103 (16)		51 (15)	73 (10)	7 (4)	12 (1)
4	1 756 (411)	910 (387)	1 497 (361)	68 (20)		654 (192)	224 (84)	101 (35)
5	791 (126)	572 (91)	650 (111)	191 (31)		301 (59)	38 (8)	112 (18)
6	1 075 (190)	424 (161)	857 (158)	52 (17)		471 (97)	55 (25)	50 (11)
7	429 (61)					153 (28)	47 (17)	12 (2)
8	325 (52)	214 (51)		57 (11)		128 (24)	12 (5)	14 (2)
9	1 295 (289)	897 (276)	1 098 (257)			467 (137)	81 (34)	178 (37)
10	436 (42)			82 (5)		145 (20)	26 (8)	13 (2)
11	541 (121)	266 (108)	373 (92)	14 (4)	144 (38)	238 (62)	99 (47)	34 (13)
12	550 (55)					210 (27)	50 (17)	12 (1)
13	809 (42)					324 (21)	55 (10)	27 (5)

Observed factor level counts (for which dvt=1) for binary risk factors in each study of the DVT case study. Entries are left blank for studies that did not measure the corresponding factor.

Table 2.2: Overview of the DVT data

Study	N	tend=1	leg=1	calfdif3=1	pit=1	vein=1	altdiagn=1	surg=1
1	1 028 (131)	562 (69)	232 (46)	311 (72)	704 (108)	155 (28)	669 (22)	81 (11)
2	814 (318)	541 (237)	169 (89)	353 (186)	419 (196)	127 (57)	217 (58)	75 (34)
3	153 (26)	82 (19)	51 (20)	59 (19)	73 (22)	25 (8)	74 (3)	24 (10)
4	1 756 (411)	664 (238)	607 (251)	426 (210)	950 (272)	283 (92)	906 (92)	198 (77)
5	791 (126)	572 (90)	353 (79)	322 (79)	490 (85)	155 (32)	300 (43)	105 (25)
6	1 075 (190)	494 (118)	217 (67)	303 (106)	357 (100)	43 (16)	448 (26)	168 (45)
7	429 (61)	203 (41)	30 (13)	96 (33)	87 (24)	33 (5)	176 (17)	25 (6)
8	325 (52)	161 (31)	47 (21)	93 (33)	97 (29)	39 (9)	114 (8)	16 (6)
9	1 295 (289)	924 (208)	583 (164)	556 (194)	799 (193)	257 (82)	782 (98)	181 (54)
10	436 (42)	222 (28)	168 (28)	66 (13)	91 (18)	1 (0)	119 (5)	58 (9)
11	541 (121)	239 (74)	152 (67)	162 (63)	270 (86)	38 (16)	313 (22)	96 (34)
12	550 (55)	176 (21)	83 (16)	114 (30)	251 (40)	28 (8)	245 (16)	39 (7)
13	809 (42)	258 (22)	75 (8)	153 (18)	196 (17)	32 (3)	399 (9)	45 (4)

Observed factor level counts (for which dvt=1) for binary risk factors in each study of the DVT case study. Entries are left blank for studies that did not measure the corresponding factor.

Two-stage IPD methods

First stage

In a two-stage method, the IPD are first analyzed separately in each study using an appropriate statistical method for binary outcome data. For example, consider where a single risk factor is of interest, then the logistic regression model is:

$$\begin{aligned} y_i &\sim \text{Bernoulli}\,(p_i) \\ \text{logit}\,(p_i) &= \alpha + \beta x_i \end{aligned} \qquad \text{(Model ♮1)}$$

with unknown parameters α (intercept) and β (slope representing the association between factor x and binary outcome y). The logit outcome probability for subject i, p_i, is then a linear function of the factor x_i. The resulting estimates from study j are denoted as $\hat{\alpha}_j$ (intercept) and $\hat{\beta}_j$ (log odds ratio). Consequently, the first step yields the intercept and the factor-outcome association estimates, and their associated within-study covariance matrix (containing the variance of the intercept var($\hat{\alpha_j}$) and each association var($\hat{\beta_j}$), as well as their respective covariances cov($\hat{\alpha}_j, \hat{\beta}_j$)) for each individual study. By utilising all the model parameter estimates, their variances and their correlation (covariance), the original IPD is reduced to AD for each study [195, 197]. If IPD are not available, such AD may alternatively be sought from study publications or study authors.In the second stage, this AD from each study are synthesized using a suitable model for meta-analysis of AD [121, 129, 130], with potential options as follows.

Second Stage

Option 1. Full (bivariate) meta-analysis AD model The AD are combined by a bivariate random-effects model that simultaneously synthesises the factor-outcome association (beta) estimates and the baseline risk (intercept) estimates whilst accounting for their correlation. The model assumes that the true underlying effect of the jth study (asymptotically) arises from a multivariate normal (MVN) distribution [271], and incorporates within- and between-study covariance. Specifically, the model fits the following marginal distributions:

$$\begin{bmatrix} \hat{\alpha}_j \\ \hat{\beta}_j \end{bmatrix} \sim \text{MVN}\left(\begin{bmatrix} \alpha \\ \beta \end{bmatrix}, \begin{bmatrix} \tau_{\alpha}^2 & \tau_{\alpha\beta} \\ \tau_{\alpha\beta} & \tau_{\beta}^2 \end{bmatrix} + \begin{bmatrix} \text{var}(\hat{\alpha_j}) & \text{cov}(\hat{\alpha}_j, \hat{\beta}_j) \\ \text{cov}(\hat{\alpha}_j, \hat{\beta}_j) & \text{var}(\hat{\beta}_j) \end{bmatrix} \right) \qquad \text{(Model 1)}$$

with unknown parameters α, β, τ_α, τ_β and $\tau_{\alpha\beta}$. Here, α and β represent the *average* baseline risk and factor-outcome association across studies, respectively, τ_α and τ_β describe their respective degree of heterogeneity between studies, and $\tau_{\alpha\beta}$ their between-study covariance.

Option 2. Traditional (univariate) meta-analysis AD model Most researchers ignore within-study and between-study covariances in parameter estimates and thus assume that $\text{cov}(\alpha_j, \beta_j)$ and $\tau_{\alpha\beta}$ equal 0 [195]. Essentially, this reduces Model 1 to a univariate meta-analysis of the factor-outcome association, and is similar to the commonly applied DerSimonian and Laird's classical random-effects meta-analysis model [70, 128], where:

$$\hat{\beta}_j \sim \mathcal{N}\left(\beta, \tau_\beta^2 + \text{var}(\hat{\beta}_j)\right) \qquad \text{(Model 2)}$$

with unknown parameters β and τ_β. This model no longer synthesises the baseline risk across studies, and just pools the factor-outcome associations.

One-stage methods

In a one-stage method, the IPD from all studies are modeled simultaneously whilst accounting for the clustering of subjects within studies. The one-stage IPD-MA framework is a (multilevel) logistic regression model with random effects. Different specifications are possible, as now described.

Option 1. Fully (bivariate) random-effects one-stage model Here, as in Model 1, random effects are specified for both the intercept and the slope, and their between-study covariance is modelled

$$\begin{aligned} y_{ij} &\sim \text{Bernoulli}\left(p_{ij}\right) \\ \text{logit}\left(p_{ij}\right) &= \alpha_j + \beta_j x_{ij} \\ \begin{bmatrix} \alpha_j \\ \beta_j \end{bmatrix} &\sim \text{MVN}\left(\begin{bmatrix} \alpha \\ \beta \end{bmatrix}, \begin{bmatrix} \tau_\alpha^2 & \tau_{\alpha\beta} \\ \tau_{\alpha\beta} & \tau_\beta^2 \end{bmatrix}\right) \end{aligned} \qquad \text{(Model 3)}$$

where i indicates observations at the individual level and j again represents the study level. Note that α_j and β_j are not explicitly estimated (in contrast to Model 1, where it represents the AD from the individual studies) but follow from the unknown parameters α, β, τ_α, τ_β and $\tau_{\alpha\beta}$. These parameters have the same interpretation as those from Model 1.

Option 2. Reduced random-effects one-stage model In a *reduced* one-stage model, independent random effects are assumed for the intercept and slope in order to avoid estimating the between-study covariance, which can often be problematic:

$$\begin{aligned} y_{ij} &\sim \text{Bernoulli}\left(p_{ij}\right) \\ \text{logit}\left(p_{ij}\right) &= \alpha_j + \beta_j x_{ij} \\ \begin{bmatrix} \alpha_j \\ \beta_j \end{bmatrix} &\sim \text{MVN}\left(\begin{bmatrix} \alpha \\ \beta \end{bmatrix}, \begin{bmatrix} \tau_{\alpha}^2 & 0 \\ 0 & \tau_{\beta}^2 \end{bmatrix} \right) \end{aligned} \tag{Model 4}$$

Option 3. Stratified one-stage model Finally, it is possible to reduce the number of assumptions by estimating a *stratified* one-stage model. This model no longer estimates an underlying average for the intercepts but rather estimates a separate intercept for each study. Thus the between-study normality assumption for the intercept term is no longer required for α_j, and there is no need to estimate a between-study covariance term. However, heterogeneity in the the factor-outcome association is still modelled using a random effect:

$$\begin{aligned} y_{ij} &\sim \text{Bernoulli}\left(p_{ij}\right) \\ \text{logit}\left(p_{ij}\right) &= \sum_{m=1}^{M} \left(\alpha_m I_{m=j}\right) + \beta_j x_{ij} \\ \beta_j &\sim \mathcal{N}\left(\beta, \tau_{\beta}^2\right) \end{aligned} \tag{Model 5}$$

where the indicator term $I_{m=j}$ indicates that a separate intercept should be estimated for each study $j = 1, \ldots, M$. Similar to Model 3 and Model 4, β_j is not explicitly estimated but follows from the unknown parameters $\alpha_1, \ldots, \alpha_M$, β and τ_{β}.

Extending the one-stage and two-stage models to examine multiple risk factors

Previously, we described models for summarizing unadjusted factor-outcome associations. Although these models are fairly straightforward to implement, it is well known that factor-outcome associations are often influenced by extraneous variables rendering exposure groups incomparable. This situation may, for instance, arise when associations are estimated from cohort and cross-sectional studies (prognostic research) or treatment-by-patient-characteristic interactions occur (intervention research). In addition, several authors have recommended that each factor should be studied for their incremental (causal or predictive) value beyond established risk factors [118, 166]. This raises the need for multivariable analyses, where the factor-outcome association under investigation is adjusted for potential confounders or other known predictors. Consequently, the methods from previous section performing a univariate (or bivariate) meta-analysis need to be extended to perform a (multivariate) meta-analysis where the factor-outcome associations (and intercept) are adjusted for K additional factors.

Extended two-stage models

For the two-stage method, multivariable logistic regression models are estimated in each study:

$$\begin{aligned} y_i &\sim \text{Bernoulli}\left(p_i\right) \\ \text{logit}\left(p_i\right) &= \alpha + \beta x_i + \sum_{k=1}^{K} \theta_k z_{ik} \end{aligned} \tag{Model $\natural$2}$$

which yields an intercept $\hat{\alpha}_j$, a risk factor-outcome association $\hat{\beta}_j$, confounder-outcome associations $\hat{\theta}_{j1}, \ldots, \hat{\theta}_{jK}$ and a within-study covariance matrix $\hat{\Sigma}_j$ for each study. A summary estimate for the regression coefficients and model intercept can be obtained by extending the bivariate random-effects model from Model 1 into a multivariate generalization [66, 129, 130, 158].

$$\begin{bmatrix} \hat{\alpha}_j \\ \hat{\beta}_j \\ \hat{\theta}_{j1} \\ \vdots \\ \hat{\theta}_{jK} \end{bmatrix} \sim \text{MVN} \left(\begin{bmatrix} \alpha \\ \beta \\ \theta_1 \\ \vdots \\ \theta_K \end{bmatrix}, \begin{bmatrix} \tau_{\alpha}^2 & \tau_{\alpha\beta} & \tau_{\alpha\theta_1} & \cdots & \tau_{\alpha\theta_K} \\ \tau_{\alpha\beta} & \tau_{\beta}^2 & \tau_{\beta\theta_1} & \cdots & \tau_{\beta\theta_K} \\ \tau_{\alpha\theta_1} & \tau_{\beta\theta_1} & \tau_{\theta_1}^2 & \cdots & \tau_{\theta_1\theta_K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \tau_{\alpha\theta_K} & \tau_{\beta\theta_K} & \tau_{\theta_1\theta_K} & \cdots & \tau_{\theta_K}^2 \end{bmatrix} + \hat{\Sigma}_j \right) \tag{Model A}$$

Usually researchers assume zero within-study and between-study correlation, and so perform a separate univariate meta-analysis to each factor-outcome and confounder-outcome association separately; that is Model 2 is fitted for each of the log odds ratio terms separately (Model B).

Extended one-stage models

The fully random-effects one-stage model with multiple risk factors is specified as follows:

$$
\begin{aligned}
y_{ij} &\sim \text{Bernoulli}\left(p_{ij}\right) \\
\text{logit}\left(p_{ij}\right) &= \alpha_j + \beta_j x_{ij} + \sum_{k=1}^{K} \left(\theta_j z_{ij}\right)_k \\
\begin{bmatrix} \alpha_j \\ \beta_j \\ \theta_{j1} \\ \vdots \\ \theta_{jK} \end{bmatrix} &\sim \text{MVN}\left(\begin{bmatrix} \alpha \\ \beta \\ \theta_1 \\ \vdots \\ \theta_K \end{bmatrix}, \begin{bmatrix} \tau_{\alpha}^2 & \tau_{\alpha\beta} & \tau_{\alpha\theta_1} & \cdots & \tau_{\alpha\theta_K} \\ \tau_{\alpha\beta} & \tau_{\beta}^2 & \tau_{\beta\theta_1} & \cdots & \tau_{\beta\theta_K} \\ \tau_{\alpha\theta_1} & \tau_{\beta\theta_1} & \tau_{\theta_1}^2 & \cdots & \tau_{\theta_1\theta_K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \tau_{\alpha\theta_K} & \tau_{\beta\theta_K} & \tau_{\theta_1\theta_K} & \cdots & \tau_{\theta_K}^2 \end{bmatrix} \right)
\end{aligned}
\qquad \text{(Model C)}
$$

Alternatively, a *reduced* one-stage model can be estimated by assuming independent random effects for $\alpha, \beta, \theta_1, \ldots, \theta_K$, i.e. the off-diagonal terms in Model C are set to 0 (Model D).

Finally, it is possible to reduce the number of random effects by stratifying the intercepts and/or predictors for which a summary estimate is not of interest. For example, one-stage *stratified* model that estimates a separate intercept for each study can be achieved as follows:

$$
\begin{aligned}
y_{ij} &\sim \text{Bernoulli}\left(p_{ij}\right) \\
\text{logit}\left(p_{ij}\right) &= \sum_{m=1}^{M} \left(\alpha_m I_{m=j}\right) + \beta_j x_{ij} + \sum_{k=1}^{K} \left(\theta_j z_{ij}\right)_k \\
\begin{bmatrix} \beta_j \\ \theta_{j1} \\ \vdots \\ \theta_{jK} \end{bmatrix} &\sim \text{MVN}\left(\begin{bmatrix} \beta \\ \theta_1 \\ \vdots \\ \theta_K \end{bmatrix}, \begin{bmatrix} \tau_{\beta}^2 & \tau_{\beta\theta_1} & \cdots & \tau_{\beta\theta_K} \\ \tau_{\beta\theta_1} & \tau_{\theta_1}^2 & \cdots & \tau_{\theta_1\theta_K} \\ \vdots & \vdots & \ddots & \vdots \\ \tau_{\beta\theta_K} & \tau_{\theta_1\theta_K} & \cdots & \tau_{\theta_K}^2 \end{bmatrix} \right)
\end{aligned}
\qquad \text{(Model E)}
$$

Stratification on all confounders may, however, not always be feasible due to sample size constraints. For this reason, we generally recommend to model separate intercept terms and to assume random effects for all predictor effects (and hence reduce model complexity by introducing additional assumptions). The underlying rationale is that accurate estimates for confounding parameters are usually not required. Although this simplification may introduce bias in all parameter estimates, baseline risks are likely most affected because they capture all unexplained variation. A non-parametric modeling approach for the intercept terms may thus better accommodate model misspecification.

Estimation procedures and approximations

In the two-stage methods, the first stage model (logistic regression in each study) is estimated using maximum likelihood (ML). In the second stage, the AD meta-analysis models are estimated using, for example, methods of moment (MOM) or restricted maximum likelihood (REML) [50, 70, 100, 128, 130]. This can be implemented in numerous software, with packages such as *lme4* and *mvmeta* in R, *Proc Mixed* in SAS and *mvmeta* in STATA. However, difficulties may arise in the first or second stage estimation. For risk factors that are binary, if zero cell counts occur in some of the included studies (e.g. when all patients with the risk factor presence also have the outcome), the likelihood function may not converge or converges in an unstable factor-outcome association. This problem is also known as (partial) separation [9, 145], and can be overcome by penalization [79, 103, 107, 141, 250] or adding a continuity correction [35, 255]. A second problem may arise when the number of included studies is small as estimation of between-study covariance may become problematic [129, 130, 198].

One-stage methods involve the estimation of a mixed effects (multilevel) model which is often high dimensional [250]. For this reason, numerical integration is often achieved through approximate methods such as adaptive Gauss-Hermite Quadrature [95, 177, 189, 191]. Although estimation becomes more precise as the number of quadrature points increases, it often gives rise to computational difficulties and convergence problems [146]. Furthermore, it has been demonstrated that the one-stage method may yield (downwardly) biased variance parameters when studies are small or limited in number [20, 41, 95, 150]. The one-stage method may also produce downwardly biased coefficient estimates when an incorrect model is specified, for instance when random effects are wrongly assumed [72]. This may increase type-II errors. Although these issues could be reduced by penalization, there is a lack of REML procedures due to the computational difficulty of the second-order Laplace approximation [150].

CASE STUDIES

In this section, we illustrate the benefits, limitations and differences of one-stage and two-stage methods in the DVT data. For all case studies, in the two-stage models we used MLE in the first stage and MLE, REML or MOM in the second stage. For the one-stage models we used adaptive Gauss-Hermite Quadrature with 1 (Laplacian approximation) and 5 quadrature points

In the first case study, we performed meta-analyses to estimate the *unadjusted* factor-outcome association for 16 risk factors using each of the models described above, and we examined the obtained log odds ratio (β), standard error (S.E.), between-study variability (τ_β) and between-study correlation ($\rho_{\alpha\beta}$). The models considered are: full bivariate two-stage meta-analysis (Model $\natural_1$ + Model 1), traditional univariate two-stage meta-analysis (Model $\natural_1$ +Model 2), fully random-effects one-stage meta-analysis (Model 3), reduced random-effects one-stage meta-analysis (Model 4) and stratified one-stage meta-analysis (Model 5). For two-stage methods, we penalized the likelihood using Jeffreys invariant prior in datasets with (partial) separation in order to stabilize study-specific estimates [79, 107].

In the second case study, we performed meta-analyses to investigate the risk factor *ddimd*, adjusted for 3 covariates (*malign*, *surg* and *calfdif3*). Hereto, we estimated the following models: extended full two-stage model (Model $\natural_2$ + Model A), extended reduced two-stage model (Model $\natural_2$ + Model B), extended full one-stage model (Model C), extended reduced one-stage model (Model D) and extended stratified one-stage model (Model E).

For all models, we calculated p-values (with $\alpha = 0.05$) and corresponding 95% confidence intervals for the estimated odds ratios, according to:

$$\hat{\beta} \pm z^\alpha \sqrt{\text{SE}(\hat{\beta})^2}$$

where z^α is the 0.975 percentile of the standardized normal distribution. Finally, we calculated 95% prediction intervals to indicate a range for the predicted odds ratio in a new study [114, 202]. Assuming the random effects are normally distributed with between-study standard deviation, then an approximate 95% prediction interval for the factor-outcome association in an unspecified study can be obtained as:

$$\hat{\beta} \pm t^\alpha_{M-2} \sqrt{\hat{\tau}^2_\beta + \text{SE}(\hat{\beta})^2}$$

where $\hat{\beta}$ is the estimate of the average factor-outcome association across studies, and t^α_{M-2} is the 0.975 percentile of the Student's t distribution with $M - 2$ degrees of freedom, where M is the number of studies in the meta-analysis.

All models were implemented in R 2.15.1 using Linux Mint 14 Nadia (MATE 64-bit) and incorporated the packages *lme4* (v0.999999-0), *mvmeta* (v0.3.4), *logistf* (v1.10) and *metamisc* (v0.0.4). Additional source code is available in the e-Appendix (supporting information files S1).

RESULTS

One-stage versus two-stage methods

Results in Table 2.3, Table 2.4 and Table S3 indicate that one- and two-stage methods often yield similar estimates for pooled factor-outcome associations, but importantly not always. For example, for the factor *par* we found an odds ratio of 1.45 (Model 1 using MLE) versus 1.32 (Model 5 using MLE). Occasionally, differences led to the one-stage and two-stage models disagreeing upon statistical significance (e.g. *eryt*). These differences mainly occurred when relatively few data were available per study (*coag* and *par*), or relatively few studies were at hand (*eryt* and *ddim*). For instance, the OR of *eryt* was 1.52 (95% CI: 0.93 to 2.47) for the univariate two-stage approach (using DerSimonian and Laird's MOM estimator), versus 1.35 (95% CI 1.03 to 1.77) for the stratified one-stage approach. Furthermore, one-stage and two-stage methods tend to provide different estimates for standard errors and between-study heterogeneity parameters, leading to different prediction intervals. For instance, the prediction interval for the odds ratio of *ddimd* ranged from 8.65 to 36.20 (Model 2 using MLE), versus 14.24 to 24.17 (Model 5 using MLE). Although usually they give similar results, the univariate two-stage method (Model 2) sometimes obtains different conclusions to the bivariate two-stage method (Model 1). For instance, for *eryt* we respectively found an odds ratio of 1.55 ($p = 0.115$) versus 1.38 ($p = 0.043$) when REML was used as estimation procedure. Finally, the bivariate two-stage method (Model 1) often gives more similar results to the one-stage method. For the factor *eryt*, we found OR = 1.37 with $p = 0.036$ using Model 1 (bivariate two-stage model), versus OR = 1.37 with $p = 0.037$ for Model 3 (bivariate one-stage model), OR = 1.39 with $p = 0.046$ for Model 4 (reduced one-stage model) and OR = 1.35 with $p = 0.029$ for Model 5 (stratified one-stage model). These estimates were all somewhat different to the results for Model 2 (univariate two-stage MoM) where OR = 1.52 with $p = 0.094$.

Estimation of correlation between random effects

As previously described, only the full one- and two-stage models (Model 1 & Model 3) estimate a parameter for the correlation between random effects. Results in Table 2.5 and 2.6 demonstrate that these models often yield correlation estimates that are close to +1 or -1, particularly when insufficient data are available and MLE is used. If correlations between random effects are assumed zero (Model 2 & Model 4), we noticed that parameter estimates may considerably change and thereby affect the calculation of p-values and prediction intervals. A good example is the unadjusted factor *coag*, where the prediction interval for the OR ranged from 0.62 to 2.47 (Model 1 with MLE) versus 0.75 to 2.23 (Model 2 with MLE), and the corresponding p-value decreased from 0.172 to

Table 2.3: Estimated unadjusted factor-outcome associations for *ddimd* in the DVT case study.

Model	Estimation	β	S.E.	τ_β	$\rho_{\alpha\beta}$	OR	95% CI	95% PI	p-value
$\natural_1$+1	MLE	2.76	0.15	0.30	0.52	15.86	11.73 to 21.45	6.98 to 36.06	< 0.001
$\natural_1$+1	REML	2.78	0.17	0.33	0.28	16.10	11.64 to 22.27	6.48 to 40.00	< 0.001
$\natural_1$+2	MLE	2.87	0.15	0.25		17.69	13.15 to 23.80	8.65 to 36.20	< 0.001
$\natural_1$+2	REML	2.89	0.17	0.31		17.97	12.88 to 25.06	7.49 to 47.04	< 0.001
$\natural_1$+2	MOM	2.89	0.17	0.32		17.98	12.87 to 25.13	7.43 to 43.54	< 0.001
3	MLE 1QP	2.87	0.15	0.28	0.07	17.70	13.14 to 23.86	8.08 to 38.78	< 0.001
3	MLE 5QP	2.85	0.14	0.25	0.58	17.35	13.15 to 22.89	8.62 to 34.91	< 0.001
4	MLE 1QP	2.88	0.15	0.29		17.79	13.15 to 24.07	8.00 to 39.56	< 0.001
4	MLE 5QP	2.85	0.19	0.41		17.37	12.06 to 25.01	5.74 to 52.55	< 0.001
5	MLE 1QP	2.92	0.11	0.00		18.55	15.01 to 22.93	14.24 to 24.17	< 0.001
5	MLE 5QP	2.92	0.11	0.00		18.46	14.94 to 22.81	14.17 to 24.04	< 0.001

Note that the IPD from 8 studies were available for estimation. Statistical significance (p-value), 95% confidence intervals (95% CI) and 95% prediction intervals (95% PI) are given for the odds ratio (OR).

Table 2.4: Estimated unadjusted factor-outcome associations for *par* in the DVT case study.

Model	Estimation	β	S.E.	τ_β	$\rho_{\alpha\beta}$	OR	95% CI	95% PI	p-value
$\natural_1$+1	MLE	0.37	0.14	0.26	-0.47	1.45	1.11 to 1.90	0.76 to 2.75	0.007
$\natural_1$+1	REML	0.38	0.14	0.29	-0.45	1.46	1.10 to 1.93	0.71 to 2.98	0.009
$\natural_1$+2	MLE	0.33	0.13	0.23		1.38	1.06 to 1.80	0.76 to 2.51	0.016
$\natural_1$+2	REML	0.33	0.14	0.27		1.39	1.05 to 1.84	0.71 to 2.73	0.020
$\natural_1$+2	MOM	0.33	0.13	0.24		1.38	1.06 to 1.80	0.76 to 2.52	0.016
3	MLE 1QP	0.32	0.13	0.23	-0.37	1.38	1.07 to 1.79	0.77 to 2.48	0.013
3	MLE 5QP								
4	MLE 1QP	0.29	0.13	0.21		1.33	1.03 to 1.71	0.78 to 2.27	0.026
4	MLE 5QP								
5	MLE 1QP	0.28	0.13	0.19		1.32	1.03 to 1.70	0.79 to 2.21	0.026
5	MLE 5QP								

Note that the IPD from 13 studies were available for estimation. Statistical significance (p-value), 95% confidence intervals (95% CI) and 95% prediction intervals (95% PI) are given for the odds ratio (OR). For some one-stage models, estimates could not be obtained because the adaptive Gauss-Hermite approximation did not converge.

Table 2.5: Estimated unadjusted factor-outcome associations for *eryt* in the DVT case study.

Model	Estimation	β	S.E.	τ_β	$\rho_{\alpha\beta}$	OR	95% CI	95% PI	p-value
$\natural_1$+1	MLE	0.32	0.15	0.10	1.00	1.37	1.02 to 1.84	0.13 to 13.97	0.036
$\natural_1$+1	REML	0.32	0.16	0.13	1.00	1.38	1.01 to 1.87	0.10 to 18.23	0.043
$\natural_1$+2	MLE	0.30	0.14	0.00		1.35	1.03 to 1.77	0.23 to 7.87	0.030
$\natural_1$+2	REML	0.44	0.28	0.39		1.55	0.90 to 2.66	0.00 to 664.30	0.115
$\natural_1$+2	MOM	0.42	0.25	0.33		1.52	0.93 to 2.47	0.01 to 303.63	0.094
3	MLE 1QP	0.31	0.15	0.10	1.00	1.37	1.02 to 1.83	0.14 to 13.02	0.037
3	MLE 5QP	0.31	0.15	0.10	1.00	1.37	1.02 to 1.83	0.14 to 13.04	0.037
4	MLE 1QP	0.33	0.17	0.14		1.39	1.01 to 1.92	0.09 to 22.31	0.046
4	MLE 5QP	0.33	0.17	0.14		1.39	1.01 to 1.93	0.09 to 22.62	0.046
5	MLE 1QP	0.30	0.14	0.00		1.35	1.03 to 1.77	0.23 to 7.80	0.029
5	MLE 5QP	0.30	0.14	0.00		1.35	1.03 to 1.77	0.23 to 7.80	0.029

Note that the IPD from 3 studies were available for estimation. Statistical significance (p-value), 95% confidence intervals (95% CI) and 95% prediction intervals (95% PI) are given for the odds ratio (OR).

Table 2.6: Estimated unadjusted factor-outcome associations for *oachst* in the DVT case study.

Model	Estimation	β	S.E.	τ_β	$\rho_{\alpha\beta}$	OR	95% CI	95% PI	p-value
$\natural_1$+1	MLE	0.10	0.18	0.20	-1.00	1.11	0.78 to 1.57	0.35 to 3.52	0.574
$\natural_1$+1	REML	0.10	0.19	0.23	-1.00	1.10	0.76 to 1.60	0.31 to 3.97	0.595
$\natural_1$+2	MLE	-0.02	0.15	0.00		0.98	0.73 to 1.31	0.52 to 1.86	0.898
$\natural_1$+2	REML	-0.02	0.15	0.00		0.98	0.73 to 1.31	0.52 to 1.86	0.898
$\natural_1$+2	MOM	-0.02	0.15	0.00		0.98	0.73 to 1.31	0.52 to 1.86	0.898
3	MLE 1QP	0.08	0.17	0.18	-1.00	1.09	0.77 to 1.53	0.37 to 3.22	0.629
3	MLE 5QP								
4	MLE 1QP	-0.03	0.15	0.00		0.98	0.73 to 1.31	0.51 to 1.85	0.866
4	MLE 5QP								
5	MLE 1QP	-0.03	0.15	0.00		0.97	0.72 to 1.30	0.51 to 1.84	0.830
5	MLE 5QP								

Note that the IPD from 4 studies were available for estimation, and that zero-cells occurred in two studies. Statistical significance (p-value), 95% confidence intervals (95% CI) and 95% prediction intervals (95% PI) are given for the odds ratio (OR). For some one-stage models, estimates could not be obtained because the adaptive Gauss-Hermite approximation did not converge.

0.078. Similar findings were obtained for the adjusted analyses (Table 2.7). Finally, results indicate that the estimated correlation between random effects tends to be less extreme when REML is used (Table 2.3). The factor *surg* is a good example, as $\rho_{\alpha\beta}$ decreased from -0.90 (MLE) to -0.65 (REML).

Estimation of stratified models

It is possible to avoid estimating correlation between random effects without assuming independence by using a stratified one-stage model, for example where a separate intercept is estimated for each study (Model 5) and, in the adjusted analyses, where predictors not of key interest are also stratified. Results indicate that the estimation of a separate intercept for each study (Model 5) tends to decrease the standard errors and between-study heterogeneity of factor-outcome associations (unless between-study correlations are +1 or -1). This, in turn, resulted in smaller prediction intervals for estimated odds ratios. For instance, the prediction interval for the unadjusted OR of *ddimd* ranged from 8.08 to 38.78 (Model 3), versus 14.24 to 24.17 (Model 5).

Estimation of one-stage models

One-stage models were estimated with 1 and 5 quadrature points, and sometimes suffered from convergence problems (e.g. *par* and *coag* in Table 2.3 where positive indefiniteness occurred when 5 quadrature points were used). Possibly, these problems are related to poor model specification. Parameter estimates were similar for 1 and 5 quadrature points in the unadjusted analyses, however, some small differences occurred in the adjusted analyses (e.g. *ddimd* in Table 2.7).

DISCUSSION

We have described several random-effects IPD-MA models that implement a one-stage or two-stage method, where one desires to evaluate a potential causal (risk) factor or predictor of outcome. We detailed how they can be estimated and also extended to adjust for other factors. Despite the conventional belief that one-stage and two-stage methods yield similar conclusions [2, 157, 233], our empirical investigation shows that this is not always the case. Specifically, we found that different estimates for pooled effects, standard errors, between-study heterogeneity and correlation between random effects can result from choosing a different method (one-stage or two-stage), choosing a different estimation procedure (MLE, REML, MOM, number of quadrature points) and choosing a different model specification (independent random effects, joint random effects, stratified

Table 2.7: Estimated adjusted factor-outcome associations in the DVT case study

Risk factor	Model	Estimation	β	S.E.(β)	τ_β	OR	p-value
ddimd (10)	$\natural_2$+A	MLE	2.62	0.18	0.40	13.67	< 0.001
	$\natural_2$+A	REML	2.64	0.20	0.44	13.80	< 0.001
	$\natural_2$+B	MLE	2.67	0.15	0.25	14.48	< 0.001
	$\natural_2$+B	REML	2.69	0.17	0.33	14.75	< 0.001
	C	MLE 1QP	2.70	0.18	0.39	14.81	< 0.001
	C	MLE 5QP	2.70	0.18	0.40	14.83	< 0.001
	D	MLE 1QP	2.67	0.16	0.33	14.42	< 0.001
	D	MLE 5QP	2.69	0.14	0.22	14.74	< 0.001
	E	MLE 1QP	2.72	0.11	0.00	15.25	< 0.001
	E	MLE 5QP	2.72	0.11	0.00	15.25	< 0.001

Factor-outcome associations are adjusted for *malign*, *surg* and *calfdif3*.

estimation). Although these differences were usually not substantial, in the DVT example they lead to discrepancies concerning the statistical significance of age, duration of symptoms, family history of thrombofilia, presence of erythema, presence of paresis and (dichotomized) D-dimer value.

Thus, importantly the choice of IPD-MA method may actually influence the conclusions about which factors are thought to be risk factors. This makes it desirable to pre-specify in a study protocol what meta-analysis method will be used, to avoid unjustified post-hoc analyses being performed to achieve statistical significance. We generally recommend that the one-stage method should be used. This method models the exact binomial distribution of the data in each study, and does not require a continuity correction when (partial) separation occurs [9, 79, 107, 145, 250]. The one-stage method may therefore produce more reliable results than the two-stage method when few studies or few subjects per study are available, as the two-stage method incorrectly assumes asymptotic normality (for the log odds ratio estimates from each study) in such scenarios [250]. The one-stage method further facilitates the adjustment for other factors, which is particularly important in non-randomised settings. In addition, one-stage models are more flexible, for example making the implementation of non-linear associations and interactions straightforward [28, 226, 233–236]. Finally, stratification in one-stage models avoids the need for estimating correlations between random effects. One can simply estimate study-specific intercepts and slopes and place the random effect only on the factor of interest.

Although we focused on IPD-MA of prognostic factors in this article, the two-stage methods can also be applied when only AD data is available for the included studies. These methods are usually preferred because sharing of IPD is often unfeasible due to, for instance, confidentiality agreements.

Results from our empirical example demonstrate that the full two-stage model, which when pooling the AD accounts for heterogeneity of baseline risk and risk factors, and their within-study and between-study correlation, tends to yield most consistent results with the one-stage models. The full two-stage method is a bivariate meta-analysis, which by additionally using the correlation between parameter estimates, is known to have benefits over a univariate me-analysis [129]. The methods presented here could further be extended using methods allowing for the combination of IPD with AD [204, 208, 254]. Potential limitations such as missing data in a subset of studies could be overcome using imputation methods that account for clustering. A Bayesian approach would be the most promising, as it would permit specification of the imputation model alongside the one-stage model, resolving several estimation limitations of the current approaches [43, 114, 253]. Furthermore, Bayesian approaches facilitate sensitivity analyses through adjusting prior specification, and permit the the robustness of fitted models to be evaluated. This is particularly useful when few studies are available and estimated parameters of one- and two-stage models may be severely biased due to estimation difficulties. Future research is needed to evaluate the performance of the described methods, and to compare their accuracy and coverage with Bayesian alternatives.

In summary, the choice of one-stage or two-stage method for performing a random-effects IPD-MA may influence the statistical identification of risk factors (predictors) for a binary outcome. When the number of studies in the meta-analysis are large and the number of events in each study are not few, we agree with Stewart *et al* [233] that a two-stage method will usually suffice. However, we generally recommend that a one-stage IPD-MA method is used as this models the exact binomial distribution, accounts for within-study parameter correlation, offers more flexibility in the model specification and avoids continuity corrections. It is therefore particularly preferable when few studies or few events in some studies are available.

ACKNOWLEDGEMENTS

We gratefully acknowledge the following authors for sharing of individual participant data from the deep vein thrombosis (DVT) studies: A.J. Ten Cate-Hoek, R. Oudega, K.G.M. Moons, R.E.G. Schutgens, D.R. Anderson, P.S. Wells, R.A. Kraaijenhagen, D.B. Toll, C. Kearon, J.L. Elf, S.M. Stevens and S.M. Bates.

3

A framework for developing, implementing and evaluating clinical prediction models in an individual participant data meta-analysis

Statistics in Medicine 2013, **32**(18): 3158–3180.

Thomas P.A. Debray
Karel G. M. Moons
Ikhlaaq Ahmed
Hendrik Koffijberg
Richard D. Riley

Abstract

The use of individual participant data (IPD) from multiple studies is an increasingly popular approach when developing a multivariable risk prediction model. Corresponding datasets, however, typically differ in important aspects, such as baseline risk. This has driven the adoption of meta-analytical approaches for appropriately dealing with heterogeneity between study populations. Although these approaches provide an averaged prediction model across all studies, little guidance exists about how to apply or validate this model to new individuals or study populations outside the derivation data. We consider several approaches to develop a multivariable logistic regression model from an IPD meta-analysis (IPD-MA) with potential between-study heterogeneity. We also propose strategies for choosing a valid model intercept for when the model is to be validated or applied to new individuals or study populations. These strategies can be implemented by the IPD-MA developers or future model validators. Finally, we show how model generalizability can be evaluated when external validation data are lacking using internal-external cross-validation, and extend our framework to count and time-to-event data. In an empirical evaluation, our results show how stratified estimation allows study-specific model intercepts which can then inform the intercept to be used when applying the model in practice, even to a population not represented by included studies. In summary, our framework allows the development (through stratified estimation), implementation in new individuals (through focused intercept choice) and evaluation (through internal-external validation) of a single, integrated prediction model from an IPD-MA in order to achieve improved model performance and generalizability.

Chapter 3

"The shoe that fits one person pinches another; there is no recipe for living that suits all cases."

– Carl Gustav Jung

Clinical prediction models are an increasingly important tool in evidence-based medical decision making [5, 170, 192]. They aim to accurately predict an individual's risk of disease being present (diagnostic prediction model) or occurring in the future (prognostic prediction model), to thereby inform clinical and therapeutic decisions, facilitate healthcare and public health policies, and aid patient counseling [56, 182, 192, 222, 289]. An example is the diagnostic model develop by Oudega *et al.*[182], which aims to predict the presence of deep-vein thrombosis in patients suspected of deep-vein thrombosis at primary care. Such prediction models are typically derived from a single dataset including individual participant data (IPD), in which the association between the presence or occurrence of the outcome of interest and a set of predictors (covariates) is estimated [5, 213, 237]. During the past decades, prediction research has become more popular and international collaboration has become more commonplace. This has led to an increased sharing of IPD and subsequently exposed the need for IPD meta-analysis (IPD-MA) to appropriately synthesize these data to develop (and validate) a single prediction model [203, 207]. Examples of IPD meta-analyses that have led (or will lead) to the development and validation of risk prediction models are abound in the literature [188, 215, 222, 248].

Prediction models resulting from IPD-MA are appealing as they may be seen to be more generalizable as compared to using IPD from just a single study population; the inclusion of multiple studies addresses a wider range of study populations and increases the variation in the characteristics of the included participants. However, by simply combining IPD to produce a prediction model averaged across all study populations, researchers might actually obfuscate the extent to which the individual studies were comparable, and can mask how the model performs in each study population separately. For example, when study differences in model parameter estimates cannot be explained by sampling variability solely, i.e. heterogeneity is present, resulting models may not generalize well and perform poorly when applied in new individuals. One of the key expressions of this heterogeneity is differences in the baseline risks, i.e. outcome prevalences (for diagnostic models) or incidences (for prognostic models), or in the predictor-outcome associations [124, 125, 169]. Potential causes of such heterogeneity in otherwise related study populations are differences in study design, inclusion and exclusion criteria, disease severity and interventions undergone [96, 279].

When an IPD-MA aimed at developing 'an average' prediction model does not appropriately handle potential heterogeneity, resulting prediction models may yield systematically biased predictions when validated or applied in new individuals or study populations. This, in turn, renders their clinical usefulness obsolete [133, 169]. Consequently, the implementation of random effects modeling that effectively account for heterogeneity across the included studies seems highly recommended [93, 202, 203]. This approach, however, also complicates external validation and implementation of the resulting prediction model, as parameters (such as intercept and predictor-outcome associations) are allowed to take different values for each included study [8, 66]. This then raises the question about which parameters should be used when the prediction model is validated or applied in new individuals or study populations that were not considered during its derivation; researchers hardly address this difficulty. Furthermore, an IPD-MA may not always improve the generalizability of clinical prediction models, as it is possible that study populations differ too much to usefully combine them; focusing on an average model across all study populations is thus misleading [202]. A framework is therefore needed that supports both the identification of the extent to which aggregation of IPD is justifiable and the optimal approach to achieve this aggregation. In addition, this framework should guide subsequent researchers and potential users how to validate or apply the model to new individuals.

Royston *et al.* proposed a framework to construct and validate a prognostic survival model from an IPD-MA [215]. This framework adopts an "internal-external cross-validation" (IECV) approach to evaluate whether derived models have good prognostic separation in independent studies, and whether the baseline survival distribution is heterogeneous across studies. Afterwards, a single final model is derived from all available IPD using flexible parametric proportional hazards modeling techniques. Although this framework appears to be a useful strategy for accounting and adjusting for heterogeneity in an IPD meta-analysis aimed at developing a single, average prediction model, it has not yet been widely implemented. In addition, the suggested framework pools the baseline hazard distribution functions, which may not be justified when heterogeneity is largely present. Finally, it remains unclear how the framework should be applied when models aim to predict binary outcomes, using multivariable logistic regression, rather than time to event.

Here, we propose several approaches to account and adjust for heterogeneity in an IPD-MA that aims to develop a novel prediction model for a binary outcome, and allow it to be externally validated or applied in new individuals. We begin by considering a range of strategies for developing a model when the included studies may have different outcome frequencies (baseline risks), that potentially require different intercepts in the model. We then describe how to apply the fitted model to a new study population by obtaining an appropriate intercept for this new study population, even when its baseline risk is unknown. In this manner, we aim to facilitate its implementation or external validation when baseline risks are heterogeneous across studies. We demonstrate that only limited information about the new study population is sufficient to adjust the derived prediction model and facilitate reliable predictions [133, 239, 270]. Furthermore, we extend the "internal-external cross-validation" approach proposed by Royston *et al.* [215] to evaluate the generalizability of derived prediction models in other study populations. This approach can also be used to identify which combination of studies yield consistent prediction models, and which studies may present problematic sources of evidence and may need to be excluded for the model development. Finally, we extend the framework to count and time-to-event data prediction models, and illustrate the approaches using a diagnostic modeling IPD-MA on the prediction of the presence of Deep Vein Thrombosis.

METHODS

This section describes our framework to develop a prediction model from an IPD-MA with a binary outcome, and optimally adjust its intercept to a new study population. This framework is summarized in Figure 3.1, and we now explain each step in detail. We begin by assuming that the included studies have similar predictor-outcome associations, but may have a heterogeneous outcome frequency or baseline risk. Consequently, three important steps can be distinguished: (1) estimation of predictor-outcome associations from the available studies whilst accounting for heterogeneity in baseline risks, (2) estimation of an appropriate model intercept when the model is to be implemented or validated in a new study population outside the IPD meta-analysis, and (3) evaluating the generalizability of the resulting model. This last step iteratively assesses the extent to which estimations of the predictor-outcome associations and model intercept from a subset of the available studies yield accurate model predictions in the remaining IPD.

Finally, the value of the framework in the presence of additional heterogeneity in the predictor-outcome associations is considered in Case Study 2.

Step 1: Estimation of predictor-outcome associations

This first step estimates the predictor-outcome associations across the available IPDs in the IPD-MA dataset, and considers several approaches to account for differences in baseline risk. For sake of simplicity, we assume that a pre-selection (based on e.g. prior knowledge or clinical expertise) of the candidate predictors has been done and that their specification (e.g. linear or non-linear forms in case of continuous predictors) in the model is predefined. We refer the reader to other sources that discuss the selection and specification of predictor variables [101, 237], and note that it is possible to evaluate different choices of model specification by assessing its performance in a validation sample. We consider the situation in which IPD from $j = 1, \ldots, M$ studies are available. The data from each study is described by K independent predictors, a dichotomous outcome $\boldsymbol{y}$, and contains N_j subjects. Let $\boldsymbol{X}_{ij}$ denote a $1 \times K$ vector with the predictors for subject $i = 1, \ldots, N_j$ in study j. Three possible logistic regression modeling approaches in this situation are stacking, random intercept effects, and stratification.

Stacking

A first, potentially naive approach may assume that all IPD were collected from a single and homogeneous population. This approach ignores the clustering of participants within different studies, and merges all their data into one dataset by means of stacking:

$$\begin{aligned} y_i &\sim \text{Bernoulli}\left(\pi_i\right) \\ \text{logit}\left(\pi_i\right) &= \alpha + \boldsymbol{\beta'} \boldsymbol{X}_i \end{aligned} \tag{3.1}$$

The common intercept α and predictor-outcome associations $\boldsymbol{\beta}$ (representing a $1 \times K$ vector) for all studies shows that clustering is being ignored. This type of meta-analysis is hard to justify when study populations have different outcome incidence or prevalence, as then the baseline risk is different for each study. It is known that ignoring such heterogeneity in baseline risk can induce bias in predictor-outcome associations [2].

Random effects modeling of the intercept

If heterogeneity in an IPD-MA only occurs in the baseline risk, it is possible to account for these differences using a random-effects logistic regression model. This approach estimates a weighted average model intercept by assuming random effects for the model intercepts across the included studies in the IPD meta-analysis [32, 230, 262]. To this purpose, it allows a separate intercept for each study and estimates the distribution of this intercept across studies. Here, we assume a normal distribution which leads to an estimated mean (i.e. the average study intercept), α, and variance (i.e. the between-study heterogeneity in intercept), τ_α^2. The corrseponding logistic regression model consists of $K + 2$ parameters, and is specified as follows:

$$\begin{aligned} y_{ij} &\sim \text{Bernoulli}\left(\pi_{ij}\right) \\ \text{logit}\left(\pi_{ij}\right) &= a_j + \boldsymbol{\beta'} \boldsymbol{X}_{ij} \\ a_j &\sim \mathcal{N}\left(\alpha, \tau_\alpha^2\right) \end{aligned} \tag{3.2}$$

By assuming random effects it becomes possible to model heterogeneity in baseline risk with relatively few parameters. Unfortuntely, it is often difficult to evaluate whether the corresponding assumptions are justifiable, particularly when a small number of studies are available in the IPD-MA. Although it is possible to relax the required assumptions by adopting a Bayesian perspective using vague priors, such strategy requires advanced statistical expertise and specialised software packages which may not always be available [144].

Stratified estimation of the intercept

Given these aforementioned limitations, it may sometimes be inappropriate to estimate an average intercept across all studies. For this reason, we propose estimating a *stratified* intercept for each study when relatively few IPD studies are at hand. This implies that a separate intercept α_j is estimated for each study, and an underlying distribution of random intercept effects is no longer assumed.

$$\begin{aligned} y_{ij} &\sim \text{Bernoulli}\left(\pi_{ij}\right) \\ \text{logit}\left(\pi_{ij}\right) &= \sum_{m=1}^{M} \left(\alpha_m I_{m=j}\right) + \boldsymbol{\beta' X}_{ij} \end{aligned} \tag{3.3}$$

where I represents an indicator variable that equals 1 when $m = j$ and 0 otherwise. By using an indicator variable to estimate a separate intercept for each study, the normality assumption from expression 3.2 is avoided, and an overall estimate for the model intercept as in the random effects approach is no longer estimated. Unfortunately, this also implies that the resulting model focuses on the studies at hand, and the choice of intercept when validating or applying the final model to new individuals (outside the IPD-MA) is not immediately obvious. How to deal with this is further addressed in step 2. It should further be noted that stratification may result into estimation difficulties when some studies have few or no events, and now involves $M + K$ instead of $K + 2$ (random effects modelling) or $K + 1$ (stacking) unknown parameters. For this reason, stratification may not be feasible when many studies with relatively few participants are at hand.

Step 2: Choosing an appropriate model intercept when implementing the model to new individuals

Although all methods in step 1 yield a unique choice of predictor-outcome associations, the presence of heterogeneity in baseline risk across the study populations of the IPD-MA may induce a set of different model intercepts. For example, the $\boldsymbol{\beta}$ estimates from the prediction model in expression 3.3 need to be combined with an intercept that is appropriate for the study population in which one wants to validate or apply the IPD model. It may be clear that the presence of heterogeneity in baseline risk complicates the implementation of a prediction model in individuals outside the IPD-MA. Although model developers could report a unique summary intercept in such scenarios (see *Average Intercept*), an alternative strategy is to allow future model implementors or validators to obtain an intercept that is optimal for their specific study population. In this section, we describe three methods for obtaining such an intercept with minimal information about the new individuals or study population. Two of these strategies solely require baseline descriptives about the study population (see *Intercept Selection* and *Intercept estimation from outcome prevalences*), whereas the third method ensures intercept optimality by re-estimating the intercept using IPD (see *Intercept estimation from new IPD*). All methods can be implemented without the original participant data, as long as some basic information about these data is reported. Summarized, this second step aims to facilitate future validations and applications of the final IPD model by presenting several strategies for obtaining a unique model intercept when baseline risks are heterogeneous across the included study populations.

Average Intercept

A straightforward approach for obtaining an appropriate model intercept may use the estimated (weighted) average from the IPD-MA, as captured by α in the stacking or random effects approaches described above. This approach was proposed by Royston *et al.*, who pool the baseline hazard distribution functions of the studies in an IPD-MA for deriving a prognostic model [215]. Although α is unavailable in the stratified approach, an estimate can be obtained by pooling the individual intercepts a_j estimates using a fixed or random effects meta-analysis as necessary. A major advantage of an average intercept based on all included studies is that it can directly be used as approximation of baseline risk in a new study population with unknown outcome incidence or prevalence. Unfortunately, this uncertainty about the new study population implies that the resulting average estimate may be very different to the actual intercept in a single population, especially when outcome incidences or prevalences do differ across patient populations, which is often the case in practice. This error in the intercept may then lead to poor predictive accuracy when the model is applied.

Intercept Selection

To avoid using an average model intercept, an alternative approach is to simply select an estimated intercept from one of the IPD studies that is most similar to the new study population. This intercept can be directly obtained from a_j (random effects approach) or α_j (stratified approach). Although we believe that this comparison should be guided by clinical expertise, it is possible to rely on a purely statistical approach. This approach could, for instance, evaluate similarity by comparing the outcome frequency of each derivation IPD with the new population where the model is to be applied. This approach is taken by Steyerberg *et al.* who develop a risk prediction model across multiple studies, and then when validating the model, they use the intercept taken from just one of the included studies, as this study had an outcome prevalence most similar to that found in clinical settings [248]. Alternatively, one could identify the closest matching IPD study by evaluating differences in baseline characteristics between the new study population and the IPD studies by comparing observed means (e.g. mean age) and proportions (e.g. % male) for each included study. Evidently, these strategies require the IPD-MA developers to report the estimated intercepts of each study population, as well as their corresponding outcome frequency or baseline characteristics. Although information about a population's outcome frequency or baseline characteristics is typically available when the model is to be externally validated in that population (as this process typically entails the collection of IPD), it may be missing when a model is to be implemented in a new population. In these scenarios, researchers could revert back to using the weighted average intercept from the random effects or stratified model [215], given that these estimates are reported.

Intercept estimation from outcome prevalences

It is also possible to calculate an estimate of the model intercept for a particular population using the outcome incidence or prevalence (proportion of patients developing the outcome) prev_{new} when known in that population. Estimates of these proportions may be obtainable from (a systematic review of) the medical literature or experts in the field, and can be translated into a model intercept by applying the logit transformation:

$$\hat{\alpha} = \ln\left(\frac{\text{prev}_{\text{new}}}{1 - \text{prev}_{\text{new}}}\right) \tag{3.4}$$

However, implementation of the resulting $\hat{\alpha}$ as the intercept when applying the prediction model is only justified when the included variables in the prediction are mean-centered for each included study, where the mean of dichotomous predictors (e.g. sex: male=1, female=0) corresponds to their prevalence (e.g. the proportion who are male). The underlying reason is that prev_{new} represents the predicted outcome risk of a random individual in the new population. If the variables in each study IPD are not mean-centered, the intercept term of their linear predictor represents a specific subgroup of individuals (such as gender=female or age=0). Mean-centering of predictor variables ensures that $\boldsymbol{\beta}'\boldsymbol{X} = 0$ on average, and thus that α represents the outcome logit risk for a random individual in the population. Although this particular individual may not exist (as individuals cannot have a mean gender between 0 and 1), it reflects the average study participant and therefore remains representative on the population level from which prev_{new} is derived. Note that because a mean-centered prediction model has population-specific predictor means in the linear predictor, it can only be implemented in a new study population when the mean predictor values are also available for that population. That is, in the new population one needs to apply the prediction model as specified by:

$$\pi_i \quad = \quad \text{logit}^{-1}\left(\hat{\alpha} + \hat{\boldsymbol{\beta}}'\left(\boldsymbol{X}_i - \overline{\boldsymbol{X}}\right)\right) \tag{3.5}$$

where the beta estimates are taken from the developed prediction model in step 1 (e.g. stratified or random-effects above) and the alpha term is from equation 3.4.

Intercept estimation from new IPD

Finally, at the time of wishing to apply the prediction model to a new study population, IPD may additionally be available from this population of interest, and these data may serve for updating or re-estimating the model intercept using methods previously described [133, 169, 237, 259]. This can generally be achieved by setting the linear predictor $\hat{\boldsymbol{\beta}}'\boldsymbol{X}$ as offset and re-estimating the corresponding intercept. For the centered approach, the mean predictor values can directly be obtained from this new IPD, and the corresponding offset is given as $\hat{\boldsymbol{\beta}}'(\boldsymbol{X} - \overline{\boldsymbol{X}})$, where $\hat{\boldsymbol{\beta}}$ is taken from the developed prediction model in step 1.

Step 3: Model evaluation using internal-external cross validation

In the previous sections we described the first two steps necessary for estimating and implementing a prediction model so that it can be considered for external validation and application in routine care. Although external validation has been proposed as the ultimate solution for evaluating a model's generalizability, corresponding IPDs are often lacking and their collection typically requires a lot of effort. Consequently, some form of internal validation seems desirable to guarantee that the derived model is accurate enough to be clinically useful. Specifically, the strategies for obtaining accurate predictor-outcome associations (step 1) and an appropriate model intercept (step 2) should lead to consistent and discriminative model predictions. Because it is possible that the IPD-MA model developers cannot present a unique model intercept due to heterogeneity in baseline risk, it would also be useful for them to investigate whether future model implementors or validators can obtain an accurate model intercept from the available evidence. Consequently, this third step is an extended form of internal model validation to evaluate its performance and generalizability when external validation data are lacking [15, 16, 169, 170, 249]. One option is to develop the model in steps 1 and 2 using just a subset of IPD studies, and keep others aside for validation. However, we consider it is important to both maximize the data available for the model development and also the model validation. In this section, we thus adapt the "internal-external cross-validation" (IECV) technique originally proposed by Royston *et al.* [215]. This technique iteratively uses $M-1$ studies from the available IPD-MA to develop a prediction model and the remaining study for its validation. In this manner, M scenarios are available to investigate consistent model performance when applied in another study population that was not included during its development. We propose the following stages in the IECV technique:

1. Select the IPD of $M-1$ studies from the meta-analysis. These data will serve as derivation data, whereas the IPD of the remaining study will serve as validation data (i.e. sample where the model is to be implemented and externally validated)
2. Estimate the predictor-outcome associations in the derivation data using one of the approaches described in step 1.
3. Choose a model intercept that is appropriate for the validation sample, using one of the approaches described in step 2. Here, the validation data may be used to borrow (limited) information about the new study population, such as the outcome prevalence or predictor mean values.
4. Combine the estimated predictor-outcome associations (from 2) and chosen model intercept (from 3) into a single model, and apply this model in the validation data.
5. Use the validation study to evaluate the performance of the derived prediction model (from 4).
6. Repeat 1–5 for each permutation of $M-1$ derivation studies.

We focus on statistical criteria to assess model performance in the validation sample, and explicitly distinguish between discrimination and calibration [58, 249]. Whereas the former reflects the ability to distinguish high-risk subjects from low-risk subjects, the latter indicates the extent to which the predicted outcome probabilities and actual probabilities agree.

An overall indication of model calibration is reflected by the ratio of predicted (expected) to observed outcomes, denoted by E/O. This ratio should ideally be 1, and deviations above (or below) this value indicate that the model intercept is too high (or too low). We also measure the calibration slope in the validation sample, b_{overall}, to evaluate whether the average strength of the predictor-outcome associations is similar in these data [60, 160, 237]. A poor calibration slope ($b_{\text{overall}} \neq 1$) usually reflects overfitting of the model in the derivation sample, but may also indicate heterogeneity of predictor-outcome associations between the derivation and validation sample. However, because the calibration slope is an overall measure of fit, it may not reveal all potential pitfalls. For this reason, it may be more useful to directly compare estimated predictor-outcome associations in the derivation and validation sample. Visual inspection of the calibration plot may further reveal how the quality of predicted risks is affected [229, 237]. This plot indicates how predicted risks diverge from observed outcomes in different deciles of predicted risks, and shows perfect predictions when the calibration curve goes through the origin and has a slope of 45°.

Finally, we assess to what extent the model is able to distinguish between patients with the outcome and patients without the outcome by means of the area under the ROC curve (AUC), also known as the C statistic [102]. This score ranges from 0.5 (no discrimination) to 1.0 (perfect discrimination). Additional insight into discrimination can be achieved through Hedges' g statistic or the overlap coefficient [211].

Results from the IECV technique can be interpreted as follows. In general, if the derived models (that is, the M models produced by omitting each IPD study in turn) all validate well across the considered permutations, all datasets can then be combined and used to develop the final prediction model. If some of the derived models do not calibrate well in the validation sample, the IECV indicates that generalizability of any model across all M studies is not guaranteed. In those scenarios, to identify the cause of the problem it is useful to examine the consistency of estimated predictor-outcome associations and model intercepts (or visually inspect the calibration plot) across the M studies, as follows.

If the E/O ratio considerably differs from 1 or calibration curves do not coincide with the reference line in many validation samples, this may suggest that the strategy chosen for obtaining a model intercept in the new study population (step 2) does not perform well. It may then be preferable to collect IPD from the new study population in order to obtain a more study-specific model intercept. Conversely, when predictor-outcome associations substantially differ between the derivation and validation sample, approaches to overcome heterogeneity in baseline risk no longer perform well and the model's generalizability may suffer. This is because the model intercept encapsulates all sources of unexplained risk, and not only difference in the incidence of the outcome. It may therefore be affected in unpredictable ways when baseline risk or predictor-outcome associations are heterogeneous. This, in turn, implies that derivation of prediction models from an IPD meta-analysis may not be feasible when predictor-outcome associations are known to be heterogeneous. This pitfall is also reflected by calibration curves that are not straight or have a slope different from 45°, and could be further examined by measuring or testing the amount of heterogeneity [70, 113, 217]. Although the inclusion of additional covariates, non-linear associations or interaction terms may reduce heterogeneity, such an approach inevitably increases the risk of overfitting. Where heterogeneity in predictor effects cannot be reduced and the IECV approach shows poor model performance and generalizability, it should signal to the researcher that a single prediction model that applies to all study populations is unlikely to be possible using the predictors available. In those scenarios, other predictor variables should be considered, or some studies could be excluded and the model built on a more homogeneous set. Then researchers need to clearly report which studies (populations) were excluded, and note that the developed model is unlikely to generalize to them.

Finally, evaluation of the AUC may further help to identify whether accurate predictions also lead to good discrimination. After all, accurate predictions may not be very useful if they are similar regardless of the developed outcome. This is particularly the case for diagnostic models, where the ultimate goal is to accurately classify subjects into their true disease states [58]. Although the AUC should ideally be 1, there are no specific guidelines about acceptable performance thresholds as these differ according to the considered prediction task.

It should be noted that results from the IECV are only useful if sufficient data are available on individual participant and study level. Specifically, if some studies in the IPD meta-analysis contain very few patients, performance statistics may become unreliable and corresponding confidence intervals may substantially inflate. Although there are some guidelines for sample size requirements in external validation studies (Vergouwe et al proposed a rule of thumb to use a minimum of 100 events and 100 nonevents), there is no clear threshold for which reliable performance statistics can be achieved [238, 276]. Similarly, if few studies are available in the IPD meta-analysis, little insight into model generalizability can be gained by applying the IECV technique, and identification of variation in baseline risk becomes difficult. For this reason, we recommend the inclusion of at least 4 or 5 studies in the development of a meta-analytical prediction model that have a reasonably large sample size and number of events.

Furthermore, it is important to realize that implementing this proposed framework requires careful planning and consideration beforehand. Assessing the performance and heterogeneity measures obtained from this process is subjective and requires in-depth knowledge of the clinical research problem. Devoid of this context, the statistical measures we present here have no direct relation to the impact a model is likely to have in routine care. For this reason, we recommend that desired performance characteristics are predefined (e.g. what minimal AUC is required? Is there a particular range of predicted probabilities for which good calibration is required?) [88, 192, 249] and evaluated alongside the consequences that would result from implementing the model in routine care [275, 277]. Furthermore, the research question needs careful thought and reporting in terms of which primary studies need to be included in the meta-analysis. In this regard, potential sources of heterogeneity should be investigated by using knowledge in the subject area or performing descriptive analyses on the key characteristics of the available studies. Then, researchers may decide which factors could contribute to heterogeneity, and whether aggregation remains justified or if study exclusion is necessary. Finally, characteristics of included and excluded studies should be adequately reported such that the final model can successfully be implemented and validated in routine care.

In summary, when developing a risk prediction model using IPD from multiple studies with binary outcomes, researchers have three main options for model development (stacked, random-effects on intercept, and stratified intercept) and must decide how to designate an intercept value when the model is applied to new individuals. The IECV is a framework for evaluating this entire strategy and the performance and generalizability of the model it produces. Evidence that the model does not generalize (validate) consistently across all M studies signals that researchers should re-evaluate their strategy, and aim to reduce any heterogeneity in predictor-outcome associations and improve the reliability of their chosen intercept.

EXTENSION TO COUNT AND TIME-TO-EVENT DATA

Although we described how our framework can be implemented for a prediction model using binary outcome data, it is fairly straightforward to extend this framework to other outcome data types. For instance, count data can be modelled using a Poisson model, where expression 3.1 becomes:

$$\begin{aligned} y_i &\sim \text{Poisson}\left(\lambda_i\right) \\ \ln\left(\lambda_i\right) &= \alpha + \boldsymbol{\beta}'\boldsymbol{X}_i \end{aligned} \tag{3.6}$$

In this expression, α represents the log of the baseline rate, and can be modelled using random effects or stratified estimation similar to expression 3.2 and 3.3. This model can further be extended to estimate proportional hazards (PH) models when time-to-event data are available such that each patient can have a different length of follow-up [38, 48, 61, 148]:

$$\begin{aligned} y_i &\sim \text{Poisson}\left(\lambda_i\right) \\ \ln\left(\lambda_i\right) &= \ln(t_i) + \alpha + \boldsymbol{\beta'} \boldsymbol{X}_i \end{aligned} \tag{3.7}$$

where $\ln(t_i)$ is a standardizing offset term for subject i with exposure time t_i. Note that this model assumes that the baseline hazard (i.e. the hazard when all covariates are zero) is a constant over the whole time period. Although it is possible to relax this assumption by adopting a Cox PH model (which still assumes proportional hazards at all times) [215], there are several limitations to this approach. Most importantly, Cox PH models have an unspecified baseline hazard which hampers prediction of survival times [210, 212, 270]. For this reason, PH models that make specific assumptions about the baseline hazard distribution are sometimes preferred. The conditional hazard function of PH models can be generalized as follows [26]:

$$h(t|\boldsymbol{X}_i) = g\left(\boldsymbol{a}, t\right) e^{\boldsymbol{\beta'} \boldsymbol{X}_i} \tag{3.8}$$

where $g(\cdot\,)$ is a function known up to a multidimensional parameter $\boldsymbol{a}$. The exponential distribution is a common example and assumes a constant hazard over time, i.e. $g(\cdot\,) = \lambda$. Here, a random baseline hazard effect can be modelled as follows:

$$\begin{aligned} h(t|\boldsymbol{X}_{ij}) &= \zeta_j \lambda e^{\boldsymbol{\beta'} \boldsymbol{X}_{ij}} \\ \zeta_j &\sim \Gamma\left(1, \theta_0\right) \end{aligned} \tag{3.9}$$

This expression is similar to the Gamma frailty model [89], where the ζ_j are study effects distributed as independent and identically distributed gamma random variables with mean 1 and variance θ_0. The variance parameter is interpretable as a measure of the heterogeneity across studies in baseline risk. When θ_0 is small, then values of ζ are closely concentrated around 1 and the study effects are small. If θ_0 is large, then values of ζ are more dispersed, inducing greater heterogeneity in the study specific baseline hazards $\zeta_j \lambda$. The study-specific baseline hazards are all proportional to λ. Other, more advanced distributions are the Weibull distribution, where $g(\cdot\,) = \lambda \gamma t^{\gamma - 1}$, or the Gompertz distribution, where $g(\cdot) = \lambda e^{\alpha t}$. Heterogeneity in baseline hazards could be introduced here in a similar manner by adding a study effect ζ_j, or by estimating a stratified baseline hazard $g(\cdot)$ for each study. An appropriate baseline hazard could then be selected from existing studies in the meta-analysis using the incidence in the new study population. Note that the baseline hazard could also be modelled using restricted cubic splines within a flexible parametric framework [214, 215]. Finally, it is important to acknowledge that estimation issues may further be complicated if the studies in the IPD-MA are subject to different censoring mechanisms.

CASE STUDIES

To demonstrate the potential value of aforementioned approaches for model development, intercept choice, and IECV, we now consider three scenarios that use the IPD of 12 studies conducted for diagnosing Deep Vein Thrombosis (DVT) in patients with a suspected DVT. The scenarios differ in the predictor variables they consider. In the first example, the modeled predictor-outcome associations are homogeneous across all studies, in the second they are strongly heterogeneous, and in the third they are weakly heterogeneous. In all scenarios the baseline risk is heterogeneous across the 12 included studies of the IPD meta-analysis. The studies are summarized in Table 3.1, and contained a total of 10 014 patients of which 1 897 (18.9%) truly have DVT. The corresponding IPD were collected between 1994 and 2007 in the United States of America, Sweden, Canada and the Netherlands (Table A.1 in the Appendix).

In each scenario we apply the three aforementioned steps. In step 1, we consider the stacking, random-effects, and stratified approaches for estimation of the predictor-outcome associations. Then, in step 2, for choosing the intercept for use in a new population following the stacking and random effects approach, the estimated average intercept α was used as final choice (see *Average Intercept*). For the stratified approach, three different strategies were evaluated: intercept *selection* based on the outcome proportion in the new study population, intercept *selection* based on similarities of baseline descriptives, and intercept *estimation* based on the outcome proportion observed in the IPD of the new study population. Finally, in step 3, we used the IECV approach for assessing the extent to which the described approaches yield generalizable prediction models. We evaluated whether model performance remained consistent in each validation study by measuring the statistics proposed in step 3 (proportion of predicted and observed outcomes, average percentage bias of the predictor-outcome associations and the area under the ROC curve) and visually inspecting the calibration plots.

All analyses were performed on a Linux system (kernel 3.2.0) with R version 2.15.2 (R Foundation for Statistical Computing, Vienna, Austria) using the *lme4* library. The corresponding source code

Figure 3.1: Recommended steps for developing, implementing, and evaluating a risk prediction model in an IPD meta-analysis

1. **Development**
   Use stratified estimation model and check for heterogeneity in baseline risk and predictor-outcome associations using multivariate meta-analysis and estimating τ_β values. Of those variables with predictive importance, consider prioritizing those with homogeneous associations across studies.
2. **Implementation in new population**
   (a) If IPD are available from the new population, use these data to estimate the model intercept to be used alongside the predictor-outcome associations from the developed model; or
   (b) If the outcome prevalence and mean predictor values from the new population are known, calculate intercept for this population using eq. 3.4; or
   (c) If the outcome prevalence from the new population is known, then identify a study with a similar outcome prevalence in the meta-analysis, and use the estimated intercept from that study; or
   (d) If no information from the new population is available, use the average intercept from the stratified or random effects model
3. **Evaluation of the model**
   (a) Use the IECV approach to evaluate the performance of the developed model in the remaining validation study for each permutation of $M - 1$ derivation studies.
   (b) Calculate E/O statistic, calibration slope and area under the ROC curve. Ideally all these values should be close to 1. Consider producing calibration plot for each study to examine consistency of E/O values across the range of predicted probabilities.
4. **Completion or Updating**
   (a) If the performance is consistently good in all studies from the IECV, produce a final fitted model using the IPD from all studies and report the strategy researchers should take for choosing an appropriate model intercept when applying the model; or
   (b) If the performance is not consistently good in all studies from the IECV and there is heterogeneity in baseline risk and/or predictor effects, consider a different strategy for modeling the intercept and try to reduce heterogeneity in predictor-outcome associations. Alternatively, identify subset of studies where the model performance is consistently good, but summarize those populations for which the model performs poorly.

Table 3.1: Baseline descriptives of the IPD in the DVT case study

ID	N	Time period	sex=1	surg=1	malign=1	calfdif3=1	ddimdich=1	dvt=1
1	1 028	2005 - 2007	37%	15%	5%	30%	46%	13%
2	814		38%	9%	11%	43%	73%	39%
3	153	1994 - 1996	48%	16%	5%	39%	24%	17%
4	1 756	1995 - 1999	37%	16%	13%	24%	52%	23%
5	532	2003 - 2005	40%	19%	4%	41%	75%	17%
6	1 075	1997 - 1999	44%	4%	5%	28%	39%	18%
7	1 768	1994 - 2001	38%	6%	8%	20%	NA	8%
8	357	2003 - 2005	39%	5%	4%	19%	69%	24%
9	1 295	2002 - 2003	36%	20%	6%	43%	69%	22%
10	436	2000 - 2001	33%	13%	6%	15%	NA	14%
11	541		44%	7%	18%	30%	49%	22%
12	259	2002 - 2006	33%	21%	7%	41%	68%	14%

The predictor variables are given as *sex* (1 = male, 0 = female), *surg* (1 = recent surgery or bedridden, 0 = no recent surgery or bedridden), *malign* (1 = active malignancy, 0 = no active malignancy), *calfdif3* (1 = calf difference ≥ 3 cm, 0 = calf difference < 3 cm), *ddimdich* (1 = D-dimer positive, 0 = D-dimer negative) and *dvt* (1 = DVT, 0 = no DVT). Predictor variables that were not measured in a particular study are indicated by NA.

is available on request.

Case Study 1: Homogeneous predictor-outcome associations

In this first scenario, we derive a prediction model by only including predictor-outcome associations that are (nearly) homogeneous in the IPD meta-analysis. In this manner, we ensure validity of the fixed effects assumption for the predictor effects of the proposed methods described in section 3. Just two variables, *sex* and *surg*, are included and to check the homogeneity assumption, we performed a multivariate meta-analysis allowing full random effects on the intercept and both predictor-outcome associations considered [129]. The corresponding model is specified as follows and was estimated using all 12 studies:

$$\begin{aligned}
y_{ij} &\sim \text{Bernoulli}\left(\pi_{ij}\right) \\
\text{logit}\left(\pi_{ij}\right) &= [a]_j + [b_{\text{sex}}]_j [\mathbf{X}_{\text{sex}}]_{ij} + [b_{\text{surg}}]_j [\mathbf{X}_{\text{surg}}]_{ij} \\
\begin{bmatrix} a \\ b_{\text{sex}} \\ b_{\text{surg}} \end{bmatrix}_j &\sim \text{MVN}\left(\begin{bmatrix} \alpha \\ \beta_{\text{sex}} \\ \beta_{\text{surg}} \end{bmatrix}, \begin{bmatrix} \tau^2_\alpha & \tau_{\alpha\beta_{\text{sex}}} & \tau_{\alpha\beta_{\text{surg}}} \\ \tau_{\alpha\beta_{\text{sex}}} & \tau^2_{\beta_{\text{sex}}} & \tau_{\beta_{\text{sex}}\beta_{\text{surg}}} \\ \tau_{\alpha\beta_{\text{surg}}} & \tau_{\beta_{\text{sex}}\beta_{\text{surg}}} & \tau^2_{\beta_{\text{surg}}} \end{bmatrix} \right)
\end{aligned} \tag{3.10}$$

Here, we found that $\hat{\alpha} = -1.80$ ($\hat{\tau}_\alpha = 0.47$ with a 95% CI of 0.42 - 0.55), $\hat{\beta}_{\text{sex}} = 0.47$ ($\hat{\tau}_{\beta_{\text{sex}}} = 0.03$ with a 95% CI of 0.01 - 0.29) and $\hat{\beta}_{\text{surg}} = 0.67$ ($\hat{\tau}_{\beta_{\text{surg}}} = 0.05$ with a 95% CI of 0.03 - 0.52). Because the between-study variability ($\hat{\tau}_\beta$) in the predictor-outcome associations for *sex* and *surg* appears negligible, we considered that assuming homogeneity was sensible and so used these predictors to derive a novel prediction model according to the approaches described in section 3. Results from the IECV are presented in Table **??**, for each of the stacking, random-effects on intercept, and stratified intercept approaches.

Consistency of estimated predictor-outcome associations

All approaches yielded similar and consistent predictor-outcome associations (estimates not shown) in the IECV. Particularly, we found that their average strength was reasonable ($0.80 < b_{\text{overall}} < 1.20$) in 8 of the 12 validation samples (Table **??**), which indicates that the modelled predictor-outcome associations were often comparable across studies. Accurate estimates of predictor-outcome associations could, however, not always be established in the validation studies. For instance, the predictor-outcome association for *sex* ($\hat{\beta}_{\text{sex,der}} = 0.49$) was unstable in study 3 ($\hat{\beta}_{\text{sex,val}} = -0.24$ with S.E. = 0.46) and in study 8 ($\hat{\beta}_{\text{sex,val}} = 0.16$ with S.E. = 0.26). It remains unclear whether the resulting discrepancy in predictor-outcome associations is due to heterogeneity or small effective sample size, but the latter is plausible given the small estimated heterogeneity for *sex* from the multivariate meta-analysis.

Quality of estimated model intercepts

Our results demonstrate that the derived prediction models do not validate well when using intercepts for a new population obtained through averaging individual intercepts of an IPD meta-analysis (i.e. through either the stacking or random effects approach). Particularly, these intercepts give an unequal proportion of predicted and observed outcomes, and considerably overestimate (E/O $\geq$ 1.2 in 4 of the 12 validation samples) or underestimate (E/O $\leq$ 0.8 in 3 of the 12 validation samples) the outcome presence. Similar results were obtained when using the stratified approach and selecting the intercept from a study with similar baseline descriptives of the new study population (i.e. matching the summary baseline characteristics from the validation study data to an IPD study used in model development, and using the latter's estimated intercept). The calibration improved greatly when the stratified approach was used and the chosen intercept was selected from an included study that had a similar observed outcome incidence (Table 3.2 and 3.3). For example, when study 1 was used as the validation data, the E/O statistic was 1.42 when using the weighted average intercept from random-effects model 3.3, but was 1.03 when using the intercept estimate for the study with the most similar incidence. However, even this approach does not guarantee good agreement between predicted and observed outcomes. Poor calibration may, for instance, arise when there are no studies with a similar outcome proportion or incidence available. This situation arose when study 2 (outcome incidence of 39% in the validation study versus 24% in the included study with the most similar incidence) or study 7 (outcome incidence of 8% versus 13%) were used as validation data in the IECV approach. In these validation studies, the outcome presence was considerably underestimated (E/O = 0.615 for study 2) and overestimated (E/O = 1.6 for study 7). Optimal agreement between predicted and observed outcomes was achieved when the intercept was estimated from the outcome proportion in the IPD for the new population by mean-centering the included predictor variables (cfr. *Intercept estimation from outcome prevalences*). Here, the E/O statistic is always close to 1, ranging between 1.00 and 1.03.

Quality of model predictions

Visual inspection of the calibration plots (Figure A.1. in the e-appendix) demonstrates that the stratified approach yields prediction models with superior calibration over the entire range of predicted probabilities when the final intercept is estimated from the outcome prevalence observed

Table 3.2: Results Case Study 1

Model Dvl	**Model Implementation**	**Model Performance**							
	Intercept choice	ID	E/O	b_{overall}	AUC	ID	E/O	b_{overall}	AUC
Stacking	overall estimate	1	1.50	0.89	0.58	7	2.60	0.94	0.57
Random Effects	estimated weighted average		1.42	0.89	0.58		2.38	0.89	0.57
Stratified	selected from outcome prev.		1.02	0.89	0.58		1.58	0.94	0.57
Stratified	selected from baseline descr.		3.02	0.89	0.58		2.95	0.94	0.57
Stratified	estimated from outcome prev.		1.03	0.89	0.58		1.03	0.94	0.57
Stacking	overall estimate	2	0.43	0.81	0.57	8	0.77	0.82	0.55
Random Effects	estimated weighted average		0.42	0.86	0.57		0.72	0.77	0.55
Stratified	selected from outcome prev.		0.62	0.81	0.57		0.96	0.82	0.55
Stratified	selected from baseline descr.		0.62	0.81	0.57		1.63	0.82	0.55
Stratified	estimated from outcome prev.		1.00	0.81	0.57		1.01	0.82	0.55
Stacking	overall estimate	3	1.18	1.32	0.65	9	0.84	0.98	0.58
Random Effects	estimated weighted average		1.14	1.22	0.65		0.80	0.98	0.58
Stratified	selected from outcome prev.		1.04	1.32	0.65		0.95	0.98	0.58
Stratified	selected from baseline descr.		1.04	1.32	0.65		0.76	0.98	0.58
Stratified	estimated from outcome prev.		1.03	1.32	0.65		1.02	0.98	0.58

Illustration of model performance in the internal-external cross-validation (Case Study 1) when dataset ID is used for validation and the remaining studies for derivation. The presented statistics are: the ratio of predicted to observed outcomes (E/O), the calibration slope (b_{overall}) and the area under the ROC curve (AUC). The standard error of the calibration slope and the AUC ranged from 0.18 to 0.69 and, respectively, from 0.02 to 0.06 [68].

Table 3.3: Results Case Study 1

Model Dvl	**Model Implementation**	**Model Performance**							
	Incercept choice	ID	E/O	b_{overall}	AUC	ID	E/O	b_{overall}	AUC
Stacking	overall estimate	4	0.77	1.24	0.60	10	1.36	1.20	0.61
Random Effects	estimated weighted average		0.75	1.24	0.60		1.32	1.21	0.61
Stratified	selected from outcome prev.		1.03	1.24	0.60		0.96	1.20	0.61
Stratified	selected from baseline descr.		0.90	1.24	0.60		1.19	1.20	0.61
Stratified	estimated from outcome prev.		1.02	1.24	0.60		1.03	1.20	0.61
Stacking	overall estimate	5	1.14	1.00	0.58	11	0.89	1.08	0.61
Random Effects	estimated weighted average		1.09	1.00	0.58		0.85	1.06	0.61
Stratified	selected from outcome prev.		0.95	1.00	0.58		1.04	1.08	0.61
Stratified	selected from baseline descr.		1.31	1.00	0.58		1.10	1.08	0.61
Stratified	estimated from outcome prev.		1.03	1.00	0.58		1.02	1.08	0.61
Stacking	overall estimate	6	1.14	0.84	0.59	12	1.41	0.76	0.55
Random Effects	estimated weighted average		1.09	0.84	0.59		1.37	0.76	0.55
Stratified	selected from outcome prev.		1.00	0.84	0.59		1.03	0.76	0.55
Stratified	selected from baseline descr.		0.95	0.84	0.59		1.23	0.76	0.55
Stratified	estimated from outcome prev.		1.03	0.84	0.59		1.03	0.76	0.55

Illustration of model performance in the internal-external cross-validation (Case Study 1) when dataset ID is used for validation and the remaining studies for derivation. The presented statistics are: the ratio of predicted to observed outcomes (E/O), the calibration slope (b_{overall}) and the area under the ROC curve (AUC). The standard error of the calibration slope and the AUC ranged from 0.18 to 0.69 and, respectively, from 0.02 to 0.06 [68].

in the new study population. Particularly, the calibration curve in these plots coincides with the 45° reference line, reflecting that predicted and actual probabilities agree for individual patients in the validation studies. Confidence intervals of the calibration curves are inflated for study 3 and 8, where the least data was available. Calibration curves for other approaches were similar (results not included), with curves shifted upwards and downwards according to underestimation (E/O < 1) and overestimation (E/O > 1) respectively of the outcome presence. Evaluation of the area under the ROC curve indicates that all approaches yielded prediction models with very similar discriminative ability. This statistic ranged from 0.55 to 0.65 across the different validation studies, suggesting that the predictors *sex* and *surg* poorly distinguish between patients with and without DVT. For instance, the interquartile ranges of predicted probabilities in validation study 3 ranged from 14 - 22 % and 13 - 19 % for cases and non-cases, respectively. In conclusion, model predictions appear to be well calibrated, but are not very informative as they are similar for cases and non-cases.

General Conclusions

For homogeneous predictor-outcome associations, we found that stratified estimation yields superior prediction model performance, particularly when the intercept is adapted to the new study population. This is best achieved by selecting the intercept from an available study in the meta-analysis that most closely matches the validation study according to the outcome proportion (prevalence), or by re-estimating the intercept from the outcome proportion or incidence in the IPD for the new (validation) population. Compared to using the average intercept, these approaches generally gave E/O ratios much closer to 1 in the validation study and yielded calibration curves that coincided with the 45° reference line. Unfortunately, derived models did not discriminate well because the included predictors *sex* and *surg* are not highly predictive. This implicates that risk predictions are quite accurate on a whole, but that the model cannot discriminate well between cases and non-cases. We therefore consider a second scenario where we include a set of strong predictors during model derivation.

Case Study 2: Strongly heterogeneous predictor-outcome associations

In the second scenario, we consider the derivation of a prediction model with important but heterogeneous predictors to investigate the impact of invalid homogeneity assumptions concerning the predictor-outcome associations across the included studies. Previous research identified *malign*, *surg*, *calfdif3* and *ddimdich* as core predictors for diagnosing DVT [66]. Consequently, we included these predictors from 10 of the 12 datasets to derive a novel prediction model, as two studies did not measure all variables. By performing a full random effects meta-analysis similar to Case Study 1, we found $\hat{\alpha} = -3.98$ ($\hat{\tau}_\alpha = 0.31$), $\hat{\beta}_{\text{malign}} = 0.38$ ($\hat{\tau}_{\beta_{\text{malign}}} = 0.35$), $\hat{\beta}_{\text{calfdif3}} = 1.05$ ($\hat{\tau}_{\beta_{\text{calfdif3}}} = 0.16$), $\hat{\beta}_{\text{surg}} = 0.25$ ($\hat{\tau}_{\beta_{\text{surg}}} = 0.09$) and $\hat{\beta}_{\text{ddimdich}} = 2.76$ ($\hat{\tau}_{\beta_{\text{ddimdich}}} = 0.41$). Clearly the heterogeneity estimates (τ values) are quite large for most variables. Results from the IECV are presented in Table 3.4 and 3.5.

Results in Table 3.4 and 3.5 demonstrate that all strategies for choosing intercepts perform poorly, as they generally give E/O ratios that are not close to 1, and thus considerably over- or underestimate the outcome prevalence when applied in other study populations. Even the strategies that performed very well in Case Study 1, that of estimating the intercept from the outcome prevalence in the validation study, or that of selecting an intercept from a study that most closely matched the outcome prevalence in the validation study, show poor performance on the whole.

Although the calibration slope b_{overall} is quite good in most validation samples, visual inspection of the calibration plots (Figure A.2. in the e-appendix) reveals that calibration curves of derived models strongly deviate from the 45° reference line. Accordingly, we may conclude that predicted probabilities do not correspond to actual outcome risks, and that the quality of model predictions is poor. This deterioration in calibration strongly contrasts with a considerable improvement in the discriminative ability of derived models. Whereas models from Case Study 1 achieved an AUC between 0.55 and 0.65 in the validation studies, the inclusion of *malign*, *surg*, *calfdif3* and *ddimdich* increased this statistic to values between 0.73 and 0.92.

In conclusion, when predictor-outcome associations in the IPD meta-analysis are strongly heterogeneous, we found that all approaches yield prediction models that generally have poor calibration when applied in the validation studies. This is likely due to model intercepts and predictor-outcome associations that do not correspond to the true intercepts and predictor-outcome associations in the validation studies because of heterogeneous predictor-outcome associations of the included variables. However, we found that the inclusion of these strong predictors did considerably improve the discriminative ability of derived prediction models. The resulting models are better able to discriminate between cases and non-cases, but yield inaccurate risk predictions, limiting their usefulness.

Case Study 3: Weakly heterogeneous predictor-outcome associations

In this last scenario, we attempt to derive a useful prediction model that both achieves good calibration (similar to Case Study 1) and good discrimination (similar to Case Study 2). To this purpose, we consider the derivation of a prediction model that includes the homogeneous predictors *sex* and *surg* from Case Study 1 and one strong predictor *calfdif3* from Case Study

Table 3.4: Results Case Study 2

Model Dvl	Model Implementation	Model Performance							
	Intercept choice	ID	E/O	b_{overall}	AUC	ID	E/O	b_{overall}	AUC
Stacking	overall estimate	1	1.42	0.88	0.80	6	0.87	0.95	0.84
Random Effects	estimated weighted average		1.35	0.84	0.80		0.83	0.92	0.84
Stratified	selected from outcome prev.		0.61	0.88	0.80		0.50	0.95	0.84
Stratified	selected from baseline descr.		1.43	0.88	0.80		1.14	0.95	0.84
Stratified	estimated from outcome prev.		1.68	0.88	0.80		1.49	0.95	0.84
Stacking	overall estimate	2	0.68	1.02	0.76	8	1.00	1.15	0.77
Random Effects	estimated weighted average		0.67	0.99	0.76		0.96	1.09	0.77
Stratified	selected from outcome prev.		0.67	1.02	0.76		1.14	1.15	0.77
Stratified	selected from baseline descr.		0.80	1.02	0.76		1.34	1.15	0.77
Stratified	estimated from outcome prev.		1.13	1.02	0.76		1.33	1.15	0.77
Stacking	overall estimate	3	0.76	1.35	0.92	9	1.25	0.95	0.76
Random Effects	estimated weighted average		0.71	1.32	0.92		1.17	0.93	0.76
Stratified	selected from outcome prev.		0.44	1.35	0.92		1.27	0.95	0.76
Stratified	selected from baseline descr.		0.81	1.35	0.92		0.70	0.95	0.76
Stratified	estimated from outcome prev.		1.39	1.35	0.92		1.37	0.95	0.76

Illustration of model performance in the internal-external cross-validation (Case Study 2) when dataset ID is used for validation and the remaining studies for derivation. The presented statistics are: the ratio of predicted to observed outcomes (E/O), the calibration slope (b_{overall}) and the area under the ROC curve (AUC). The standard error of the calibration slope and the AUC ranged from 0.07 to 0.24 and, respectively, from 0.02 to 0.04 [68].

Table 3.5: Results Case Study 2

Model Dvl	**Model Implementation**	**Model Performance**							
	Intercept choice	ID	E/O	b_{overall}	AUC	ID	E/O	b_{overall}	AUC
Stacking	overall estimate	4	0.80	1.16	0.85	11	0.90	0.96	0.83
Random Effects	estimated weighted average		0.78	1.15	0.85		0.84	0.92	0.83
Stratified	selected from outcome prev.		0.80	1.16	0.85		0.70	0.96	0.83
Stratified	selected from baseline descr.		0.88	1.16	0.85		1.03	0.96	0.83
Stratified	estimated from outcome prev.		1.35	1.16	0.85		1.41	0.96	0.83
Stacking	overall estimate	5	1.66	1.01	0.73	12	1.95	0.97	0.76
Random Effects	estimated weighted average		1.61	0.99	0.73		1.95	0.96	0.76
Stratified	selected from outcome prev.		2.03	1.01	0.73		1.34	0.97	0.76
Stratified	selected from baseline descr.		1.27	1.01	0.73		1.51	0.97	0.76
Stratified	estimated from outcome prev.		1.40	1.01	0.73		1.57	0.97	0.76

Illustration of model performance in the internal-external cross-validation (Case Study 2) when dataset ID is used for validation and the remaining studies for derivation. The presented statistics are: the ratio of predicted to observed outcomes (E/O), the calibration slope (b_{overall}) and the area under the ROC curve (AUC). The standard error of the calibration slope and the AUC ranged from 0.07 to 0.24 and, respectively, from 0.02 to 0.04 [68].

2. By performing a full multivariate random effects meta-analysis similar to Case Study 1, we found $\hat{\alpha} = -2.25$ ($\hat{\tau}_{\alpha} = 0.47$), $\hat{\beta}_{\text{sex}} = 0.37$ ($\hat{\tau}_{\beta_{\text{sex}}} = 0.06$), $\hat{\beta}_{\text{surg}} = 0.56$ ($\hat{\tau}_{\beta_{\text{surg}}} = 0.15$) and $\hat{\beta}_{\text{calfdif3}} = 1.28$ ($\hat{\tau}_{\beta_{\text{calfdif3}}} = 0.19$). The estimated τ values indicate that these predictor-outcome associations are weakly to moderately heterogeneous. Results from the IECV are presented in Table 3.6 and 3.7, and indicate that stratified estimation (where the final intercept is estimated from the outcome prevalence in the new study population, or selected from an available study in the meta-analysis that most closely matches the validation study according to the outcome proportion) again yields prediction models with superior performance. Specifically, this approach resulted into E/O ratios close to 1 in all validation studies. Furthermore, visual inspection of the calibration plots (Figure 3.2, 3.3 and 3.4) revealed good agreement, across the whole range, between predicted and actual outcome probabilities in at least 9 of the 12 validation studies (studies 1, 2, 4, 6, 9, 11 and 12). Studies 3, 8 and 10 showed poor calibration at predicted probabilities around 0.4, but as these studies also involved relatively small numbers of participants and events, it is difficult to know whether this is due to chance or a truly poor prediction performance in these settings. To be cautious, one could consider discarding these studies when fitting the final model, but our judgment was to leave them in. Finally, the discriminative ability of derived models was relatively good, and ranged between 0.64 and 0.76 across the validation studies. Consequently, the inclusion of weakly to moderately heterogeneous predictors resulted into prediction models that both discriminate and calibrate well in new patient populations.

DISCUSSION

An increasing number of prediction models are derived from an IPD-MA. Very little guidance currently exists about how researchers should account for the inherent potential for between-study heterogeneity, and how to implement the model in practice when outcome frequencies (baseline risks) differ across included study populations. As a consequence, many prediction models ignore clustering of participants and thus effectively assume they are using IPD from a single study. This straightforward stacking of IPDs is often not justified and, as we show in our Case Study 1 (Table 3.2 and 3.3), may lead to inconsistent model performance and considerably reduced generalizability. We therefore considered two other approaches to account for heterogeneity of baseline risk (random effects or stratified estimation), and evaluated several techniques to implement the developed model in a new clinical setting where the baseline risk is potentially unknown.

When there is homogeneity in predictor-outcome associations, stratified estimation of the model intercept helps to improve generalizability. This approach allows to derive a near-optimal intercept from reported outcome incidences when predictor variables are centered around their local means

Table 3.6: Results Case Study 3

Model Dvl	**Model Implementation**	**Model Performance**							
	Intercept choice	ID	E/O	b_{overall}	AUC	ID	E/O	b_{overall}	AUC
Stacking	overall estimate	1	1.51	0.89	0.68	7	2.27	1.04	0.69
Random Effects	estimated weighted average		1.41	0.90	0.68		2.07	1.01	0.69
Stratified	selected from outcome prev.		0.91	0.89	0.68		1.38	1.04	0.69
Stratified	selected from baseline descr.		2.73	0.89	0.68		2.65	1.04	0.69
Stratified	estimated from outcome prev.		1.15	0.89	0.68		1.15	1.04	0.69
Stacking	overall estimate	2	0.51	0.74	0.65	8	0.76	0.92	0.66
Random Effects	estimated weighted average		0.49	0.74	0.65		0.71	0.93	0.66
Stratified	selected from outcome prev.		0.70	0.74	0.65		1.02	0.92	0.66
Stratified	selected from baseline descr.		0.70	0.74	0.65		1.46	0.92	0.66
Stratified	estimated from outcome prev.		1.03	0.74	0.65		1.08	0.92	0.66
Stacking	overall estimate	3	1.28	1.36	0.76	9	0.98	0.97	0.69
Random Effects	estimated weighted average		1.23	1.37	0.76		0.92	0.99	0.69
Stratified	selected from outcome prev.		1.02	1.36	0.76		1.10	0.97	0.69
Stratified	selected from baseline descr.		1.18	1.36	0.76		0.78	0.97	0.69
Stratified	estimated from outcome prev.		1.14	1.36	0.76		1.10	0.97	0.69

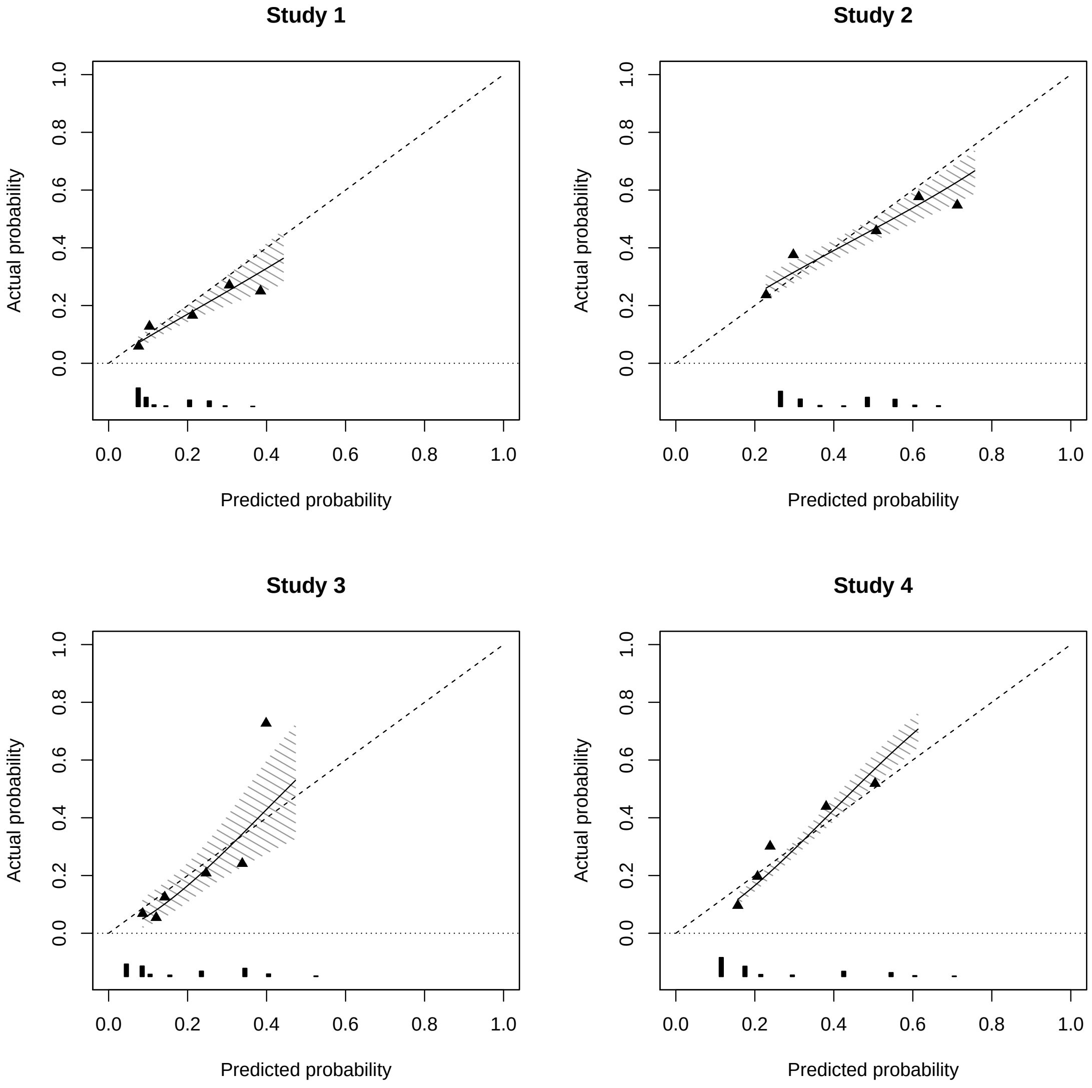
Illustration of model performance in the internal-external cross-validation (Case Study 3) when dataset ID is used for validation and the remaining studies for derivation. The presented statistics are: the ratio of predicted to observed outcomes (E/O), the calibration slope (b_{overall}) and the area under the ROC curve (AUC). The standard error of the calibration slope and the AUC ranged from 0.10 to 0.35 and, respectively, from 0.02 to 0.05 [68].

Table 3.7: Results Case Study 3

Model Dvl	**Model Implementation**	**Model Performance**							
	Intercept choice	ID	E/O	b_{overall}	AUC	ID	E/O	b_{overall}	AUC
Stacking	overall estimate	4	0.70	1.35	0.74	10	1.12	0.68	0.64
Random Effects	estimated weighted average		0.69	1.43	0.74		1.06	0.69	0.64
Stratified	selected from outcome prev.		0.99	1.35	0.74		0.68	0.68	0.64
Stratified	selected from baseline descr.		0.87	1.35	0.74		1.01	0.68	0.64
Stratified	estimated from outcome prev.		1.07	1.35	0.74		1.10	0.68	0.64
Stacking	overall estimate	5	1.28	0.72	0.65	11	0.88	0.91	0.70
Random Effects	estimated weighted average		1.22	0.74	0.65		0.83	0.92	0.70
Stratified	selected from outcome prev.		1.00	0.72	0.65		0.91	0.91	0.70
Stratified	selected from baseline descr.		1.30	0.72	0.65		1.17	0.91	0.70
Stratified	estimated from outcome prev.		1.14	0.72	0.65		1.10	0.91	0.70
Stacking	overall estimate	6	1.10	1.02	0.70	12	1.59	0.91	0.67
Random Effects	estimated weighted average		1.04	1.03	0.70		1.53	0.92	0.67
Stratified	selected from outcome prev.		0.87	1.02	0.70		1.43	0.91	0.67
Stratified	selected from baseline descr.		0.85	1.02	0.70		1.25	0.91	0.67
Stratified	estimated from outcome prev.		1.12	1.02	0.70		1.17	0.91	0.67

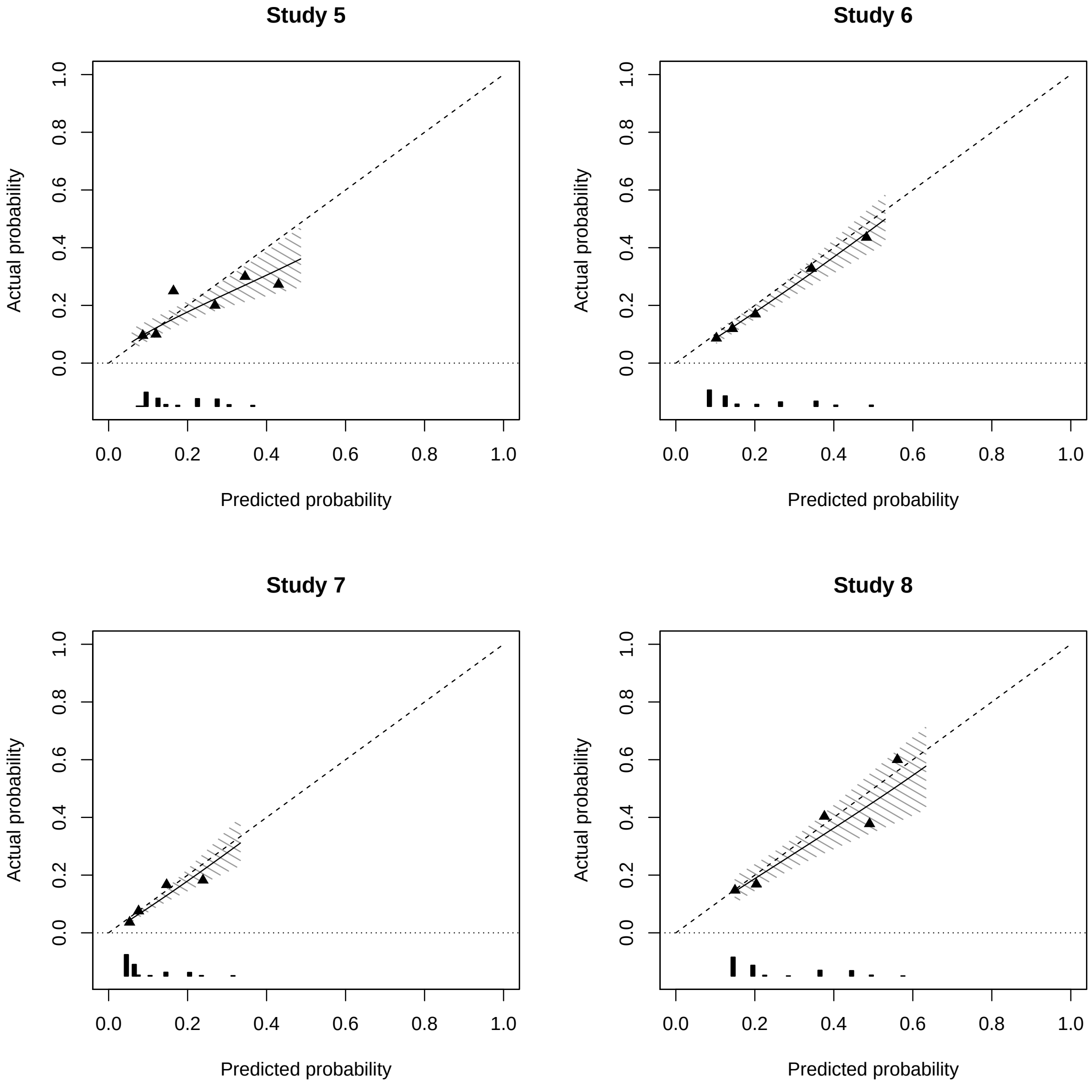
Illustration of model performance in the internal-external cross-validation (Case Study 3) when dataset ID is used for validation and the remaining studies for derivation. The presented statistics are: the ratio of predicted to observed outcomes (E/O), the calibration slope (b_{overall}) and the area under the ROC curve (AUC). The standard error of the calibration slope and the AUC ranged from 0.10 to 0.35 and, respectively, from 0.02 to 0.05 [68].

Figure 3.2: Calibration plots for study 1–4 (Case Study 3)



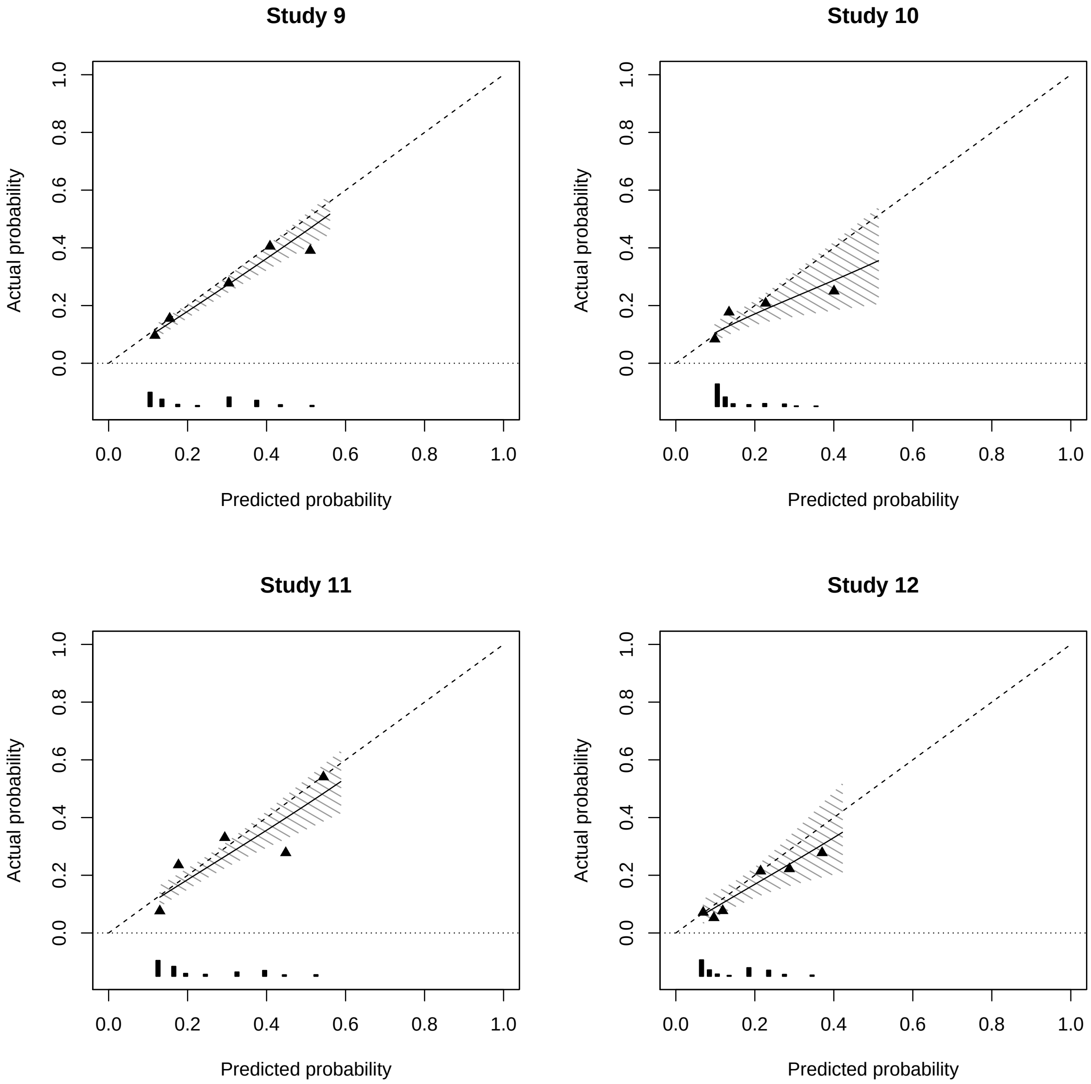
Calibration plots of models derived by stratified estimation of the intercept (where the final intercept is estimated from the outcome proportion in the validation study) in the validation studies of Case Study 3. The triangles indicate groups of observations with similar predicted probabilities and their corresponding outcome proportion. Note that a maximum of 8 groups can be generated because the included predictor variables *sex*, *surg* and *calfdif3* are dichotomous.

Figure 3.3: Calibration plots for study 5–8 (Case Study 3)



Calibration plots of models derived by stratified estimation of the intercept (where the final intercept is estimated from the outcome proportion in the validation study) in the validation studies of Case Study 3. The triangles indicate groups of observations with similar predicted probabilities and their corresponding outcome proportion. Note that a maximum of 8 groups can be generated because the included predictor variables *sex*, *surg* and *calfdif3* are dichotomous.

Figure 3.4: Calibration plots for study 9–12 (Case Study 3)



Calibration plots of models derived by stratified estimation of the intercept (where the final intercept is estimated from the outcome proportion in the validation study) in the validation studies of Case Study 3. The triangles indicate groups of observations with similar predicted probabilities and their corresponding outcome proportion. Note that a maximum of 8 groups can be generated because the included predictor variables *sex*, *surg* and *calfdif3* are dichotomous.

(see *Intercept estimation from outcome prevalences*). Alternatively, an estimated intercept can be selected from existing studies in the meta-analysis using the outcome incidence or prevalence in the new study population (see *Intercept Selection*). When no information about the population of interest is available, using the average intercept (for instance obtained by random effects or stacking) presents a workable solution, but generally this may cause poor calibration when baseline risks strongly differ (see *Stacking* and *Random effects modeling of the intercept*). In such situations, the IECV ("internal-external cross-validation") technique may be particularly helpful to identify the generalizability of derived prediction models across other study populations [215]. It allows the model fit and its predictive ability to be appraised across several studies, and ultimately allows a single (final) prediction model to be built using as much of the data as possible. It also identifies which populations (if any) the model is not suitable for, and helps ascertain the strategy for choosing an intercept, an additional validation step to gain insight into the future generalizability of the newly constructed model.

Some important limitations need to be considered to fully appraise the findings of this study. Firstly, the inclusion of homogeneous predictors may not always yield highly discriminative prediction models. Weakly heterogeneous but strong predictors may therefore be included to improve discrimination at the cost of model calibration. Heterogeneity may further be reduced by including additional covariates, non-linear associations or interaction terms, or by applying bootstrap and shrinkage techniques [23, 59, 149, 237, 244]. Secondly, when many but relatively small studies are available, stratified estimation may no longer be feasible due to its inherent model complexity. In such scenarios, random intercept effects modeling may considerably reduce the amount of unknown parameters whilst still allowing individual study intercepts. Thirdly, our case studies indicate that IPD-MA developers should report estimated model intercepts and corresponding outcome frequencies of included studies when their baseline risks are heterogeneous. In this manner, the derivation of an appropriate model intercept can be facilitated when the model is to be implemented or externally validated in new study populations. Note that it is possible to further improve the intercept choice by estimating an appropriate intercept from characteristics of the new study population. Further research might therefore consider a Bayesian approach to this framework and the selection of an intercept. Finally, it is often difficult to obtain IPD with the same and prognostically important information, especially if datasets were originally collected for a different purpose. Consequently, missing data is likely to be a common challenge in IPD-MA, and advanced imputation methods may be required to appropriately address their hierarchical nature. Future research will investigate the performance of several imputation methods, adopting a frequentist or Bayesian perspective.

In conclusion, in this article we have recommended steps for developing, implementing, and evaluating a risk prediction model when IPD from multiple studies are available (Figure 3.1). For model development, stratified estimation appears to be the most promising approach, which accounts for clustering of patients within studies and thereby allows a separate intercept per study. For implementation and external validation of this model, the predictor-outcome associations can be combined with the population's intercept as estimated from the outcome prevalence in the new population, or by taking the estimated intercept for one of the studies included in the model development whose outcome incidence closely matches that in the new population. Alternatively, it is possible to implement the population's intercept as estimated from IPD available for this population. Performance of the model and intercept strategy can be evaluated using the IECV approach. A reliable model that is generalizable across all studies is facilitated by homogeneity in predictor-outcome associations; however restricting inclusion to just homogeneous predictors may cause the model to have poor discrimination and so weakly heterogeneous predictors might also be considered. Further research is needed to evaluate how between-study differences in predictor-outcome associations could be addressed appropriately.

ACKNOWLEDGEMENTS

We gratefully acknowledge the following authors for sharing of individual participant data from the deep vein thrombosis (DVT) studies: A.J. Ten Cate-Hoek, R. Oudega, K.G.M. Moons, R.E.G. Schutgens, D.R. Anderson, P.S. Wells, R.A. Kraaijenhagen, D.B. Toll, C. Kearon, J.L. Elf, S.M. Stevens and S.M. Bates. Finally, we thank the editor, associate editor, and two anonymous reviewers from *Statistics in Medicine* for their constructive comments that helped improve this article.

4

Individual Participant Data Meta-Analysis with systematically missing predictors: an empirical example

Thomas P. A. Debray
Hendrik Koffijberg
Shahab Jolani
Yvonne Vergouwe
Daan Nieboer
Ewout W. Steyerberg
Stef van Buuren
Karel G. M. Moons

Abstract

Individual participant data meta-analyses (IPD-MA) are increasingly used for developing multivariable risk prediction models. Unfortunately, some predictors may not have been measured in each study and are systematically missing in the IPD-MA. As a consequence, predictor effects can no longer be estimated in each study, certainly when the clustering of subjects within studies need to be accounted for.

We used a case study to develop a multivariable logistic regression model for predicting the presence of Deep Venous Thrombosis (DVT) in subjects with a suspected DVT. Hereto, 12 datasets with a total of 8 974 subjects (1 733 events) were used for model development. These data contain 15 predictors of which 5 are systematically missing. We evaluated four approaches to deal with the resulting missingness. The first approach simply excludes studies with any systematically predictors (FCA). The second approach simply ignores the missing predictors when developing a model (FPA). The third approach imputes missing data by stacking the study datasets and applying multiple imputation ignoring the clustering of subjects within studies (TMI). Finally, the fourth approach allows for a study-specific intercept in the imputation model to account for clustering (SMI). An external validation study with 1 028 subjects (131 events) was used for evaluating the performance of the derived prediction models in terms of their discrimination and calibration.

We found a *c*-statistic of 0.82 (FCA), 0.68 (FPA), 0.82 (TMI) and 0.82 (SMI) in the validation sample. The calibration-in-the-large was 0.48 (FCA), -0.25 (FPA), -0.07 (TMI) and 0.11 (SMI). The calibration slope was 0.81 (FCA), 0.85 (FPA), 0.85 (TMI) and 0.4 (SMI). Finally, we found that estimates for between-study heterogeneity inflated when applying imputation strategies (as compared to full case analysis).

Our study demonstrates that an IPD-MA with systematically missing predictors does not need to discard studies or predictors, and may benefit from imputation strategies. Ignoring systematically missing predictors generally leads to poorest model performance, particularly when important predictors are excluded.

Chapter 4

"We demand rigidly defined areas of doubt and uncertainty!"

– Douglas Adams, *The Hitchhiker's Guide to the Galaxy*

An important aim in diagnostic and prognostic research is the development of clinical prediction models. These models predict whether a certain outcome is present (diagnosis) or will occur (prognosis) in a subject by relying on several predictors. These may range from individual characteristics, signs and symptoms, to results of more invasive or costly measures such as imaging, electrophysiology, blood, urine, coronary plaque or even genetic markers [170, 201, 247]. The development of a novel prediction model typically utilizes a so-called individual participant dataset (IPD). This dataset contains the predictor values and final diagnosis of several subjects, and is ideally obtained from a prospective cohort study [170]. However, during the past decades, the popularity of prediction research has increased and international collaboration has become more commonplace. Individual participant datasets are therefore frequently combined when developing or validating a novel prediction model. This strategy is also known as individual participant data meta-analysis (IPD-MA) [6, 68, 188, 215, 222, 248].

A key issue in an IPD-MA is the presence of between-study heterogeneity, i.e. clustering of participants within studies. Heterogeneity typically manifests as differences in baseline risks, that is, outcome prevalences (for diagnostic models) or incidences (for prognostic models), or in the predictor-outcome associations. Recently, Debray *et al.* proposed a framework for developing, implementing, and validating a risk prediction model when IPD from multiple studies are available [68]. This framework accounts for between-study heterogeneity in baseline risk by estimating a stratified intercept term (or baseline hazard) for each study. It also proposes to pursue homogeneity in predictor-outcome associations during model development as to improve the model's generalizability across all studies. For implementation and external validation of this model, the estimated predictor-outcome associations can then be combined with an intercept term that is appropriate for the local circumstances.

Unfortunately, it may arise that the studies from an IPD-MA measured different subject characteristics or performed different (biomarker) tests, e.g. due to cost constraints. When combining the corresponding datasets, some predictors are no longer complete and become systematically missing [45, 253]. As a consequence, researchers often choose to exclude entire studies with one or more missing predictors from the IPD-MA [248]. Alternatively, systematically missing predictors are ignored during model development. It may be clear that this approach is undesirable as available evidence is not optimally used, and certainly if the missing predictors are known to be important.

In this article, we investigate four simple approaches for developing a prediction model from an IPD-MA when some predictors are systematically missing. Each approach can be implemented with standard software packages, and makes different assumptions about the missing data mechanisms. The first approach excludes entire studies where predictors are systematically missing during model development. The second approach does not include systematically missing predictors in the prediction model. The third approach treats the IPD-MA as a single dataset, and uses traditional multiple imputation strategies for dealing with missing data. Finally, the fourth approach implements an extension of the third approach that accounts for clustering of participants within studies. We illustrate each approach in an empirical example of predicting Deep Venous Thrombosis (DVT).

METHODS

In this section, we perform a case study to illustrate four approaches for developing a logistic regression model in an IPD-MA with systematically predictors. We consider 2 scenarios in which different amounts of studies are at hand, and measure several performance statistics in an independent validation sample. All approaches were implemented in R version 2.15.1, using *mice* 2.17 (multiple imputation) and *lme4* 0.999999-2. The corresponding source code is available on request.

Case Study

In order to illustrate the approaches we describe a clinical example involving the diagnosis of Deep Vein Thrombosis (DVT) presence. DVT is a blood clot that forms in a vein in the body and may lead to blockage in the lungs, preventing oxygenation of the blood and potentially causing death. Clinical DVT diagnosis is not straightforward. For this reason, multivariable diagnostic prediction models have been developed to predict the probability of presence of DVT in suspected patients [182, 261, 281]. These models use the results from history taking, physical examination and blood tests as predictors for DVT presence. They may subsequently be used to safely exclude DVT without having to perform further testing.

In this case study, we use an IPD meta-analysis of 13 studies conducted for diagnosing DVT in patients with a suspected DVT. The IPD-MA contains a total of 10 002 subjects of which 1 864 (18.6%) truly have DVT (Table A.1 in the Appendix) [17, 18, 24, 44, 73, 137, 138, 140, 182, 223, 232, 260, 261, 282]. The following 10 predictors were measured in all studies: male gender (*sex*), active malignancy (*malign*), paresis (*par*), recent surgery or bedridden (*surg*), localized tenderness deep venous system (*tend*), entire leg swollen (*leg*), calf difference >= 3 cm (*cdif3*), pitting edema (*pit*), vein distension (*vein*) and alternative diagnosis present (*adiag*). Finally, 5 predictors were systematically missing in one or more studies: D-dimer positive (*ddimd*), no leg trauma present (*notraum*), family history of thrombofilia (*coag*), oral contraceptive use (*oachst*) and history of previous DVT (*hdvt*)

Approaches

We consider four approaches to develop a prediction model with a binary outcome from an IPD-MA when some predictors are systematically missing across studies. In general, the presence of missing data can be described by three mechanisms with different assumptions about the probability of missingness. When this probability is identical for all subjects, predictors are missing completely at random (MCAR). Conversely, missing at random (MAR) occurs when the probability of missingness depends on the observed information. Finally, missing not at random (MNAR) occurs when the probability of missingness depends on the (potentially missing) predictor itself or on other predictors that have not been measured.

In the presence of missing data it is common to assume MAR and to apply multiple imputation. This approach generates several copies of the original dataset and replaces missing values by values drawn from a multivariate distribution (joint modeling) [220] or from a set of conditional densities (fully conditional specification) [267, 268, 286]. The imputed datasets are then analyzed separately and resulting model estimates are pooled using Rubin's rule [216].

In an IPD-MA with sporadically missing data it is generally recommended to apply multiple imputation separately for each study IPD or to implement a hierarchical imputation model [45, 285, 287]. These approaches allow each IPD to have a different covariance structure, and can thereby accommodate for between-study heterogeneity. Particularly hierarchical imputation models are appealing because they allow to share information between the available IPD. Unfortunately, implementation of these approaches is not feasible in the presence of systematically missing predictors. This is because study-specific associations are no longer identifiable for all relevant predictors. Researchers therefore often have to fall back on simpler but possibly flawed approaches. We list some common approaches below, and present a simple imputation algorithm that accounts for clustering of subjects within studies.

- Full case analysis (FCA). The most common approach to deal with missing data is to exclude studies where important predictors have not been measured. This approach actually assumes that the occurrence of systematic missingness in predictors is MCAR on the study level.
- Full predictor analysis (FPA). An alternative approach may simply discard systematically missing predictors during model development. This approach has the advantage that no assumptions are made about the missing data mechanisms. Unfortunately, it also implies

that the performance and generalizability of the prediction model may degrade, particularly when missing predictors are known to be important.

- Traditional Multiple Imputation (TMI). More sophisticated approaches assume that the missingness mechanisms depend on the observed data only (MAR). Because it is not possible to impute each dataset separately (as systematically missing predictors will remain present), one may choose to stack all study IPD and treat them as a single dataset during imputation [139]. This traditional approach, however, does not account for clustering of subjects within studies, and assumes a common covariance structure for all IPD. In addition, the traditional imputation approach may lead to inconsistencies when subsequent model development actually does account for clustering (e.g. by estimating a stratified intercept term). It is widely acknowledged that imputation models should be *congenial* with the analysis model [45, 178]. Generally speaking, this implies that predictors should be treated similarly during imputation of missing data and the estimation of a prediction model. For instance, if a prediction model estimates stratified a intercept term to account for between-study heterogeneity, the imputation model should adopt the same strategy. Unfortunately, traditional imputation strategies cannot accommodate for between-study heterogeneity and may therefore lead to bias in estimated predictor effects and standard errors [45, 139].

- Stratified Multiple Imputation (SMI). Finally, it is possible to address the pitfalls of TMI by extending the imputation model with a clustering term. This can be achieved fairly straightforward by creating a categorical variable with the study identification number [90, 284]. The corresponding variable is typically dummy-coded and can be used as additional predictor(s) in the imputation model. As a consequence, the imputation model includes a study-specific (stratified) intercept term and thereby accommodates for clustering of subjects within studies. This implies that the imputation of a missing predictor allows a study-specific prevalence (or mean for continuous predictors) for that predictor. The estimation of stratified intercept terms is, however, problematic when predictors are systematically missing [45]. This is because study-specific intercept terms are no longer identifiable in studies where predictors are systematically missing (i.e. the study-specific mean or prevalence is unknown). Fortunately, it is possible to evaluate which studies would have similar study-specific intercept terms, even when they cannot be estimated. This can be achieved by constructing a baseline table with the prevalence (or mean) of each predictor in each study, and imputing missing entries in this table (due to systematic missingness) using predictive mean matching. Studies with a similar prevalence (or mean) can then be identified for each systematically missing predictor and subsequently be treated as a single study. This can be achieved by creating a categorical grouping variable for each predictor in the imputation model. Each missing predictor can therefore be imputed using a different grouping of studies, thereby preserving a substantial degree of clustering. The general strategy of stratified imputation is to combine studies by predicted prevalence of systematically missing predictors for each imputation model of a systematically missing predictor, in such a way that resulting groups are no longer affected by systematic missingness. If no distinct groups can be formed (e.g. when all studies need to be combined to avoid systematic missingness), stratified imputation will simply ignore clustering of subjects and collapse to TMI.

Scenarios

We consider two scenarios to develop a multivariable logistic regression model in an IPD-MA for the prediction of DVT presence. In the first scenario, 12 studies ($N = 8\,974$, Table 4.1 and 4.2) were available for model development. Conversely, the IPD-MA of the second scenario only includes 6 studies ($N = 4\,466$, ID 2, 5, 6, 10, 11, 13 in Table 4.1 and 4.2) of which 5 studies are affected by systematic missingness. In each scenario, we first performed data completion by applying FCA (full case analysis), FPA (full predictor analysis), TMI (traditional multiple imputation) or SMI (stratified multiple imputation). We subsequently developed a multivariable risk prediction model with a predefined set of 8 predictors (Table 4.3) [182] and a stratified intercept term to allow baseline risk to differ across studies [68].

Model evaluation

We evaluated the performance of the developed prediction models in an independent dataset that was not used by the aforementioned scenarios. The corresponding validation sample contains 1 028 subjects of which 131 were diagnosed with DVT. For each prediction model, an appropriate intercept term for the validation sample was selected from the development study with the most similar outcome prevalence (as compared to 13% in the validation sample) [68]. Model performance was then quantified in terms of discrimination and calibration [83, 101, 237]. Discrimination is the ability to distinguish high-risk subjects from low-risk subjects, and is typically quantified by the concordance (c) statistic (which should ideally be 1) [77, 116]. Calibration reflects the extent to which the predicted probabilities and actual probabilities agree, and can be quantified by the calibration-in-the-large and calibration slope statistics. The calibration-in-the-large quantifies whether the average of predictions corresponds with the average outcome frequency, and ideally equals 0. Conversely, the calibration slope reflects whether predicted risks are appropriately scaled with respect to each other over the entire range of predicted probabilities, and ideally equals 1 [60, 160]. A calibration slope below 1 usually reflects overfitting of the model in the development sample, but may also indicate inconsistency of predictor effects between the development and

Table 4.1: Baseline descriptives of the IPD in the DVT case study (scenario 1)

ID	N	dvt=1	sex=1	malign=1	par=1	surg=1	tend=1	leg=1	cdif3=1
2	814	39%	38%	11%	4%	9%	66%	21%	43%
3	153	17%	48%	5%	8%	16%	54%	33%	39%
4	1 756	23%	37%	13%	6%	11%	38%	35%	24%
5	791	16%	38%	5%	14%	13%	72%	45%	41%
6	1 075	18%	44%	5%	5%	16%	46%	20%	28%
7	429	14%	36%	11%	3%	6%	47%	7%	22%
8	325	16%	39%	4%	4%	5%	50%	14%	29%
9	1 295	22%	36%	6%	14%	14%	71%	45%	43%
10	436	10%	33%	6%	3%	13%	51%	39%	15%
11	541	22%	44%	18%	6%	18%	44%	28%	30%
12	550	10%	38%	9%	2%	7%	32%	15%	21%
13	809	5%	40%	7%	3%	6%	32%	9%	19%

Variables in the DVT data. Prevalences for systematically missing variables (underlined) were imputed using predictive mean matching. *dvt* (1=DVT presence, 0=no DVT)

Table 4.2: Baseline descriptives of the IPD in the DVT case study (scenario 1)

	pit=1	vein=1	adiag=1	ddimd=1	notraum=1	coag=1	oachst=1	hdvt=1
2	51%	16%	27%	73%	67%	18%	5%	0%
3	48%	16%	48%	69%	67%	18%	10%	3%
4	54%	16%	52%	52%	85%	4%	5%	9%
5	62%	20%	38%	72%	82%	24%	10%	18%
6	33%	4%	42%	39%	80%	5%	10%	15%
7	20%	8%	41%	73%	69%	18%	5%	0%
8	30%	12%	35%	66%	67%	18%	10%	3%
9	62%	20%	60%	69%	85%	24%	10%	24%
10	21%	0%	27%	52%	80%	19%	10%	9%
11	50%	7%	58%	49%	69%	3%	10%	24%
12	46%	5%	45%	66%	80%	3%	10%	3%
13	24%	4%	49%	39%	82%	4%	10%	9%

Variables in the DVT data. Prevalences for systematically missing variables (underlined) were imputed using predictive mean matching. *dvt* (1=DVT presence, 0=no DVT)

Table 4.3: Overview of heterogeneity in estimated regression coefficients, depicted by the square root of τ^2 and by the ICC.

Scenario 1	**FCA**		**FPA**		**TMI**		**SMI**	
	τ	ICC	τ	ICC	τ	ICC	τ	ICC
(Intercept)	0.50	(0.07)	0.58	(0.09)	0.48	(0.04)	0.47	(0.06)
sex	0.02	(0.00)	0.07	(0.00)	0.09	(0.00)	0.05	(0.00)
oachst†	0.04	(0.00)			0.24	(0.01)	0.20	(0.01)
malign	0.03	(0.00)	0.44	(0.05)	0.43	(0.03)	0.41	(0.04)
surg	0.01	(0.00)	0.20	(0.01)	0.14	(0.00)	0.14	(0.00)
notraum†	0.23	(0.01)			0.15	(0.00)	0.17	(0.01)
vein	0.12	(0.00)	0.20	(0.01)	0.26	(0.01)	0.27	(0.02)
cdif3	0.21	(0.01)	0.21	(0.01)	0.18	(0.01)	0.19	(0.01)
ddimd†	0.07	(0.00)			0.50	(0.04)	0.43	(0.05)
Scenario 2	**FCA‡**		**FPA**		**TMI/SMI***			
	τ	ICC	τ	ICC	τ	ICC		
(Intercept)			0.72	(0.13)	0.74	(0.12)		
sex			0.07	(0.00)	0.13	(0.00)		
oachst†					0.33	(0.03)		
malign			0.50	(0.06)	0.42	(0.04)		
surg			0.17	(0.01)	0.20	(0.01)		
notraum†					0.14	(0.01)		
vein			0.21	(0.01)	0.33	(0.02)		
cdif3			0.13	(0.00)	0.15	(0.00)		
ddimd†					0.52	(0.06)		

Estimates for τ were obtained by fitting a mixed effects model with full random effects on the intercept and predictoroutcome associations. Standard errors are not presented because confidence intervals around variance components are not symmetrically distributed (but instead follow a scaled X^2 distribution), and because *lme4* does not provide such estimates for exactly the same reason.

† These predictor variables were systematically missing in one or more studies.

‡ Between-study heterogeneity could not be assessed as only one study was left in the anlyses.

* TMI and SMI are the same approach because no distinct groups could be formed.

validation sample [132, 239, 249, 270, 275]. The quality of predicted risks was further investigated by visual inspection of the calibration plot. This plot indicates how predicted risks diverge from observed outcomes in different deciles of predicted risks and shows perfect predictions when the calibration curve goes through the origin and has a slope of 45°.

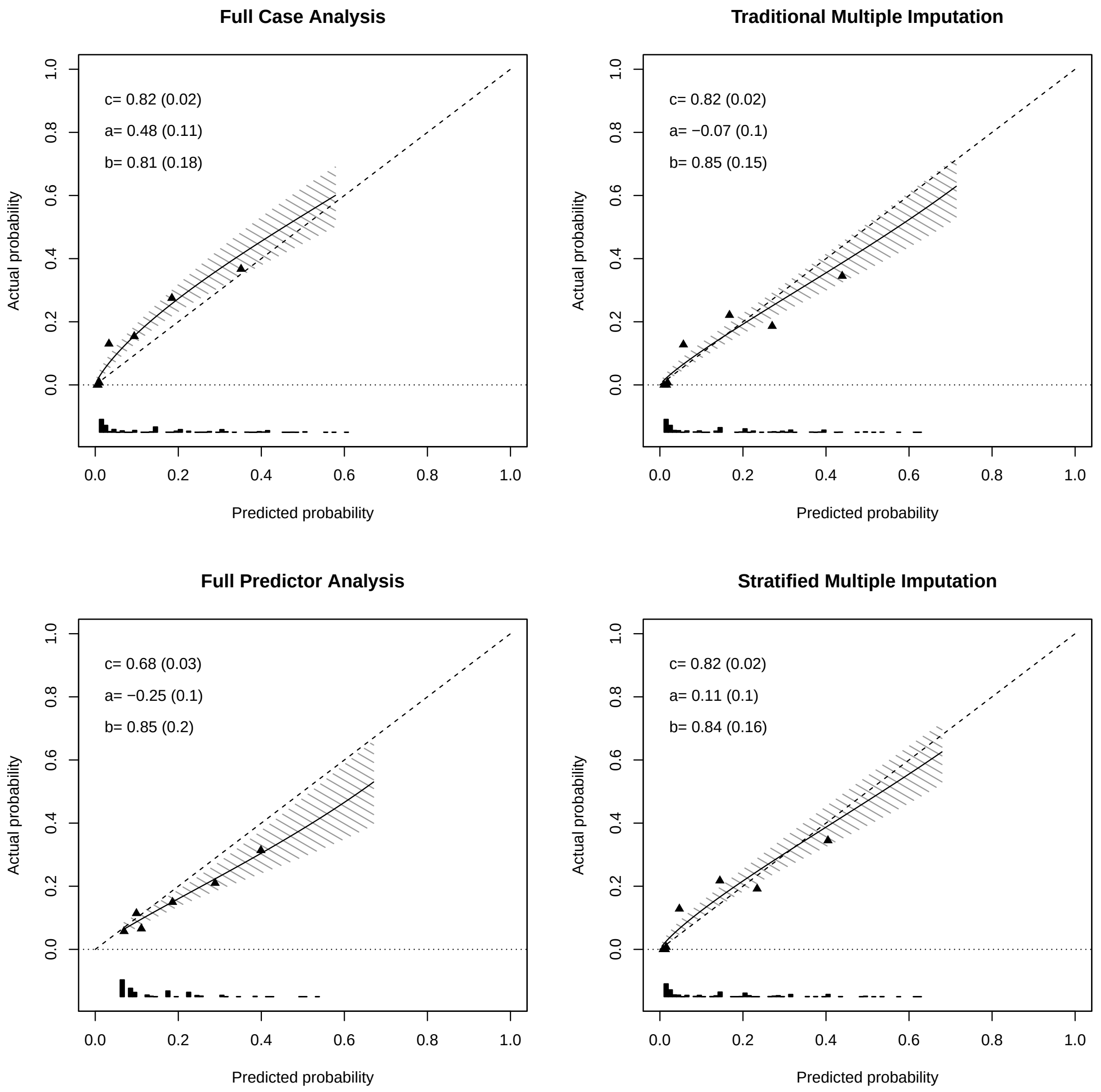
Finally, we evaluated the extent to which aforementioned approaches affect the perception of between-study heterogeneity in predictor effects. Hereto, we used the completed development sample to perform a full one-stage multivariate meta-analysis using the model's predictors (model C in [67]). This implies that 9 (FCA, TMI and SMI) and respectively 6 (FPA) predictors were included as joint random effects in a mixed effect model. The between-study heterogeneity for each regression coefficient β_j was then quantified in terms of its variance across studies, i.e. τ_j^2 [68], and its Intra-Class Correlation (ICC). The ICC quantifies dependence among subjects [183] and can be calculated as $\tau_j^2/(\sum_j \tau_j^2 + \pi^2/3)$ for a logistic regression coefficient β_j.

CASE STUDY RESULTS

Full case analysis

Results from scenario 1 (Figure 4.1) illustrate that FCA yields adequate discrimination, even though 9 studies are discarded during model development and 3 842 subjects (826 events) remain for estimation purposes. Model calibration was reasonable (calibration slope = 0.81), but substantially deteriorated (calibration slope = 0.67) in scenario 2 where only one study remained for model development (Figure 4.2). Indeed, visual inspection of the calibration plot indicated that the calibration curve of the prediction model strongly deviated from the 45° reference line. As a consequence, the prediction model in scenario 2 may be prone to overfitting, or may be affected substantially by between-study heterogeneity of predictor effects. Finally, we noticed that the calibration-in-the-large was relatively poor in both scenarios, and corresponded to a predicted outcome prevalence of 9% in the validation sample (observed: 13%). Results in Table 4.3 further indicate that the between-study heterogeneity of predictor effects in the development sample was relatively low, except for the intercept term ($\beta_0 = -5.11$ with $\tau_0 = 0.50$) and the predictors *notraum* ($\beta_{\text{notraum}} = 0.58$ with $\tau_{\text{notraum}} = 0.23$) and *vein* ($\beta_{\text{vein}} = 0.46$ with $\tau_{\text{vein}} = 0.12$). Unfortunately, evaluation of between-study heterogeneity was not possible in scenario 2, as only one study remained for model development.

Figure 4.1: Calibration plots of estimated models (Scenario 1).

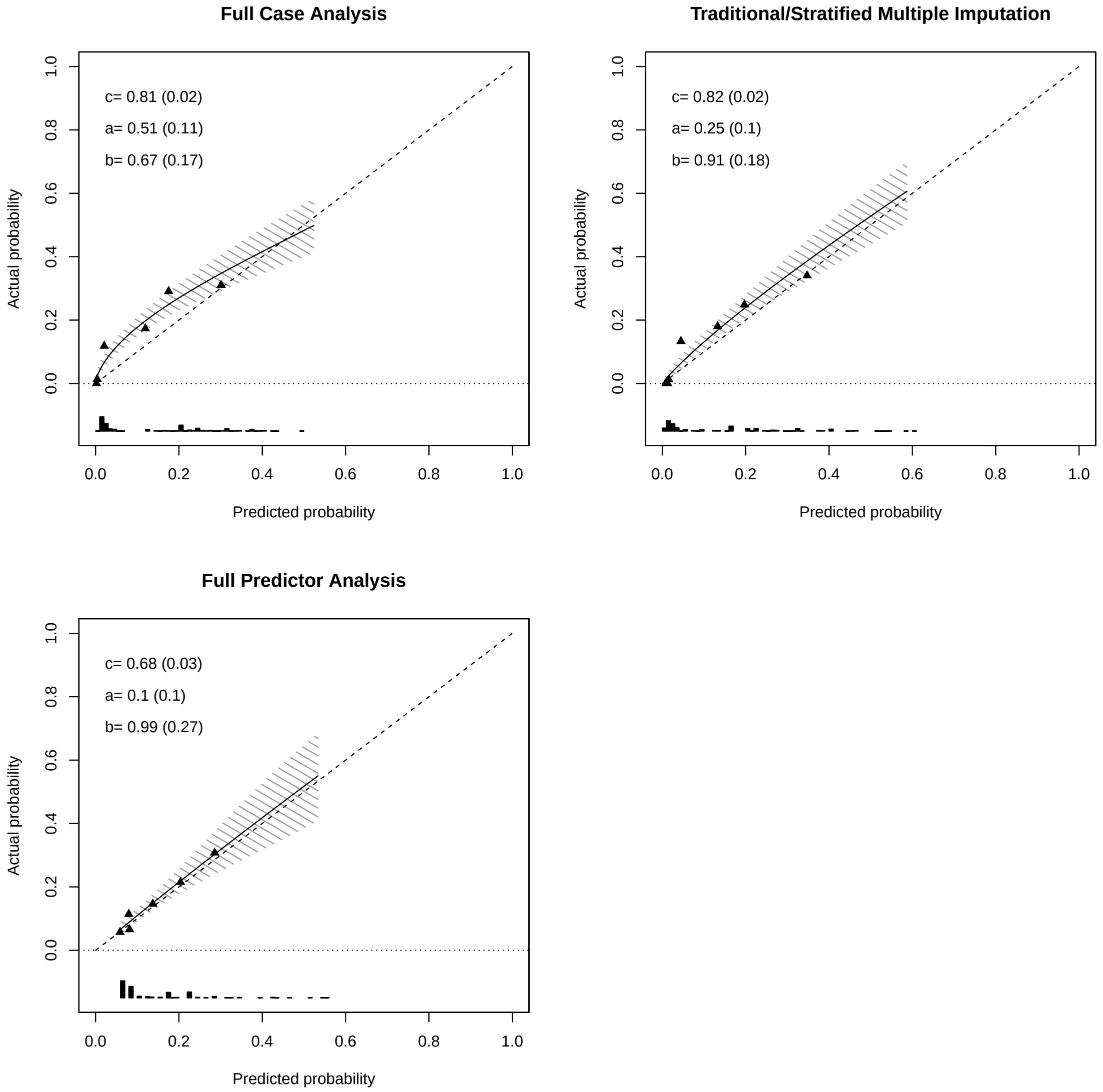


c = concordance statistic, a = calibration-in-the-large, b = calibration slope.
The calibration plots were calculated in the external validation sample.

Full predictor analysis

Results demonstrate that ignoring systematically missing predictors leads to prediction models with a good calibration, but a rather poor discriminative ability. In particular, we found that the c-statistic decreased to 0.68 in scenario 1 and 2. The calibration-in-the-large was -0.25 (scenario 1) and 0.1 (scenario 2), corresponding to a predicted outcome prevalence in the validation sample

Figure 4.2: Calibration plots of estimated models (Scenario 2).



c = concordance statistic, a = calibration-in-the-large, b = calibration slope.
The calibration plots were calculated in the external validation sample.

of 16% and respectively 12%. The calibration slope was reasonable (0.85) in the first scenario, and good (0.99) in the second scenario. Finally, for both scenarios we found a substantial amount of heterogeneity in the intercept term and the predictors *malign*, *surg* and *vein* of the development sample.

Figure 4.3: Example dataset with dummy-coding for SMI

studyID	D_{ddimd1}	D_{ddimd2}	$D_{notraum1}$	dvt	sex	ddimd	notraum	...
3	0	0	0	0	0	NA	1	...
3	0	0	0	1	0	NA	0	...
8	0	0	0	0	0	1	NA	...
9	0	0	1	0	1	1	1	...
12	0	0	0	0	1	NA	NA	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱
2	0	1	0	0	0	1	NA	...
5	0	1	0	0	1	1	1	...
7	0	1	0	0	0	NA	NA	...
7	0	1	0	0	1	NA	NA	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱
4	1	0	1	0	1	1	1	...
4	1	0	1	0	0	1	1	...
6	1	0	0	1	1	1	1	...
10	1	0	0	0	0	NA	NA	...
11	1	0	0	0	0	1	0	...
13	1	0	0	0	0	NA	NA	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱

When applying SMI, an imputation model needs to be specified for each systematically missing predictor variable. Each imputation model uses a dummy variable D to group studies according to the (predicted) prevalence of the missing predictor variable. For instance, when imputing *ddimd* the following groups can be formed that are no longer affected by systematic missingness for any predictor variable: $[3, 8, 9, 12]$, $[2, 5, 7]$ and $[4, 6, 10, 11, 13]$. The imputation model of *ddimd* is then as follows: $ddimd = f(D_{ddimd1}, D_{ddimd2}, dvt, sex, notraum, \ldots)$. Conversely, when imputing *notraum* we could group $[2, 3, 5, 6, 7, 8, 10, 11, 12, 13]$ and $[4, 9]$ and subsequently use the imputation model $notraum = f(D_{notraum1}, dvt, sex, ddimd, \ldots)$. Note that in *mice*, the imputation models can be specified by the investigator in the *predictorMatrix*.

Traditional Multiple imputation

When all datasets are stacked and imputed by treating them as a single study (TMI), resulting prediction models perform similar to FCA. However, because imputation does not discard any datasets from the analyses, model intercept terms are available from additional studies and the calibration-in-the-large may improve. This situation occurred in scenario 1, where the calibration-in-the-large decreased from 0.48 (FCA) to -0.07 (TMI). The relative performance of TMI substantially improved when fewer studies were available in the IPD-MA (scenario 2). In particular, we found that the calibration was no longer adequate when applying FCA (calibration slope = 0.67), but remained nearly optimal when applying TMI (calibration slope = 0.91). Finally, results in Table 4.3 indicate that estimates for τ substantially increased as compared to FCA. For instance, τ_{ddimd} increased from 0.07 (FCA) to 0.50 (TMI), and τ_{malign} increased from 0.03 (FCA) to 0.43 (TMI) in scenario 1.

Stratified Multiple Imputation

The implementation of stratified multiple imputation was only possible in scenario 1 (Table 4.1 and 4.2), and an illustration of the corresponding grouping of studies is depicted in Figure 4.3. For scenario 2, there was only one study with information on *oachst* and SMI therefore collapsed to TMI. For both scenarios, we found that stratified imputation performed similar to traditional imputation and yielded a *c*-statistic of 0.82. For scenario 1, the calibration-in-the-large was 0.11 and the calibration slope was 0.84. The degree of between-study heterogeneity in predictor effects from the development sample was similar as compared to TMI.

DISCUSSION

In this study, we used an empirical example to demonstrate that the presence of systematically missing predictor variables in an IPD-MA can be tackled by relatively simple solutions and standard software packages when developing a novel prediction model. In particular, our results showed that full case analysis does not necessarily compromise model performance, and that imputation strategies are preferred when systematically missing predictors are strong or few studies are at hand. These findings are in line with previous research [45, 131] and underline the value of imputation strategies in prediction modeling studies.

Various aspects need to be addressed to appreciate these results. First, the implementation of full case analysis is the most straightforward but may not always be recommendable. In particular, we found that full case analysis substantially hampers model development when after study exclusion only few studies are left for the analysis. This situation arose in scenario 2 where only one study remained for model estimation. As a consequence, it was no longer possible to assess the between-study heterogeneity of predictor effects, and the prediction model became prone to overfitting. In addition, the calibration-in-the-large of estimated models using full case analysis was substantially than when imputation strategies were used. This is mainly because no appropriate intercept term could be selected when evaluating the prediction model's performance in the validation sample.

Second, we did not evaluate imputation techniques that employ a hierarchical approach. These approaches borrow information across studies whilst accounting for clustering of subjects within subjects. They are likely to yield better performance when covariance structures between the (missing) predictors substantially differ across studies. Unfortunately, hierarchical imputation approaches for dealing with systematically missing predictor variables have received limited attention, and software implementations are currently lacking. As a consequence, researchers necessarily fall back on traditional imputation strategies that assume a common covariance structure across the IPD-MA studies. These traditional strategies may then support a self-fulfilling prophecy when assessing the between-study heterogeneity of predictor effects. Our results, however, demonstrate that traditional (and stratified) imputation strategies can actually inflate estimates of between-study variance. Thus, we did not find any evidence that traditional imputation strategies promote similarity of IPD-MA datasets.

Third, we used a predefined set of predictors to develop a multivariable risk prediction model. It is not clear how variable selection should optimally be applied in an IPD-MA, and approaches for this purpose are currently lacking. In addition, Debray *et al* recently demonstrated that the generalizability of prediction models increases when included predictor effects are homogeneous across studies [68]. Results from our case study, however, indicated substantial between-study heterogeneity in some of the included predictor effects. This may have affected model performance in the independent validation sample. It may therefore be valuable to investigate how predictor selection algorithms may be implemented to identify a useful set of homogeneous predictor-outcome associations.

Fourth, the implementation of stratified multiple imputation may not be feasible when many predictors are systematically missing. This is because predicting the prevalence (or mean) of missing predictors may no longer be possible due to a lack of observed prevalances (or means). Furthermore, predictions for missing prevalences (or means) are only performed once in SMI. Although this facilitates the identification of studies with a similar prevalence (or mean) for a predictor, it also implies that the uncertainty of these predictions are not taken into account. Further research is needed to evaluate how these issues can be addressed.

General conclusions Traditional imputation of systematically missing predictors appears an efficient approach when developing a multivariable risk prediction model. Exclusion of affected studies during model development generally leads to similar model performance if a sufficient amount of studies remains available in the IPD-MA. Finally, ignoring systematically missing predictors leads to poorest model performance, particularly when important predictors are ignored during model development.

ACKNOWLEDGEMENTS

We gratefully acknowledge the following authors for sharing of individual participant data from the deep vein thrombosis (DVT) studies: A.J. Ten Cate-Hoek, R. Oudega, K.G.M. Moons, R.E.G. Schutgens, D.R. Anderson, P.S. Wells, R.A. Kraaijenhagen, D.B. Toll, C. Kearon, J.L. Elf, S.M. Stevens and S.M. Bates.

Part III

Aggregate Data Meta-Analysis

5

Incorporating published univariable associations in diagnostic and prognostic modeling

BMC Medical Research Methodology 2012, **12**(1): 121.

Thomas P. A. Debray
Hendrik Koffijberg
Difei Lu
Yvonne Vergouwe
Ewout W. Steyerberg
Karel G. M. Moons

Abstract

Diagnostic and prognostic literature is overwhelmed with studies reporting univariable predictor-outcome associations. Currently, methods to incorporate such information in the construction of a prediction model are underdeveloped and unfamiliar to many researchers. This article aims to improve upon an adaptation method originally proposed by Greenland (1987) and Steyerberg (2000) to incorporate previously published univariable associations in the construction of a novel prediction model. The proposed method improves upon the variance estimation component by reconfiguring the adaptation process in established theory and making it more robust. Different variants of the proposed method were tested in a simulation study, where performance was measured by comparing estimated associations with their predefined values according to the Mean Squared Error and coverage of the 90% confidence intervals. Results demonstrate that performance of estimated multivariable associations considerably improves for small datasets where external evidence is included. Although the error of estimated associations decreases with increasing amount of individual participant data, it does not disappear completely, even in very large datasets. In conclusion, the proposed method to aggregate previously published univariable associations with individual participant data in the construction of a novel prediction models outperforms established approaches and is especially worthwhile when relatively limited individual participant data are available.

"Invention, it must be humbly admitted, does not consist in creating out of void but out of chaos."

– Mary Shelley

Recent medical literature has shown an increasing interest in clinical prediction models obtained from cross-sectional studies (diagnostic models) as well as case-control, cohort and randomized controlled data (prognostic models) [167, 170, 192, 237, 280]. Such models combine multiple predictors or markers that are independently associated with the presence (in case of diagnosis) or future occurrence (in case of prognosis) of a particular outcome. Typically, logistic regression is used to model these binary outcomes. Alternatively, Cox proportional hazards regression may be applied to account for the time-to-event.

The development of a novel prediction model requires a dataset with a sufficient amount of participants to obtain accurate associations and to make reliable predictions. Also, larger numbers of participants increase the statistical power when selecting predictive subject characteristics to be included in predictive models. Although numerous prediction models are constructed from a single dataset, it is possible to increase the amount of evidence available by incorporating information from the literature.

The availability of individual participant data (IPD) is commonly recommended as gold standard for combining existing information with newly collected data [203, 234]. However, this situation is often unfeasible due to practical constraints [126, 236], for instance when studies were conducted several years ago. Fortunately, numerous papers contain baseline population characteristics from which univariable predictor-outcome associations can be derived. Consequently, these associations represent an appealing source of evidence when developing a novel prediction model [27, 52, 76, 118, 166, 196, 206, 237, 252].

Greenland and Steyerberg have recently proposed adaptation methods to incorporate previously published univariable predictor-outcome associations as prior evidence in a regression analysis [91, 243]. These methods combine the result of a univariable meta-analysis with the results of a univariable and multivariable logistic regression analysis on the IPD. Although these quantitative approaches may considerably improve the quality of a model's regression coefficients and its resulting performance, they are not yet frequently used in practice [207, 219].

Here we present an improved alternative to the methods proposed by Greenland and Steyerberg that aims to further increase the accuracy and precision of the multivariable associations estimated using external evidence. This method improves upon the variance estimation component by reconfiguring the adaptation process in established theory and making it more robust. We present two variants of our method and test their performance in a simulation study. We illustrate the proposed methods' application in a clinical example involving the prediction of peri-operative mortality after elective abdominal aortic aneurysm surgery [246].

METHODS

This method is intended to address the specific situation where IPD have been collected to evaluate the effect of a number of predictors on a dichotomous outcome using logistic regression analysis. Here, univariable and multivariable associations (logistic regression coefficients) are estimated and denoted as β_u and β_m . Particularly, two sources of associations are assumed to be available, namely the IPD of the study at hand (I) and aggregated data from the literature (L). The univariable and multivariable associations estimated in the derivation data are denoted as $\hat{\beta}_{\mathrm{u}|\mathrm{I}}$ and $\hat{\beta}_{\mathrm{m}|\mathrm{I}}$. For the literature, only univariable associations are available ($\hat{\beta}_{\mathrm{u}|\mathrm{L}}$). It is assumed that the study at hand and the studies forming the literature are both random samples from a common underlying patient population.

Previously, Greenland proposed a method to incorporate univariable associations reported in the literature when developing a novel multivariable prediction model from newly collected data [91]. This method attempts to approximate a situation where the individual participant data from all the previously published datasets was available for all the candidate covariates. It uses the calculated change from univariable to multivariable association in the newly collected data and uses this difference to estimate the multivariable association that would have been reported in the previous literature using the IPD from the previous studies:

$$\hat{\beta}_{\mathrm{m}|\mathrm{L}} = \hat{\beta}_{\mathrm{u}|\mathrm{L}} + \left(\hat{\beta}_{\mathrm{m}|\mathrm{I}} - \hat{\beta}_{\mathrm{u}|\mathrm{I}}\right) \tag{5.1}$$

The proposed estimate for the variance of $\hat{\beta}_{\mathrm{m}|\mathrm{L}}$ is given as follows [91, 94].

$$\widehat{\mathrm{Var}}\left(\hat{\beta}_{\mathrm{m}|\mathrm{L}}\right) = \widehat{\mathrm{Var}}\left(\hat{\beta}_{\mathrm{u}|\mathrm{L}}\right) + \left[\widehat{\mathrm{Var}}\left(\hat{\beta}_{\mathrm{m}|\mathrm{I}}\right) - \widehat{\mathrm{Var}}\left(\hat{\beta}_{\mathrm{u}|\mathrm{I}}\right)\right] \tag{5.2}$$

Here, $\hat{\beta}_{\mathrm{u}|\mathrm{L}}$ can be obtained through a meta-analysis involving fixed or random effects, and $\hat{\beta}_{\mathrm{m}|\mathrm{L}}$ is the (asymptotically) unbiased estimate of the multivariable association $\hat{\beta}_\mathrm{m}$. Subsequently, Steyerberg *et al.* extended this method by defining a weight c to reflect inconsistencies and variability in previous research [243]:

$$\hat{\beta}_{\mathrm{m|L}} = \hat{\beta}_{\mathrm{m|I}} + c\left(\hat{\beta}_{\mathrm{u|L}} - \hat{\beta}_{\mathrm{u|I}}\right) \tag{5.3}$$

Previous simulations have however shown that the original unweighted method ($c = 1$ in expression 5.3) has a similar performance.

Concerns and proposed solutions

Although aforementioned formulas are relatively simple to apply, the calculation of $\widehat{\mathrm{Var}}(\hat{\beta}_{\mathrm{m|L}})$ in expression 5.2 clearly contrasts with the theoretical variance component:

$$\begin{aligned}\mathrm{Var}\left(\hat{\beta}_{\mathrm{m|L}}\right) =& \mathrm{Var}\left(\hat{\beta}_{\mathrm{u|L}}\right) + \mathrm{Var}\left(\hat{\beta}_{\mathrm{m|I}}\right) + \mathrm{Var}\left(\hat{\beta}_{\mathrm{u|I}}\right) \\ &+ 2\,\mathrm{Cov}\left(\hat{\beta}_{\mathrm{u|L}}, \hat{\beta}_{\mathrm{m|I}}\right) - 2\,\mathrm{Cov}\left(\hat{\beta}_{\mathrm{m|I}}, \hat{\beta}_{\mathrm{u|I}}\right) \\ &- 2\,\mathrm{Cov}\left(\hat{\beta}_{\mathrm{u|L}}, \hat{\beta}_{\mathrm{u|I}}\right)\end{aligned} \tag{5.4}$$

Although it is possible to assume that estimated associations from the literature and IPD at hand are independent, i.e. $\mathrm{Cov}(\hat{\beta}_{\mathrm{u|L}}, \hat{\beta}_{\mathrm{m|I}}) = \mathrm{Cov}(\hat{\beta}_{\mathrm{u|L}}, \hat{\beta}_{\mathrm{u|I}}) = 0$, the remaining assumption that $\mathrm{Cov}(\hat{\beta}_{\mathrm{m|I}}, \hat{\beta}_{\mathrm{u|I}}) = \mathrm{Var}(\hat{\beta}_{\mathrm{u|I}})$ seems unrealistic. Particularly, this assumption requires that the univariable and multivariable association in the IPD at hand are strongly correlated and neglects $\mathrm{Var}(\hat{\beta}_{\mathrm{m|I}})$, as $\mathrm{Cov}(\hat{\beta}_{\mathrm{m|I}}, \hat{\beta}_{\mathrm{u|I}}) = \rho(\hat{\beta}_{\mathrm{m|I}}, \hat{\beta}_{\mathrm{u|I}})\,\mathrm{Var}(\hat{\beta}_{\mathrm{m|I}})\,\mathrm{Var}(\hat{\beta}_{\mathrm{u|I}})$. Consequently, expression 5.2 may yield biased variance estimates of adapted multivariable associations. Although it is even possible that $\widehat{\mathrm{Var}}(\hat{\beta}_{\mathrm{m|L}})$ becomes negative when $\widehat{\mathrm{Var}}(\hat{\beta}_{\mathrm{m|I}}) < \widehat{\mathrm{Var}}(\hat{\beta}_{\mathrm{u|I}})$, this is unlikely to happen because adjustment of logistic regression coefficients is expected to result in a loss of precision [209].

In order to obtain asymptotically unbiased estimates for $\mathrm{Var}(\hat{\beta}_{\mathrm{m|L}})$, we incorporate the distribution of estimated associations. A pragmatic parametric family for the distribution of associations is the normal distribution, where we assume that $\hat{\beta}_{\mathrm{u|I}} \sim \mathcal{N}(\mu_{\mathrm{u|I}}, \sigma^2_{\mathrm{u|I}})$, $\hat{\beta}_{\mathrm{m|I}} \sim \mathcal{N}(\mu_{\mathrm{m|I}}, \sigma^2_{\mathrm{m|I}})$ and $\hat{\beta}_{\mathrm{u|L}} \sim \mathcal{N}(\mu_{\mathrm{u|L}}, \sigma^2_{\mathrm{u|L}})$. Then, the adaptation from univariable to multivariable association, i.e. $\hat{\beta}_{\mathrm{m|I}} - \hat{\beta}_{\mathrm{u|I}}$ in expression 5.1, is also normally distributed. The distribution of this adaptation is further denoted as $\mathcal{N}\left(\mu_\delta, \sigma_\delta^2\right)$, such that $\hat{\beta}_{\mathrm{m|L}}$ can be estimated by:

$$\hat{\mu}_{\mathrm{u|L}} + \hat{\mu}_\delta \tag{5.5}$$

with a standard error estimate of

$$\sqrt{\hat{\sigma}^2_{\mathrm{u|L}} + \hat{\sigma}^2_\delta} \tag{5.6}$$

The probabilistic adaptation from univariable to multivariable association $\mathcal{N}(\mu_\delta, \sigma_\delta^2)$ can be estimated from the IPD at hand using bootstrap sampling [62]. This procedure applies repeated sampling with replacement of subjects from the derivation dataset. Hence, it allows generating numerous datasets (bootstrap samples) where the adaptation can be estimated. Unfortunately, the bootstrap procedure may become unstable when the effective sample size is small, and yield regression coefficients with extreme values [9, 145, 184]. This, in turn, may strongly affect the quality of estimated adaptations and result in poor estimates of $\beta_{\mathrm{m|L}}$. For this reason, we propose to shrink the adaptation by implementing a Bayesian prior for the univariable and multivariable associations of the IPD at hand. Recently, Gelman *et al.* proposed a weakly default prior distribution that is based on the Cauchy distribution and assumes a probability of 70.48% for associations between -5 and 5. This distribution is less conservative than the uniform prior distribution (which assumes higher probabilities for extreme associations), and yields estimates that make more sense and have predictive performance better than maximum likelihood estimates [87]. The weakly informative prior distribution for generalized linear modeling was recently implemented in R, and is available in the package *arm*.

Finally, the summary of univariable associations from the literature $\mathcal{N}(\mu_{\mathrm{u|L}}, \sigma^2_{\mathrm{u|L}})$ is originally estimated by applying a fixed effects meta-analysis [106, 179]. Because this estimate may be unstable when few studies are available, Steyerberg *et al.* proposed using the univariable associations from the literature (published as $\hat{\beta}_{\mathrm{u|L}}$) and the IPD at hand (estimated as $\hat{\beta}_{\mathrm{u|I}}$) [243]. When the homogeneity assumptions made by the adaptation method are violated, it is possible to assume random effects to further improve the robustness of estimated associations.

Given aforementioned concerns, we propose two variants (Table 5.1) of the adaptation method which we further denote as the *Improved Adaptation Method.* The first variant (*no prior*) decreases the bias of $\widehat{\mathrm{Var}}(\hat{\beta}_{\mathrm{m|L}})$ by effectively removing the unrealistic assumptions about the covariance between univariable and multivariable associations in the IPD at hand. This variant also attempts to reduce the impact of heterogeneity by allowing random effects in the pooling of literature associations. The second variant (*weakly informative prior*) aims to further improve the quality of estimated multivariable associations by implementing a weakly informative prior distribution for estimating the univariable and multivariable associations in the IPD at hand. For this purpose, its logistic regression analyses use independent Cauchy distributions on all regression coefficients, each centered at 0 and with scale parameter 10 for the constant term and 2.5 for all other coefficients. In this manner, estimates for the adaptation from univariable to multivariable association become more robust.

SIMULATION STUDY

We performed a simulation study to assess the quality of estimated multivariable associations. Hereto, we considered the situation in which IPD and literature data are described by two predictors and a dichotomous outcome. Arbitrary values were predefined for the independent association between these predictors and their respective outcome, with $\mathrm{b}_0 = -3.43$, $\mathrm{b}_1 = 1.45$ and $\mathrm{b}_2 = 1.18$ (where we chose $x_1, x_2 \sim \mathcal{N}(0,1)$ and $\rho\,(x_1, x_2) = 0$, i.e. x_1 and x_2 are not correlated) which we further refer to as the reference model. The outcome y for each subject $i = 1, \ldots, N$ is generated as follows, and corresponds to an average incidence of 9%.

$$y = \begin{cases} 1, & \text{if } u < \operatorname{logit}^{-1}(-3.43 + 1.45\,x_1 + 1.18\,x_2) \\ 0, & \text{if } u \geq \operatorname{logit}^{-1}(-3.43 + 1.45\,x_1 + 1.18\,x_2) \end{cases}$$

where $u \sim \mathcal{U}(0,1)$. We applied aforementioned methods (Table 5.1) to update only the multivariable association of the first predictor b_1. In each scenario, data for four literature studies as well as an IPD are generated with different degrees of comparability. For this purpose, we used the reference model (fixed effects) to generate the IPD and source datasets of the univariable associations from the literature. We investigated the impact of sample size by evaluating different choices for N_I (100, 200, 500 and 1 000) and N_L (500 and 2 000). Note that $N_\mathrm{I} = 100$ violates the rule of thumb that logistic models should be used with a minimum of 10 outcome events per predictor variable [184]. We also evaluated the performance for the scenario in which the key assumption of study exchangeability is violated. Hereto, we introduced random variation in b_1 of the reference model when generating data for the literature studies:

$$y = \begin{cases} 1, & \text{if } u < \operatorname{logit}^{-1}(-3.43 + (\mathrm{b}_{1|\mathrm{L}})_j\,x_1 + 1.18\,x_2) \\ 0, & \text{if } u \geq \operatorname{logit}^{-1}(-3.43 + (\mathrm{b}_{1|\mathrm{L}})_j\,x_1 + 1.18\,x_2) \end{cases}$$

Table 5.1: Overview of approaches

		No MA	**GS-AM**	**I-AM** (v1)	**I-AM** (v2)
Step 1	Estimate associations in IPD				
	Implemented	Yes	Yes	Yes	Yes
	Association type	m	u+m	u+m	u+m
	Prior distribution	none	none	none	weakly informative
Step 2	Summarize univariable associations				
	Implemented	No	Yes	Yes	Yes
	Source	-	I+L	I+L	I+L
	Pooling Method	-	random effects	random effects	random effects
Step 3	Estimate adaptation from univariable to multivariable association				
	Implemented	No	Yes	Yes	Yes
	Assumptions	-	(1)+(2)	(1)	(1)
	Estimation procedure	-	analytic	bootstrap	bootstrap
	Prior distributions	-	none	none	weakly informative
Step 4	Apply adaptation to summary estimate from the literature and estimate $\beta_{\mathrm{m}\|\mathrm{L}}$				
	Implemented	No	Yes	Yes	Yes

No MA = No meta-analysis; GS-AM = Greenland/Steyerberg Adaptation Method; I-AM = Improved Adaptation Method (variant 1 & 2). The assumptions about the variance component are as follows: (1) estimated associations in the individual participant data (IPD) are independent from estimated associations in the literature, and (2) $\mathrm{Cov}(\hat{\beta}_{\mathrm{m|I}}, \hat{\beta}_{\mathrm{u|I}}) = \mathrm{Var}(\hat{\beta}_{\mathrm{u|I}})$.

where $u \sim \mathcal{U}(0,1)$ and $(\mathrm{b}_{1|\mathrm{L}})_j \sim \mathcal{N}\left(1.45, \sigma_{\mathrm{h}}^2\right)$ with $j = 1, \ldots, 4$. Consequently, differences in multivariable associations from the literature appear due to sampling variance and heterogeneity across study populations originated from one source of variability (e.g. due to a focus of studies on primary versus secondary care, younger versus older patients etc). Multivariable associations from the IPD at hand remain homogeneous with the study population ($\mathrm{b}_{1|\mathrm{I}} = 1.45$). The scenarios are illustrated in Figure 5.1, which also demonstrates that the sampling process substantially affects the bias and variance of the univariable and multivariable associations.

Finally, the updated multivariable association $\hat{\beta}_1$ obtained with each method is compared with the predefined association b_1 from the reference model. We evaluate the frequentist properties of the estimated associations in terms of the percentage bias (PB) and the Mean Squared Error (MSE) [46], where

$$\mathrm{PB}\left(\hat{\beta}_1\right) = \frac{\overline{\hat{\beta}}_1 - \mathrm{b}_1}{\mathrm{b}_1} \times 100\% \tag{5.7}$$

and

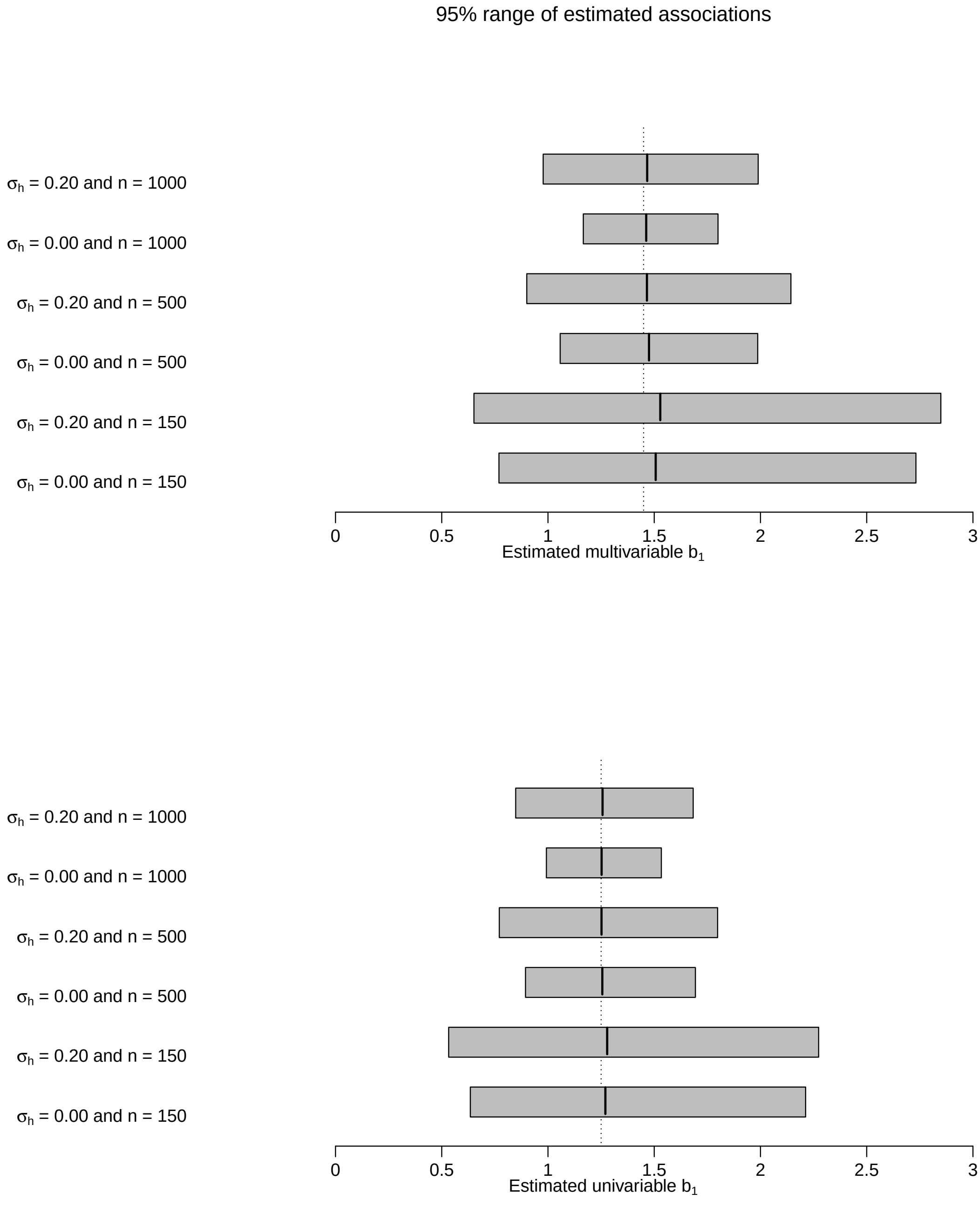
$$\mathrm{MSE}\left(\hat{\beta}_1\right) = \left(\overline{\hat{\beta}}_1 - \mathrm{b}_1\right)^2 + \left(\mathrm{SE}\left(\hat{\beta}_1\right)\right)^2 \tag{5.8}$$

In addition, we calculate the coverage of the 90% confidence intervals (90% CI coverage) and quantify how often invalid variance estimates are obtained (i.e. $\widehat{\mathrm{Var}}(\hat{\beta}_1) < 0$) for the Greenland/Steyerberg adaptation method. We simulated different degrees of available evidence and heterogeneity, and repeated each scenario 500 times. The corresponding results are presented in Table 5.2, 5.3, 5.4 and 5.5. An implementation in R of aforementioned methods is available on request.

No meta-analysis (classical approach)

Results demonstrate that the classical approach to logistic regression, ignoring published univariable evidence from previous studies, considerably overestimates multivariable associations, particularly when the IPD at hand is very small. Although the percentage bias and MSE of $\hat{\beta}_1$ decreases in larger datasets, it does not completely disappear. Similar to previous research, we found that the bias of estimated regression coefficients increases when collinearity occurs and effective sample sizes are small [155]. The coverage of the 90% confidence interval was adequate for all scenarios considered.

Figure 5.1: Comparison of estimated associations



Graphic presentation of multivariable (with true value 1.45) and corresponding univariable (with true value 1.25) associations estimated in an IPD of size n. This dataset is generated according to $x_1, x_2 \sim \mathcal{N}(0, 1)$ with $\Pr(y = 1) = \mathrm{logit}^{-1}(-3.43 + \mathrm{b}_1 x_1 + 1.18 x_2)$ and $\mathrm{b}_1 \sim \mathcal{N}(1.45, \sigma_\mathrm{h}^2)$. Each interval is based on 10 000 repetitions

Table 5.2: Results simulation study

Scenario				**No meta-analysis**			**Greenland/Steyerberg Adaptation Method**			
N_{I}	N_{L}	σ_{h}	$\rho(x_1, x_2)$	PB	MSE	coverage	PB	MSE	coverage	(*)
100	500	0	0	15.07 %	0.613	89.0 %	8.87 %	0.219	89.2 %	8
200	500	0	0	6.58 %	0.186	90.0 %	2.34 %	0.063	90.8 %	1
500	500	0	0	3.65 %	0.061	90.4 %	1.00 %	0.024	90.0 %	0
1 000	500	0	0	1.31 %	0.028	90.2 %	0.84 %	0.014	91.2 %	0
100	500	0	0.50	20.39 %	0.888	91.2 %	5.75 %	0.166	94.4 %	7
200	500	0	0.50	8.22 %	0.226	91.0 %	1.63 %	0.037	93.0 %	0
500	500	0	0.50	1.89 %	0.073	87.6 %	0.45 %	0.019	92.2 %	0
1 000	500	0	0.50	0.88 %	0.031	92.2 %	0.33 %	0.011	93.8 %	0
100	500	0.20	0	10.89 %	0.440	92.4 %	5.17 %	0.140	90.4 %	8
200	500	0.20	0	6.54 %	0.177	92.0 %	3.81 %	0.060	91.6 %	1
500	500	0.20	0	1.23 %	0.049	93.8 %	0.34 %	0.024	92.2 %	0
1 000	500	0.20	0	0.94 %	0.029	89.2 %	0.89 %	0.017	90.4 %	0

In each scenario, an IPD of N_{I} subjects is available and the literature associations are based on 4 studies of N_{L} subjects each. Between-study heterogeneity of literature associations is parameterized by σ_{h}. Correlation between the predictor variables x1 and x2 is indicated by $\rho(x_1, x_2)$. The following statistics of $\hat{\beta}_1$ are presented: percentage bias (PB), Mean Squared Error (MSE) and coverage of the 90% confidence interval (coverage). We also assessed how often the Greenland/Steyerberg adaptation method estimated a negative variance for $\hat{\beta}_1$ (*).

Table 5.3: Results simulation study

Scenario				No meta-analysis			Greenland/Steyerberg Adaptation Method			
N_{I}	N_{L}	σ_{h}	$\rho(x_1, x_2)$	PB	MSE	coverage	PB	MSE	coverage	(*)
100	2 000	0	0	47.95 %	4.9 e+01	93.2 %	37.63 %	4.3 e+01	86.2 %	21
200	2 000	0	0	5.60 %	0.184	90.2 %	3.31 %	0.058	89.8 %	1
500	2 000	0	0	2.36 %	0.064	87.2 %	1.10 %	0.017	89.2 %	0
1 000	2 000	0	0	1.17 %	0.027	90.0 %	0.58 %	0.009	90.2 %	0
100	2 000	0	0.50	20.05 %	0.856	89.6 %	5.68 %	0.139	92.0 %	11
200	2 000	0	0.50	6.99 %	0.206	90.8 %	2.67 %	0.035	92.2 %	1
500	2 000	0	0.50	2.44 %	0.063	90.8 %	0.75 %	0.011	92.8 %	0
1 000	2 000	0	0.50	1.62 %	0.032	89.4 %	0.26 %	0.007	91.6 %	0
100	2 000	0.20	0	16.17 %	0.654	92.6 %	7.67 %	0.201	89.8 %	16
200	2 000	0.20	0	6.63 %	0.177	93.0 %	3.74 %	0.057	89.2 %	1
500	2 000	0.20	0	2.33 %	0.056	92.8 %	1.23 %	0.021	89.6 %	0
1 000	2 000	0.20	0	2.02 %	0.027	92.2 %	1.07 %	0.014	87.4 %	0

In each scenario, an IPD of N_{I} subjects is available and the literature associations are based on 4 studies of N_{L} subjects each. Between-study heterogeneity of literature associations is parameterized by σ_{h}. Correlation between the predictor variables x1 and x2 is indicated by $\rho(x_1, x_2)$. The following statistics of $\hat{\beta}_1$ are presented: percentage bias (PB), Mean Squared Error (MSE) and coverage of the 90% confidence interval (coverage). We also assessed how often the Greenland/Steyerberg adaptation method estimated a negative variance for $\hat{\beta}_1$ (*).

Table 5.4: Results simulation study

Scenario				**Improved Adaptation Method (no prior)**			**Improved Adaptation Method (weakly informative prior)**		
N_{I}	N_{L}	σ_{h}	$\rho(x_1, x_2)$	PB	MSE	coverage	PB	MSE	coverage
100	500	0	0	1.3 e+12 %	1.8 e+23	97.8 %	-1.98 %	0.065	89.6 %
200	500	0	0	18.13 %	3.671	94.4 %	-1.44 %	0.043	89.0 %
500	500	0	0	2.21 %	0.026	91.0 %	-0.54 %	0.021	89.0 %
1 000	500	0	0	1.34 %	0.014	90.6 %	-0.11 %	0.013	90.0 %
100	500	0	0.50	-80.77 %	3.9 e+04	98.4 %	1.41 %	0.048	96.2 %
200	500	0	0.50	4.55 %	0.091	94.2 %	0.32 %	0.031	93.6 %
500	500	0	0.50	0.89 %	0.020	90.8 %	-0.32 %	0.019	91.4 %
1 000	500	0	0.50	0.55 %	0.012	92.8 %	-0.19 %	0.011	93.8 %
100	500	0.20	0	-3.7 e+02 %	5.6 e+04	98.0 %	-4.02 %	0.056	89.8 %
200	500	0.20	0	-11.08 %	0.801	95.6 %	-0.18 %	0.039	91.6 %
500	500	0.20	0	1.53 %	0.026	92.2 %	-1.13 %	0.022	90.8 %
1 000	500	0.20	1.42 %	0.018	90.4 %	0.02 %	0.016	89.8 %	

In each scenario, an IPD of N_{I} subjects is available and the literature associations are based on 4 studies of N_{L} subjects each. Between-study heterogeneity of literature associations is parameterized by σ_{h}. Correlation between the predictor variables x1 and x2 is indicated by $\rho(x_1, x_2)$. The following statistics of $\hat{\beta}_1$ are presented: percentage bias (PB), Mean Squared Error (MSE) and coverage of the 90% confidence interval (coverage).

Table 5.5: Results simulation study

Scenario				Improved Adaptation Method (no prior)			Improved Adaptation Method (weakly informative prior)		
N_{I}	N_{L}	σ_{h}	$\rho(x_1, x_2)$	PB	MSE	coverage	PB	MSE	coverage
100	2 000	0	0	1.6 e+12 %	1.5 e+23	98.2 %	-1.09 %	0.058	89.6 %
200	2 000	0	0	54.36 %	2.1 e+02	94.2 %	-0.12 %	0.036	88.2 %
500	2 000	0	0	2.31 %	0.020	91.4 %	-0.07 %	0.015	88.8 %
1 000	2 000	0	0	1.16 %	0.010	89.2 %	-0.03 %	0.009	87.4 %
100	2 000	0	0.50	3.5 e+12 %	1.3 e+23	98.4 %	1.67 %	0.045	95.4 %
200	2 000	0	0.50	5.94 %	0.120	93.8 %	2.02 %	0.029	92.2 %
500	2 000	0	0.50	1.18 %	0.011	92.0 %	0.45 %	0.010	92.2 %
1 000	2 000	0	0.50	0.45 %	0.007	91.6 %	0.02 %	0.007	91.4 %
100	2 000	0.20	0	1.5 e+03 %	3.9 e+04	98.2 %	-2.66 %	0.046	91.0 %
200	2 000	0.20	0	13.89 %	0.754	94.8 %	0.26 %	0.037	88.8 %
500	2 000	0.20	0	2.46 %	0.023	89.4 %	-0.08 %	0.019	88.6 %
1 000	2 000	0.20	0	1.62 %	0.015	86.6 %	0.37 %	0.013	85.8 %

In each scenario, an IPD of N_{I} subjects is available and the literature associations are based on 4 studies of N_{L} subjects each. Between-study heterogeneity of literature associations is parameterized by σ_{h}. Correlation between the predictor variables x1 and x2 is indicated by $\rho(x_1, x_2)$. The following statistics of $\hat{\beta}_1$ are presented: percentage bias (PB), Mean Squared Error (MSE) and coverage of the 90% confidence interval (coverage).

Greenland/Steyerberg adaptation method

The multivariable associations estimated with the Greenland/Steyerberg Adaptation method were far more accurate than those estimated with the classical approach, especially when little actual data were available. Estimated associations remain, however, too extreme compared to the associations from the reference model. The coverage of the 90% confidence interval was good for most scenarios, although we observed over-coverage when collinearity was present, and under-coverage when the literature studies were very large and heterogeneous. Unfortunately, we also noticed that some estimates for $\mathrm{Var}(\hat{\beta}_{\mathrm{m|L}})$ were negative when IPDs were small, and particularly when the literature studies were large (such that $\mathrm{Var}(\hat{\beta}_{\mathrm{u|L}})$ becomes negligible). Finally, the presence of heterogeneity in the literature associations did not influence the accuracy of estimated associations. This finding can however be explained by the fact that heterogeneity was only introduced in the spread of the literature associations.

Improved adaptation method (no prior)

When no shrinkage was applied for the associations of the IPD at hand, estimated multivariable associations had the largest error, particularly when few data were available. Regression coefficients in bootstrap samples were often non-identifiable (results not shown), resulting in unstable estimates and over-coverage of multivariable regression coefficients. When the size of the IPD at hand increased, this approach performed similar to the improved adaptation method with a weakly informative default prior and the approach proposed by Greenland and Steyerberg.

Improved adaptation method (weakly informative prior)

Results demonstrate that estimated associations were most accurate when a weakly informative prior was used during estimation of the adaptation. Even when the rule of thumb that logistic models should be used with a minimum of 10 outcome events per predictor variable is clearly violated, this approach yielded superior estimates of b_1 that were very similar to estimates obtained from large amounts of IPD. Finally, we observed over-coverage of the 90% confidence interval when collinearity was present, and under-coverage when the literature studies were very large and heterogeneous with the IPD at hand.

APPLICATION

We applied the methods discussed above to an empirical dataset of the prediction of peri-operative mortality (in-hospital or within 30 days) after elective abdominal aortic aneurysm surgery [246]. The study was exempted from ethical approval under Dutch law. Individual participant data were available for 238 subjects (including 18 deaths) and consisted of the predictors age, gender, cardiac co-morbidity (history of myocardial infarction, congestive heart failure, and ischemia on the ECG), pulmonary co-morbidity (COPD, emphysema or dyspnea) and renal co-morbidity (elevated pre-operateive creatinine level). Univariable literature data were available from 15 studies with 15 821 subjects including 1 153 deaths in total [65]. We incorporated the univariable evidence from the literature data to estimate the multivariable associations of four of these predictors. Similar to the simulation study, we applied standard logistic regression modeling (no meta-analysis), the Greenland/Steyerberg Adaptation method and the improved adaptation method. The corresponding results are presented in Table 5.6.

No meta-analysis (classical approach)

The poor quality of estimated associations can be illustrated by their substantial variance. The predictor ‘Female Sex’ is a good example, since the 90% confidence interval of its multivariable association was estimated as $[-1.30, 2.00]$.

Greenland/Steyerberg adaptation method

The Greenland/Steyerberg Adaptation method yielded notably different multivariable associations. For instance, whereas the classical approach estimated a multivariable association of 0.74 ($\mathrm{OR}_{\mathrm{adj}} = 2.10$) for the predictor ‘History of MI’, this estimate was shrunk to 0.26 ($\mathrm{OR}_{\mathrm{adj}} = 1.20$) by the adaptation method. Here, the considerable difference in univariable associations between the individual dataset and the literature is a major cause of shrinkage. Finally, the variance of multivariable associations was much smaller when published evidence from the literature was incorporated.

Improved adaptation method (no prior)

We noticed a substantial increase in the variance of estimated adaptations due to the occurrence of non-identifiability in some of the bootstrap samples. These findings illustrate the need for a

Table 5.6: Calculation of Adapted Associations in the Application

	Sex	MI	CHF	ischemia
Adaptation				
Greenland/Steyerberg	0.02; 0.13	-0.76; 0.07	-0.74; 0.05	-0.72; 0.08
Improved (no prior)	0.04; 0.39	-0.69; 0.15	-0.67; 0.16	-0.72; 0.41
Improved *	0.05; 0.12	-0.65; 0.07	-0.63, 0.05	-0.67; 0.11
Univariable association				
Greenland/Steyerberg	0.35; 0.03	1.02; 0.07	1.58; 0.12	1.52; 0.10
Improved (no prior)	0.35; 0.03	1.02; 0.07	1.58; 0.12	1.52; 0.10
Improved *	0.34; 0.03	1.00; 0.07	1.52; 0.11	1.48; 0.09
Multivariable association				
No meta-analysis	0.30; 0.75	0.74; 0.32	1.04; 0.35	0.99; 0.38
Greenland/Steyerberg	0.36; 0.16	0.26; 0.14	0.84; 0.17	0.80; 0.18
Improved (no prior)	0.38; 0.42	0.33; 0.22	0.91; 0.28	0.80; 0.51
Improved *	0.39; 0.15	0.35; 0.14	0.90; 0.16	0.81; 0.21

Illustration of the adaptation methods for four independent associations for predicting perioperative mortality (in-hospital or within 30 days) after elective abdominal aortic aneurysm surgery. The following estimates are presented: adaptation from univariable to multivariable association (with mean $\hat{\mu}_\delta$ and variance $\hat{\sigma}^2_\delta$), summary of univariable associations from the literature and IPD (with mean $\hat{\mu}_\mathrm{u}$ and variance $\hat{\sigma}^2_\mathrm{u}$) and adapted multivariable association (with mean $\hat{\mu}_\mathrm{m}$ and variance $\hat{\sigma}^2_\mathrm{m}$). Multivariable estimates were obtained through independent adaptation of the corresponding univariable associations, and are adjusted for the following variables: female sex (Sex), age in decades, history of myocardial infarction (MI), congestive heart failure (CHF), ischemia on electrocardiogram, renal co-morbidity and lung co-morbidity.
* This improved adaptation method employs a weakly informative prior

prior distribution that shrinks the associations of the individual dataset and thereby robustifies the adaptation.

Improved adaptation method (weakly informative prior)

Multivariable associations were similar but not equal to those estimated with the Greenland/Steyerberg Adaptation method. For instance, the multivariable association of the predictor ‘History of MI’ was shrunk to a lesser extent by both variants of the improved adaptation method. Furthermore, the variance of estimated adaptations and multivariable associations decreased considerably by implementing a weakly informative prior distribution.

Chapter 5

DISCUSSION

The incorporation of previously published univariable associations from single diagnostic or prognostic test, predictor or marker studies, into the development of a novel prediction model is both feasible and beneficial. A simple method for this purpose was proposed by Greenland and Steyerberg using the change from univariable to multivariable association observed in the IPD to adapt the univariable associations from the literature. We present an improved adaptation method and demonstrate its additional value in a simulation study. Particularly when the individual dataset is relatively small, this method estimates multivariable associations with a smaller MSE, and obtains better coverage of their 90% confidence intervals. Major performance gain is obtained by shrinking the associations from the individual dataset when calculating the adaptation. When no shrinkage was applied (no prior), non-identifiability occurred in some of the bootstrap samples and estimated adaptations were no longer normally distributed. Since we know that extreme associations are very rare in medical sciences, the use of a weakly informative default prior is justified [87], resulting in improved accuracy and precision of the adaptation and hence also the multivariable associations under study.

Several issues must be considered when evaluating these findings: Firstly, performance was evaluated here through the estimation of an association in a small prediction model. Our method may perform better in larger models where correlations between univariable and multivariable associations may be less strong, but this remains untested. Secondly, advanced Bayesian approaches for summarizing the evidence from the literature were not considered. Although these approaches might further improve the accuracy and coverage of multivariable associations, they are less readily compared with meta-analytical models and require more modeling expertise.

Third, the assumption that studies from the literature are exchangeable with the data at hand might not always hold. Simulations showed an under-coverage of the estimated 90% confidence interval when comparability between the considered associations was low, indicating that incorporating strongly heterogeneous evidence from the literature into prediction modeling remains problematic. In those scenarios, the change from univariable to multivariable association in the IPD at hand may no longer be representative for associations from the literature. Evidently, the incorporation of strongly heterogeneous evidence (for example indicated by the I^2 statistic) from the literature into the development of a novel prediction model remains questionable [92, 113]. In addition, aggregating published results may not be desirable if publication bias is present or suspected. Fortunately, the use of random effects when summarizing the associations from the literature seems to counter this problem to some extent.

Fourth, we did not consider the situation in which multivariable (rather than univariable) associations are available from the literature. Although their incorporation may be difficult due to the diversity of considered predictors, it could further improve the quality of estimated associations. The synthesis process of associations from the literature should then account for differences in model specification and included associations. Future research will investigate how these challenges can be assessed [66].

Finally, our simulation study only evaluated the performance of estimated multivariable predictor-outcome associations. Although Steyerberg *et al.* showed that improved estimates may increase the quality of the prediction model [243], this relation was not assessed here. It is possible that all adaptation methods perform similar in a prediction task. However, we showed that the Improved Adaptation Method with a weakly informative prior may further reduce the bias of multivariable associations when datasets are small. It may be clear that for strong predictors, this improvement may have a meaningful impact when making predictions. Additional research is needed to evaluate the extent to which improved predictor-outcome associations result in an improved model performance.

CONCLUSIONS

Our study demonstrates that the MSE in multivariable associations of a novel prediction model is largest when external evidence, in this case previously published univariable predictor-outcome associations, is ignored. Although this error decreases with increasing amount of IPD, it does not disappear completely, even in very large datasets. Therefore, it is valuable to incorporate any existing univariable evidence from the literature unless this evidence is strongly heterogeneous. Even when the individual dataset is relatively large compared to the literature, the proposed method will still result in an estimate closer to the underlying multivariable association than the standard method ignoring the literature. The improved and original adaptation methods are robust approaches for this purpose. Whereas the latter method is simpler to apply, the former is more vigorous in small datasets and provides the most stable estimates.

6

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

Statistics in Medicine 2012, **31**(23): 2697–2712.

Thomas P. A. Debray
Hendrik Koffijberg
Yvonne Vergouwe
Karel G. M. Moons
Ewout W. Steyerberg

Abstract

During 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 account for within- and between-study covariance. These approaches are applied to a collection of 15 datasets of patients with Traumatic Brain Injury, and to 5 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 to 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.

Chapter 6

"The multitude of books is making us ignorant."

– Voltaire

It is well known that many prediction models do not generalize well across patient populations [30, 136, 167, 237, 238, 259]. This quandary may occur, e.g., when prediction models are developed from small data sets, when too many predictors were studied compared to 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 due to, for example, cost constraints, ethical considerations or inclusion problems.

It is remarkable that despite the scarcity of individual participant data, 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 [11, 159], over 25 for predicting long-term outcome in neurotrauma patients [186], 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 [70]. More elaborate approaches, e.g. for synthesizing the accuracy of diagnostic tests [193], have also recently emerged but remain largely lacking in prediction research, despite the fact that the potential gains are arguably even greater [109]. The closest existing equivalent techniques focus upon updating of existing prediction models that are being applied to a different setting [167, 237, 239, 243, 272]. Approaches for using prior knowledge in prediction research are underdeveloped [109]. Some published approaches rely on evidence that is typically not published, such as covariance matrices or regression coefficients, or lack a formal statistical foundation [25, 63].

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 individual participant data 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 [22, 112]. 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.* [207] and explore three approaches to aggregate the published prediction models (with similar predictors) with individual participant data (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 [152, 248]. In addition, we illustrate their application in a genuine example involving the prediction of Deep Vein Thrombosis (DVT).

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 Aggregate Data (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.* [207], 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 re-estimate 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:

$$\begin{aligned} \Pr(y_s = 1) &= \text{logit}^{-1}(\beta_{0,\text{adj}} + \text{offset}) \\ \text{offset} &= \hat{\beta}_{1,\text{UPD}} x_{s1} + \ldots + \hat{\beta}_{K,\text{UPD}} x_{sK} \end{aligned} \tag{6.1}$$

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:

$$\widehat{\text{Cov}}(\hat{\beta}_{0,\text{adj}}, \hat{\beta}_{i,\text{UPD}}) = \frac{\hat{\sigma}_{0,\text{adj}}}{\hat{\sigma}_{0,\text{UPD}}} \widehat{\text{Cov}}(\hat{\beta}_{0,\text{UPD}}, \hat{\beta}_{i,\text{UPD}}) \text{ where } i = 1, \ldots, K \tag{6.2}$$

All approaches were implemented in R 2.14.1 [190]. The corresponding source code is available on request.

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 (WLS) 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 $w_{ij} = 1/(\hat{\sigma}^2_{ij} + \tau^2_j)$ with τ^2_j 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 below.

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 [70, 130]. 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- 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* [25].

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 [254] and bivariate random effects model [271] for this purpose can be generalized as follows:

$$\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}_l \sim \mathrm{MVN}^{K+1}\left(\mu_{\mathrm{re}}, (\Sigma_{\mathrm{re}})_l\right) \tag{6.3}$$

with

$$(\Sigma_{\mathrm{re}})_l = \Sigma_{\mathrm{bs}} + \Sigma_l \tag{6.4}$$

and

$$\Sigma_{\text{bs}} = \begin{pmatrix} \tau_0^2 & \tau_{01} & \dots & \tau_{0K} \\ \tau_{01} & \tau_1^2 & \dots & \tau_{1K} \\ \vdots & \vdots & \ddots & \vdots \\ \tau_{0K} & \tau_{1K} & \dots & \tau_K^2 \end{pmatrix} \tag{6.5}$$

and

$$\Sigma_l = \begin{pmatrix} \sigma_0^2 & \text{Cov}\left(\beta_0, \beta_1\right) & \dots & \text{Cov}\left(\beta_0, \beta_K\right) \\ \text{Cov}\left(\beta_0, \beta_1\right) & \sigma_1^2 & \dots & \text{Cov}\left(\beta_1, \beta_K\right) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}\left(\beta_0, \beta_K\right) & \text{Cov}\left(\beta_1, \beta_K\right) & \dots & \sigma_K^2 \end{pmatrix}_l \tag{6.6}$$

In the expressions above, 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, i.e. $l = \{1, \dots, M, \text{IPD}\}$.

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 $\mu_{\text{UPD}} = \mu_{\text{re}}$ and covariance matrix $\Sigma_{\text{UPD}} = \left(\sum_{l=1}^{M+1} \left(\Sigma_{\text{re}}\right)_l^{-1}\right)^{-1}$. Their corresponding log-likelihood is given by $\ell\left(\mu_{\text{re}}, \Sigma_{\text{bs}}\right) = \sum \ell_l\left(\mu_{\text{re}}, \Sigma_{\text{bs}}\right)$ where $\ell_l\left(\mu_{\text{re}}, \Sigma_{\text{bs}}\right) = \ln\left(\Pr\left(\beta_{0l}, \dots, \beta_{Kl} | \mu_{\text{re}}, \left(\Sigma_{\text{re}}\right)_l\right)\right)$ and $\Pr\left(\beta_{0l}, \dots, \beta_{Kl} | \mu_{\text{re}}, \left(\Sigma_{\text{re}}\right)_l\right) \sim \mathcal{N}^{K+1}\left(\mu_{\text{re}}, \left(\Sigma_{\text{re}}\right)_l\right)$. 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 semi-definite.

Although Σ_l is fully defined for the individual participant data, 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 $\hat{\Sigma}_{\text{IPD}}$, according to

$$\hat{\Sigma}_{\phi\psi l} = \widehat{\text{Cov}}(\hat{\beta}_{\phi l}, \hat{\beta}_{\psi l}) = \frac{\widehat{\text{Cov}}(\hat{\beta}_{\phi,\text{IPD}}, \hat{\beta}_{\psi,\text{IPD}}) \hat{\sigma}_{\phi l}\, \hat{\sigma}_{\psi l}}{\hat{\sigma}_{\phi,\text{IPD}}\, \hat{\sigma}_{\psi,\text{IPD}}} \tag{6.7}$$

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 $\hat{\Sigma}_l = \text{diag}(\hat{\sigma}^2_{0l}, \hat{\sigma}^2_{1l}, \ldots, \hat{\sigma}^2_{Kl})$. 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 amongst 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 [25, 127]. 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 [127].

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 univariate meta-analysis. 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 $\tau^2_1 = \tau_{0,1} = \tau_{1,2} = \ldots = \tau_{1,K} = 0$. 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 [204]. 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.

Bayesian Inference

The approaches for performing a univariate and multivariate meta-analysis 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 multivariate meta-analysis approach previously described:

$$\mu_{\text{PRIOR}} = \mu_{\text{re}} \tag{6.8}$$

$$\Sigma_{\text{PRIOR}} = \left(\sum_{j=1}^{M} \left(\Sigma_{\text{re}}\right)_j^{-1} \right)^{-1} \tag{6.9}$$

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 $\Pr\left(\beta_{0,\text{IPD}}, \ldots, \beta_{K,\text{IPD}} | \mu_{\text{IPD}}, \Sigma_{\text{IPD}}\right) \sim \mathcal{N}^{K+1}\left(\mu_{\text{IPD}}, \Sigma_{\text{IPD}}\right)$. We propose to construct a conjugate prior distribution for μ_{IPD} with $\Pr\left(\mu_{\text{IPD}}\right) \sim \mathcal{N}^{K+1}\left(\mu_{\text{PRIOR}}, \Sigma_{\text{PRIOR}}\right)$ such that the posterior density $\Pr\left(\mu_{\text{IPD}} | \beta_{0,\text{IPD}}, \ldots, \beta_{k,\text{IPD}}, \Sigma_{\text{IPD}}\right) \sim \mathcal{N}^{K+1}\left(\mu_{\text{POST}}, \Sigma_{\text{POST}}\right)$ can be determined analytically:

$$\mu_{\text{UPD}} = \left(\Sigma_{\text{PRIOR}}^{-1} + \Sigma_{\text{IPD}}^{-1}\right)^{-1} \left(\Sigma_{\text{PRIOR}}^{-1}\, \mu_{\text{PRIOR}} + \Sigma_{\text{IPD}}^{-1}\, \mu_{\text{IPD}}\right) \tag{6.10}$$

$$\Sigma_{\text{UPD}} = \left(\Sigma_{\text{PRIOR}}^{-1} + \Sigma_{\text{IPD}}^{-1}\right)^{-1} \tag{6.11}$$

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.

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 Traumatic Brain Injury (TBI) patients (Table 6.1). TBI is a leading cause of death and disability worldwide with a substantial economic burden [122, 147]. It is difficult to establish a reliable prognosis on admission [134]. This requires the consideration of multiple and easily accessible risk factors in multivariable prognostic models [4, 175, 237, 248]. Many prognostic models with admission data are readily available from the literature [175]. However, most models were developed on relatively small sample sizes originating from a single center or region and lack external validation [175, 186]. Therefore, their aggregation might improve the generalization of novel prognostic models.

Table 6.1: Datasets of IMPACT database

Study Code	**Study Year**	**Study Type**	**Patients**
TCDB	1984 – 1987	Obs.	603
UK4	1986 – 1988	Obs.	791
HIT I	1987 – 1989	RCT	350
HIT II	1989 – 1991	RCT	819
TIUS	1991 – 1994	RCT	1 041
TINT	1992 – 1994	RCT	1 118
PEGSOD	1993 – 1995	RCT	1 510
SLIN	1994 – 1996	RCT	409
EBIC	1995	Obs.	822
SKB	1996	RCT	126
SAPHIR	1995 – 1997	RCT	919
CERESTAT	1996 – 1997	RCT	517
NABIS	1994 – 1998	RCT	385
APOE	1996 – 1999	Obs.	756
PHARMOS	2001 – 2004	RCT	856

RCT = Randomized Controlled Trial; Obs = Observational Study

Application Setup

To test the potential value of our approaches we used 15 series of individual participant data collected in the International Mission for Prognosis and Analysis of Clinical Trials in TBI (IMPACT) project (Table A.2 in the Appendix) [21, 80, 152–154, 172–174, 256, 294]. 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 6.2 and 6.3) [175, 248]. In this manner, we aimed to simulate scenarios in which a common set of core predictors is available and can be aggregated with individual participant data. 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 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

Table 6.2: Overview of estimated logistic regression coefficients in the IMPACT data.

Characteristics	**Coding**	**Regression coefficient**
Baseline risk		β_0
Age, years		β_1
Motor Score	Localizes/obeys	Ref.
	None	β_2
	Extension	β_3
	Abnormal flexion	β_4
	Normal flexion	β_5
	Untestable/missing	β_6
Pupillary reactivity	Both pupils reacted	Ref.
	One pupil reacted	β_7
	No pupil reacted	β_8

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 [39, 99].

Practical Example

As an illustration, we used the HIT I study [21] as IPD, the HIT II study [256] as validation data, and the prediction models of the remaining studies as previously published evidence (Table 6.4 and 6.5). 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 [113]. 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 due to 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:

Table 6.3: Estimated regression coefficients (and standard error) from the IMPACT data.

	Logistic regr. coefficients for favorable vs. unfavorable outcome after 6 Months TBI							
	TINT	TIUS	SLIN	SAPHIR	PEGSOD	HIT I	UK4	TCDB
$\hat{\beta}_0$	-2.5 (0.2)	-3.1 (0.3)	-2.1 (0.3)	-2.4 (0.2)	-2.8 (0.2)	-2.7 (0.5)	-2.1 (0.3)	-2.2 (0.3)
$\hat{\beta}_1$	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.1 (0.0)
$\hat{\beta}_2$	1.5 (1.0)	1.4 (0.8)	NA	0.7 (0.2)	1.5 (0.2)	1.4 (0.4)	1.3 (0.4)	2.0 (0.4)
$\hat{\beta}_3$	1.8 (0.2)	1.9 (0.3)	1.7 (0.4)	1.5 (0.3)	2.6 (0.3)	2.5 (0.5)	1.7 (0.4)	2.1 (0.4)
$\hat{\beta}_4$	1.1 (0.2)	1.6 (0.2)	0.6 (0.3)	0.5 (0.2)	1.5 (0.2)	2.0 (0.5)	1.2 (0.5)	0.7 (0.3)
$\hat{\beta}_5$	0.5 (0.2)	0.8 (0.2)	0.3 (0.3)	0.2 (0.2)	0.8 (0.2)	0.8 (0.4)	0.4 (0.3)	0.5 (0.3)
$\hat{\beta}_6$	NA	NA	NA	0.3 (1.2)	NA	1.1 (0.8)	0.9 (0.2)	-0.1 (0.5)
$\hat{\beta}_7$	0.8 (0.2)	0.3 (0.2)	1.1 (0.3)	1.2 (0.2)	0.5 (0.2)	0.4 (0.4)	0.8 (0.3)	0.7 (0.4)
$\hat{\beta}_8$	1.3 (0.2)	1.3 (0.2)	2.1 (0.8)	NA	1.1 (0.1)	2.2 (0.4)	2.0 (0.3)	1.5 (0.3)
	SKB	EBIC	HIT II	NABIS	C-STAT	PHARMOS	APOE	
$\hat{\beta}_0$	-1.8 (0.7)	-3.1 (0.3)	-2.7 (0.3)	-2.1 (0.4)	-2.5 (0.4)	-1.5 (0.2)	-3.2 (0.3)	
$\hat{\beta}_1$	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	
$\hat{\beta}_2$	0.6 (0.6)	1.6 (0.3)	1.1 (0.2)	1.0 (0.3)	0.9 (0.4)	0.5 (0.3)	1.3 (1.2)	
$\hat{\beta}_3$	0.6 (0.7)	1.9 (0.4)	2.1 (0.4)	1.7 (0.4)	1.5 (0.3)	1.3 (0.3)	NA	
$\hat{\beta}_4$	1.3 (0.8)	1.5 (0.4)	1.6 (0.3)	1.8 (0.4)	1.1 (0.3)	1.0 (0.2)	NA	
$\hat{\beta}_5$	-0.2 (0.7)	1.3 (0.3)	0.5 (0.3)	0.8 (0.3)	0.1 (0.3)	0.6 (0.2)	1.3 (0.6)	
$\hat{\beta}_6$	-0.7 (0.7)	1.1 (0.3)	1.0 (0.3)	0.8 (0.7)	NA	0.5 (0.2)	1.2 (0.2)	
$\hat{\beta}_7$	1.1 (0.5)	1.0 (0.3)	0.4 (0.2)	1.0 (0.4)	1.5 (0.3)	0.5 (0.2)	0.9 (0.4)	
$\hat{\beta}_8$	NA	1.4 (0.2)	1.3 (0.2)	1.2 (0.3)	1.9 (0.3)	0.5 (0.4)	2.0 (0.4)	

Note: NA = Not Available

- Standard Logistic Regression (SLR) modeling ignoring the literature studies
- Full IPD modeling (FULL) is a standard logistic regression analysis using all available IPD datasets (except for the validation study). The resulting model is used as "gold standard" for comparing the aggregated models.
- Univariate meta-analysis (UMA)
- Multivariate meta-analysis (MMA)
- Bayesian Inference (BI)

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 above, 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 6.4, 6.5 and [66]) 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 to 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 due 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.

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 5 previously published prediction models [81, 86, 182, 251, 281, 282] with one IPD set, and evaluated different strategies for coping with missing predictor values and within-study covariance. We

Table 6.4: Estimated regression coefficients (and standard error) in the TBI example.

	SLR	**FULL**	**UMA** [†]	**MMA** [†]	**BI** [†]
$\hat{\beta}_0$	-2.66 (0.47)	-2.52 (0.07)	-2.67 (0.12)	-2.67 (0.12)	-2.65 (0.12)
$\hat{\beta}_1$	0.03 (0.01)	0.04 (0.00)	0.04 (0.00)	0.04 (0.00)	0.04 (0.00)
$\hat{\beta}_2$	1.36 (0.37)	1.22 (0.07)	1.20 (0.13)	1.21 (0.10)	1.19 (0.11)
$\hat{\beta}_3$	2.53 (0.54)	1.88 (0.08)	1.81 (0.12)	1.81 (0.09)	1.83 (0.09)
$\hat{\beta}_4$	1.95 (0.47)	1.21 (0.07)	1.17 (0.12)	1.17 (0.10)	1.19 (0.09)
$\hat{\beta}_5$	0.80 (0.42)	0.60 (0.06)	0.60 (0.09)	0.60 (0.07)	0.59 (0.07)
$\hat{\beta}_6$	1.08 (0.77)	0.98 (0.08)	0.82 (0.13)	0.81 (0.11)	0.81 (0.11)
$\hat{\beta}_7$	0.42 (0.35)	0.80 (0.06)	0.83 (0.10)	0.83 (0.07)	0.81 (0.07)
$\hat{\beta}_8$	2.15 (0.42)	1.48 (0.06)	1.46 (0.12)	1.44 (0.12)	1.51 (0.12)
AUC	0.745 (0.017)	0.749 (0.017)	0.749 (0.017)	0.749 (0.017)	0.749 (0.017)
BS	0.206 (0.008)	0.207 (0.007)	0.203 (0.007)	0.203 (0.007)	0.203 (0.007)

In this example, 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. Missing within-study covariance is restricted to zero. 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 BS were estimated according to $\text{sd}[(p_s - o_s)^2]/\sqrt{N}$.

[†] Uninformative regression coefficients are used for missing estimates in the literature models ($\hat{\beta}_\phi = 0$ with $\hat{\sigma}^2_\phi = 100$)

Table 6.5: Estimated regression coefficients (and standard error) in the TBI example.

	SLR	**FULL**	**UMA** ‡	**MMA** ‡	**BI** ‡
$\hat{\beta}_0$	-2.66 (0.47)	-2.52 (0.07)	-2.67 (0.12)	-2.67 (0.12)	-2.65 (0.12)
$\hat{\beta}_1$	0.03 (0.01)	0.04 (0.00)	0.04 (0.00)	0.04 (0.00)	0.04 (0.00)
$\hat{\beta}_2$	1.36 (0.37)	1.22 (0.07)	1.20 (0.13)	1.21 (0.10)	1.19 (0.11)
$\hat{\beta}_3$	2.53 (0.54)	1.88 (0.08)	1.81 (0.12)	1.81 (0.09)	1.83 (0.09)
$\hat{\beta}_4$	1.95 (0.47)	1.21 (0.07)	1.17 (0.12)	1.17 (0.10)	1.21 (0.08)
$\hat{\beta}_5$	0.80 (0.42)	0.60 (0.06)	0.60 (0.09)	0.60 (0.07)	0.59 (0.07)
$\hat{\beta}_6$	1.08 (0.77)	0.98 (0.08)	0.81 (0.13)	0.81 (0.11)	0.81 (0.11)
$\hat{\beta}_7$	0.42 (0.35)	0.80 (0.06)	0.83 (0.10)	0.83 (0.07)	0.79 (0.10)
$\hat{\beta}_8$	2.15 (0.42)	1.48 (0.06)	1.46 (0.12)	1.44 (0.12)	1.47 (0.08)
AUC	0.745 (0.017)	0.749 (0.017)	0.749 (0.017)	0.749 (0.017)	0.749 (0.017)
BS	0.206 (0.008)	0.207 (0.007)	0.203 (0.007)	0.203 (0.007)	0.202 (0.007)

In this example, 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. Missing within-study covariance is restricted to zero. 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 BS were estimated according to $\text{sd}[(p_s - o_s)^2]/\sqrt{N}$.

‡ Mean imputation for missing estimates in the literature models (with $\hat{\sigma}^2_\phi = \sum_{j=1}^{M} \hat{\sigma}^2_{\phi j}$)

Table 6.6: Performance of aggregated prediction models, expressed by means of the Area under the Receiver Operator Characteristic curve (AUC) and the Brier Score (BS).

	UK4			
	$N_{\text{IPD}} = 500\,(N_{\text{VAL}} = 291)$		$N_{\text{IPD}} = 200\,(N_{\text{VAL}} = 591)$	
	AUC (SE)	BS (SE)	AUC (SE)	BS (SE)
SLR	0.813 (0.022)	0.165 (0.010)	0.801 (0.011)	0.172 (0.006)
FULL	0.822 (0.020)	0.174 (0.009)	0.821 (0.008)	0.176 (0.003
UMA	0.821 (0.020)	0.162 (0.009)	0.820 (0.008)	0.164 (0.005)
MMA	0.820 (0.020)	0.162 (0.009)	0.820 (0.008)	0.164 (0.005)
BI	0.820 (0.020)	0.162 (0.009)	0.820 (0.008)	0.164 (0.005)
	HIT II			
	$N_{\text{IPD}} = 500\,(N_{\text{VAL}} = 319)$		$N_{\text{IPD}} = 200\,(N_{\text{VAL}} = 619)$	
	AUC (SE)	BS (SE)	AUC (SE)	BS (SE)
SLR	0.739 (0.021)	0.201 (0.008)	0.728 (0.013)	0.207 (0.007)
FULL	0.744 (0.020)	0.205 (0.007)	0.742 (0.010)	0.207 (0.004)
UMA	0.744 (0.020)	0.199 (0.008)	0.743 (0.010)	0.199 (0.005)
MMA	0.745 (0.020)	0.198 (0.008)	0.743 (0.010)	0.199 (0.005)
BI	0.745 (0.019)	0.198 (0.008)	0.743 (0.010)	0.199 (0.005)

For multivariate meta-analysis and Bayesian Inference, we used uninformative regression coefficients when missing. Missing within-study correlations were assumed to equal 0.

used an IPD ($N = 1\,028$) from the Amsterdam-Maastricht-Utrecht Study on thromboEmbolism (AMUSE-1) [44] and aggregated these data with the prediction models described below. A detailed description of the predictors can be found in the online Appendix [66]. After aggregation, we validated the original and aggregated models in an independent dataset of 791 participants (Table A.1 in the Appendix) [261].

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$_{\text{K}} x_K$ (eg. 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:

$$\Pr(\text{DVT presence}) = \text{logit}^{-1}(\beta_{\text{adj}0} + \beta_{\text{adj}1}\text{score}) \tag{6.12}$$

The resulting slope $\hat{\beta}_{\text{adj}1}$ is then multiplied with the reported weights to obtain an estimate for the

Table 6.7: Performance of aggregated prediction models, expressed by means of the Area under the Receiver Operator Characteristic curve (AUC) and the Brier Score (BS).

	EBIC			
	$N_{\text{IPD}} = 500\,(N_{\text{VAL}} = 322)$		$N_{\text{IPD}} = 200\,(N_{\text{VAL}} = 622)$	
	AUC (SE)	BS (SE)	AUC (SE)	BS (SE)
SLR	0.810 (0.019)	0.179 (0.010)	0.801 (0.013)	0.185 (0.007)
FULL	0.814 (0.019)	0.176 (0.009)	0.814 (0.010)	0.176 (0.004)
UMA	0.815 (0.019)	0.176 (0.009)	0.814 (0.010)	0.176 (0.004)
MMA	0.815 (0.019)	0.176 (0.009)	0.814 (0.010)	0.177 (0.005)
BI	0.814 (0.019)	0.176 (0.009)	0.814 (0.010)	0.177 (0.005)
	PHARMOS			
	$N_{\text{IPD}} = 500\,(N_{\text{VAL}} = 356)$		$N_{\text{IPD}} = 200\,(N_{\text{VAL}} = 656)$	
	AUC (SE)	BS (SE)	AUC (SE)	BS (SE)
SLR	0.642 (0.024)	0.237 (0.007)	0.627 (0.017)	0.243 (0.007)
FULL	0.653 (0.022)	0.242 (0.008)	0.656 (0.009)	0.242 (0.004)
UMA	0.654 (0.023)	0.236 (0.008)	0.657 (0.009)	0.236 (0.004)
MMA	0.654 (0.024)	0.236 (0.008)	0.657 (0.009)	0.236 (0.004)
BI	0.654 (0.024)	0.236 (0.008)	0.657 (0.009)	0.236 (0.004)

For multivariate meta-analysis and Bayesian Inference, we used uninformative regression coefficients when missing. Missing within-study correlations were assumed to equal 0.

original regression coefficients, and $\hat{\beta}_{\text{adj0}}$ is used as estimate for the model intercept. Conservative estimates for the corresponding standard errors can be obtained by assuming

$$\sigma_{\text{adj1}} = \left(\sum_{j=1}^{M} \sigma_j^{-2} \right)^{-1/2} \tag{6.13}$$

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 directly obtained from $\hat{\sigma}_{\text{adj0}}$. Alternatively, reported p-values of regression coefficients can be converted into standard errors by assuming normality [12]. 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 $\hat{\sigma}_{\text{adj0}} = 0.15$ and $\hat{\sigma}_{\text{adj1}} = 0.05$, 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 $= 0.125$).

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 4 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 6.8. These were then pooled. In order to appraise the quality of the derived model (which only included 4 core predictors), we also fitted a more complex prediction model where we considered the 8 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 models. Finally, we compared the simplified aggregated models to a more extensive model derived with univariate meta-analysis using the 8 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).

Results in the DVT case study

Results in Table 6.9 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 8 predic-

Table 6.8: 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**						
	Wells $N = 593$	M-Wells $N = 530$	Gagne $N = 276$	Hamilton $N = 309$	Oudega $N = 1\,295$	IPD (4) $N = 1\,028$	IPD (8) $N = 1\,028$
(Intercept)	-2.7 (0.2)	-2.8 (0.2)	-1.7 (0.1)	-2.7 (0.2)	-5.5 (NA)	-4.0 (0.3)	-4.7 (0.4)
altdiagn	-1.1 (0.2)	-1.1 (0.2)	-1.8 (0.2)				
calfdif3	0.5 (0.2)	0.5 (0.2)	0.7 (0.2)	0.4 (0.2)	1.1 (0.3)	0.9 (0.2)	0.9 (0.2)
ddimdich					3.0 (0.9)	2.4 (0.3)	2.4 (0.3)
eryt				0.4 (0.2)			
histdvt		0.5 (0.2)	0.6 (0.2)	0.9 (0.2)			
leg	0.5 (0.2)	0.5 (0.2)					
malign	0.5 (0.2)	0.5 (0.2)	1.7 (0.2)	0.9 (0.2)	0.4 (0.2)	0.8 (0.4)	0.7 (0.4)
notraum					0.6 (0.2)		0.6 (0.3)
oachst			1.2 (0.2)		0.8 (0.2)		-12.4 (535)
par	0.5 (0.2)	0.5 (0.2)		0.9 (0.2)			
pit	0.5 (0.2)	0.5 (0.2)					
sex				0.4 (0.2)	0.6 (0.2)		0.6 (0.2)
surg	0.5 (0.2)	0.5 (0.2)	0.5 (0.2)	0.4 (0.2)	0.4 (0.2)	-0.1 (0.4)	0.0 (0.4)
tend	0.5 (0.2)	0.5 (0.2)					
vein	0.5 (0.2)	0.5 (0.2)			0.5 (0.2)		0.2 (0.3)

M-Wells = Modified Wells; IPD (4) and IPD (8) represent the models derived from the AMUSE-1 study, with 4 and 8 core predictors respectively.

tors 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 4 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 $\rho(\hat{\beta}_{\text{ddimdich}}, \hat{\beta}_{\text{malign}}) = -0.97$ [198]. Finally, we noticed that standard errors of aggregated regression coefficients tend to be smallest when estimated with Bayesian Inference.

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 [133, 167, 243, 271]. The case-studies demonstrate that the proposed methods are particularly useful when few participant data are at hand. Although the aggregation methods perform similarly in most scenarios, multivariate 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 [129], 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 [127, 198]. 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 to 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

Table 6.9: Multivariable regression coefficients (and standard error) of the aggregated prediction models in the DVT application.

Approach		**Logistic regression coefficients for DVT outcome**					**Model performance**	
Model	Imputation	(Intercept)	malign	surg	calfdif3	ddimdich	AUC	BS
SLR		-4.0 (0.3)	0.8 (0.4)	-0.1 (0.4)	0.9 (0.2)	2.4 (0.3)	0.72 (0.02)	0.12 (0.01)
UMA	Uninformative	-4.0 (0.1)	0.8 (0.2)	0.5 (0.1)	0.6 (0.1)	2.4 (0.3)	0.73 (0.02)	0.12 (0.01)
MMA	Uninformative	-3.5 (0.1)	0.8 (0.2)	0.4 (0.1)	0.6 (0.1)	2.0 (1.0)	0.73 (0.02)	0.12 (0.01)
BI	Uninformative	-3.3 (0.1)	0.5 (0.1)	0.5 (0.1)	0.7 (0.1)	1.6 (0.2)	0.74 (0.02)	0.12 (0.01)
UMA	Mean	-4.1 (0.1)	0.8 (0.2)	0.5 (0.1)	0.6 (0.1)	2.6 (0.2)	0.73 (0.02)	0.12 (0.01)
MMA	Mean	-4.0 (0.1)	0.7 (0.2)	0.4 (0.1)	0.7 (0.1)	2.4 (0.5)	0.74 (0.02)	0.12 (0.01)
BI	Mean	-3.9 (0.1)	0.7 (0.2)	0.4 (0.1)	0.8 (0.1)	2.3 (0.2)	0.74 (0.02)	0.12 (0.01)

SLR modeling is a standard logistic regression analysis ignoring evidence from the literature, univariate meta-analysis (UMA) ignores within- and between-study covariance, multivariate meta-analysis (MMA) and Bayesian Inference (BI) restrict missing within-study covariance to zero. 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. We considered two imputation approaches for non-included/missing regression coefficients: $\hat{\beta}_\phi = 0$ with $\hat{\sigma}^2_\phi = 100$ (uninformative) and mean imputation with $\hat{\sigma}^2_\phi = \sum_{j=1}^{M} \hat{\sigma}^2_{\phi j}$ (mean).

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 data set 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% [113].

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 over- 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 [241]. 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.

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 since 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 [57, 185]. For this reason, we also considered model accuracy in terms of the Brier Score.

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.

ACKNOWLEDGEMENTS

We would like to thank Stan Buckens for his input and comments during the review process. We also thank those researchers who agreed to share their individual participant data from the International Mission for Prognosis and Analysis of Clinical Trials (IMPACT) project to facilitate this article. Finally, we gratefully acknowledge the following investigators for sharing of individual participant data from the deep vein thrombosis (DVT) studies: A.J. Ten Cate-Hoek, R. Oudega, K.G.M. Moons and D.B. Toll.

7

Meta-analysis and aggregation of multiple published prediction models

Second revision at Statistics in Medicine

Thomas P. A. Debray
Hendrik Koffijberg
Daan Nieboer
Yvonne Vergouwe
Ewout W. Steyerberg
Karel G. M. Moons

Abstract

Published clinical prediction models are often ignored during the development of novel prediction models despite similarities in populations and intended usage. The plethora of prediction models that arise from this practice may still perform poorly when applied in other populations. Incorporating prior evidence might improve the development of prediction models, and make them potentially better generalizable. Unfortunately, aggregation of prediction models is not straightforward and methods to combine differently specified models are currently lacking. We propose two approaches, Model Averaging and Stacked Regressions, for aggregating previously published prediction models when a validation dataset is available. These approaches yield user-friendly stand-alone models that are adjusted for the new validation data. Both approaches rely on weighting to account for model performance and between-study heterogeneity, but adopt a different rationale (averaging versus combination) to combine the models. We illustrate their implementation in two clinical datasets and compare them with established methods for prediction modeling in a simulation study. Results from the clinical datasets and simulation studies demonstrate that aggregation yields prediction models with an improved discrimination and calibration in a vast majority of scenarios, and results in equivalent performance (compared to developing a novel model from scratch) when validation samples are relatively large. In conclusion, model aggregation is a promising extension of model updating when several models are available from the literature, and a validation dataset is at hand. The aggregation methods do not require existing models to have similar predictors and can be applied when relatively few data are at hand.

"If I have seen further it is by standing on the shoulders of Giants."

– Isaac Newton, *The Correspondence Of Isaac Newton*

The past decades has seen a great emphasis on explicitly modelling diagnoses and prognoses in medicine, with gravid appreciation for prediction models [11, 159, 170, 171, 237]. In the cardiovascular domain, well known prediction models are the Framingham [289], SCORE [56], ASSIGN [292], EuroSCORE [176], PROCAM [19] and Wells' scores [281, 283]. Unfortunately, many prediction models perform more poorly than anticipated when taken from the research settings in which they were developed and applied in routine care [74]. This deficiency may occur when prediction models were developed from relatively small datasets or used inappropriate modelling strategies, leading to poorly estimated predictor effects and over-optimism [241, 242]. However, model performance does not necessarily improve when prediction models are developed from larger studies such as Individual Participant Data (IPD) meta-analyses [68]. This is because baseline risks and predictor effects may vary across different patient populations, i.e. between-study heterogeneity occurs. Several authors have therefore recommended that model development should be proceeded by external validation studies assessing the performance of developed prediction models in a new sample before they are implemented in guidelines or applied in practice [30, 169, 237, 259].

When an external validation study shows disappointing results, researchers often reject the original prediction model and develop a new one from their own data [132, 171, 239]. This practice is an unfortunate habit as it makes prediction research particularistic and prior knowledge is not optimally used. Moreover, validation studies are often smaller than development studies, such that the accuracy of the new model (from the validation sample) in future patients can actually be worse than applying the original model. Finally, redevelopment results in the publication of multiple models predicting the same outcomes for the same (or similar) patients or individuals. These models often incorporate different predictors, adding to the incompatibility and confusion. The user must then choose between a cacophony of existing models for which performance may be obscure, which is far from straightforward. For example, there are over 60 published models aiming to predict outcome after breast cancer [11], over 25 for predicting long-term outcome in neurotrauma patients [186], over 14 for identifying patients at risk of prolonged stay at the Intensive Care Unit [75] and over 12 for predicting the risk of cardiovascular disease in patients with type 2 diabetes [269].

An alternative solution to redevelopment is to update existing prediction models with the external validation sample at hand [16, 132, 169, 239, 270]. The updated model is then based on both the development and validation data, further improving its performance in the new population. The proliferation of published prediction models shows a tendency to adopt this strategy, as it typically requires less data than model redevelopment. Updating methods are, however, not a panacea against poorly-conceived and underpowered prediction research [239, 276]. In addition, the limited scope of updating methods does not accomodate the use of evidence from other potentially useful models. For this reason, some form of evidence synthesis during model updating may substantially improve its performance in local or contemporary circumstances, and further contribute to its generalizability.

Meta-analysis has recently emerged as a crux to unravel the incompatibility and confusion of results arising from heterogeneous studies [70, 106, 109, 252]. This practice is now commonly applied in therapeutic research where effect estimates for a particular intervention from different trials are combined and synthesised. We therefore anticipate that a meta-analysis of previously published prediction models would help breaking the cycle of under-powered model development, poor generalizability and redevelopment [65, 66].

Here we present two approaches that extend the classical paradigm of model validation and updating by applying model aggregation. These approaches combine the literature models into a so-called 'meta-model' that weights the predictor-outcome associations from the original models according to their performance in the validation sample and is adjusted for the local circumstances. The first approach, Model Averaging, generates a meta-model that focalizes on the best performing literature model. Conversely, the second approach is called Stacked Regressions and yields an optimal linear combination of the literature models. Both approaches may discard literature models when they are not deemed relevant for the validation population. Although we target a binary prediction task, the proposed approaches could easily be extended to other outcome types. Through a series of case studies we demonstrate that these approaches can be used to combine multiple models that exist for the same outcome or target population, and improve the identification of new predictors.

EMPIRICAL EXAMPLE

In order to illustrate the methods below we describe a genuine clinical example involving the prediction of Deep Vein Thrombosis (DVT). DVT is a blood clot that forms in a vein in the body (usually in the lower leg or thigh). A (part of such) clot can break off and be carried through the bloodstream to the lungs and there cause a blockage (pulmonary embolism), preventing oxygenation of the blood and potentially causing death. Clinical DVT diagnosis is not straightforward. For this reason, multivariable diagnostic prediction models have been developed during the past decades [182, 261, 281]. These models predict the probability of presence of DVT in suspected patients using various patient characteristics obtained from history taking and physical examination to safely exclude DVT without having to perform further testing. Physicians may doubt whether or when to use such a diagnostic prediction model as most of these models have not previously been validated and their performance may change when applied to the heterogeneous reality of routine care. Consequently, a validation study may indicate which models are indeed useful and allow recalibration if necessary.

We previously identified 5 published prediction models for diagnosing DVT (Table 7.1) [66], and collected a validation sample of 1 028 subjects. Three of these models represent score charts (Wells, Modified Wells, Hamilton), whereas the remaining two models represent logistic regression models (Gagne, Oudega). For the logistic regression models, the predicted probability of DVT presence can be calculated as $\text{logit}^{-1}(\text{LP}) = 1/(1+\exp(-\text{LP}))$, where LP represents the linear predictor. For instance, the linear predictor of the Oudega model is given as:

$$\begin{aligned}\text{LP}_{\text{Oudega}} = & -5.47 + 0.42\, x_{\text{malign}} + 0.38\, x_{\text{surg}} + 1.13\, x_{\text{calfdif3}} + 0.48\, x_{\text{vein}} \\ & + 0.75\, x_{\text{oachst}} + 0.59\, x_{\text{sex}} + 0.60\, x_{\text{notraum}} + 3.01\, x_{\text{ddimdich}}\end{aligned}$$

such that a female subject ($x_{\text{sex}} = 0$) with an active malignancy ($x_{\text{malign}} = 1$), no recent surgery ($x_{\text{surg}} = 0$), a calf difference $\geq$ 3cm ($x_{\text{calfdif3}} = 1$), no vein distension ($x_{\text{vein}} = 0$), not using oral contraceptives or hst ($x_{\text{oachst}} = 0$), no leg trauma ($x_{\text{notraum}} = 1$) and positive D-dimer test ($x_{\text{ddimdich}} = 1$) has a linear predictor of -0.31 and a corresponding DVT probability of 42%. Note that the score charts can also be interpreted as a regression model as they typically assume linearity of predictor effects.

Below, we briefly discuss how the literature models would typically be identified, validated and updated. Afterwards, we describe how the updated literature models can be combined into a new summary model that captures all the available evidence.

CLASSICAL PARADIGM: MODEL VALIDATION AND UPDATING

The generalizability of prediction models is typically evaluated in so-called external validation studies where individuals are ‘different but related’ to the development sample [169]. Because many models predict the same outcome for a similar patient population, it is increasingly common to identify such models by means of a systematic review and validate them all togheter [1, 11,

Table 7.1: Overview of 5 previously published, updated and aggregated prediction models for predicting DVT.

Orig.	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\theta}_5$	$\hat{\theta}_6$	$\hat{\theta}_7$	$\hat{\theta}_8$	$\hat{\theta}_9$	$\hat{\theta}_{10}$	$\hat{\theta}_{11}$	$\hat{\theta}_{12}$	$\hat{\theta}_{13}$	$\hat{\theta}_{14}$
Wells		1	1	1	1	1	1	1	1	-2					
M-Wells		1	1	1	1	1	1	1	1	-2	1				
Gagne		2.0		0.6			0.8			-2.1	0.7	1.4			
Hamilton		2	2	1			1				2		1		
Oudega	-5.5	0.4		0.4			1.1		0.5			0.8	0.6	0.6	3.0
Updated models															
Wells †	-1.1	0.5	-0.4	-0.1	-0.6	0.3	0.4	0.3	0.1	-2.5					
M-Wells †	-1.2	0.6	-0.3	-0.0	-0.6	0.3	0.5	0.3	0.1	-2.5	0.7				
Gagne	-1.7	1.7		0.5			0.7			-1.8	0.6	1.2			
Hamilton †	-2.9	0.9	0.1	0.1			1.2				0.8		0.6		
Oudega	-4.6	0.3		0.3			0.9		0.4			0.6	0.5	0.5	2.4
Aggregated models															
MA	-4.6	0.3		0.3			0.9		0.4	-0.0	0.0	0.6	0.5	0.5	2.4
SR	-3.6	1.2		0.5			1.0		0.3	-1.0	0.4	1.1	0.3	0.3	1.6

M-Wells = Modified Wells; MA = Model Averaging; SR = Stacked Regressions

For each model, the original and updated regression coefficients (or points) are provided: $\hat{\theta}_0$ (Model intercept), $\hat{\theta}_1$ (malign), $\hat{\theta}_2$ (par), $\hat{\theta}_3$ (surg), $\hat{\theta}_4$ (tend), $\hat{\theta}_5$ (leg), $\hat{\theta}_6$ (calfdif3), $\hat{\theta}_7$ (pit), $\hat{\theta}_8$ (vein), $\hat{\theta}_9$ (altdiagn), $\hat{\theta}_{10}$ (histdvt), $\hat{\theta}_{11}$ (oachst), $\hat{\theta}_{12}$ (sex), $\hat{\theta}_{13}$ (notraum) and $\hat{\theta}_{14}$ (ddimdich). Because not all models included the same predictors, some parameters are left blank (but equal, in fact, zero). Updating was achieved by estimating a calibration intercept and slope in the validation sample ($N = 1\,028$).

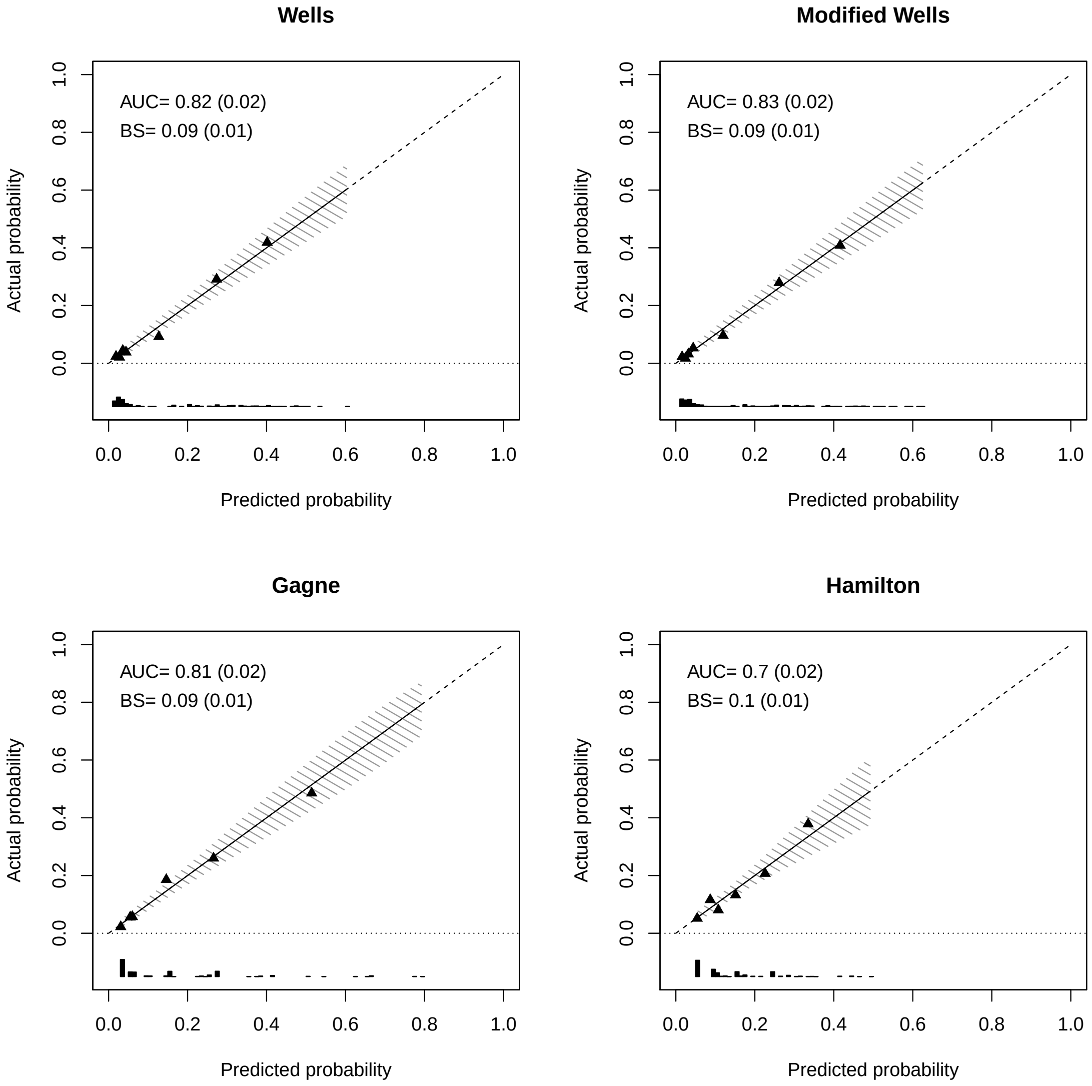
† For the Wells, Modified Wells and Hamilton rule, individual regression coefficients were re-estimated because the original model represented a score chart.

55, 75, 159, 186, 224]. The systematic review required typically begins with a literature search of electronic databases such as Medline and Embase [85, 105, 123, 291]. A critical appraisal can help to identify those literature models that may indeed be useful for the intended outcome or patient population, and to exclude the models that are deemed irrelevant or of poor quality. Although there are no explicit guidelines for such prognostically-orientated appraisal at this stage, several comprehensive item lists have been proposed that are based on existing methodological recommendations for conducting and reporting prediction research [34, 55, 104, 186].

Once all relevant prediction models have been identified, their performance is evaluated in the validation sample. This is achieved by applying the model to the available subjects, and comparing the predicted risks to the observed outcomes. The performance of the evaluated models can be quantified in terms of calibration and discrimination. Calibration reflects the extent to which the predicted probabilities agree with observed event rates, whereas discrimination is the ability to distinguish high-risk patients from low-risk patients. The Area under the Receiver Operating Characteristic curve (AUC) is a common summary measure of discrimination and is strongly related to the Brier Score (BS), an overall performance measure [77, 99, 116, 160, 237, 249]. Other measures are also available (e.g. calibration-in-the-large, calibration slope, R^2 and Goodness-of-fit), and may be equally useful [116, 221, 239]. In the empirical example, we found the following AUC for the original models when applied in the validation sample: 0.67 (Hamilton), 0.76 (Wells), 0.77 (Modified Wells), 0.81 (Gagne) and 0.82 (Oudega).

Finally, the model(s) showing the most appealing characteristics in the validation sample can be selected for updating. Here, the previously published prediction models are recalibrated by re-estimating some of their parameters in the validation sample [132, 169, 237, 239, 274]. The most straightforward strategy of model updating is to adjust its intercept such that the mean predicted probability is equal to the observed event rate (intercept update). Additional updating methods vary from overall adjustment of the model intercept and the overall calibration slope (logistic calibration), adjustment of a particular regression coefficient, to the re-estimation of included or the addition of completely new predictors to the exsiting model (model revision). It is, however, important to realize that extensive updating strategies use more information from the validation sample at hand and may therefore lead to overfitting. In addition, extensive updating strategies adust the model to the validation sample and therefore reduce its evaluated external validity to internal validity. For this reason, updating strategies should be carefully conducted when the validation sample is relatively small. In the empirical example, we updated all available models using logistic calibration (Gagne and Oudega) or model revision (Wells, Modified Wells and Hamilton). The updated models are presented in Table 7.1 and their resulting calibration in the validation sample is depicted in Figure 7.1 and 7.2. Here, the Wells models achieved the best performance, and the AUC increased from 0.76 to 0.82 (Wells) and from 0.77 to 0.83 (Modified

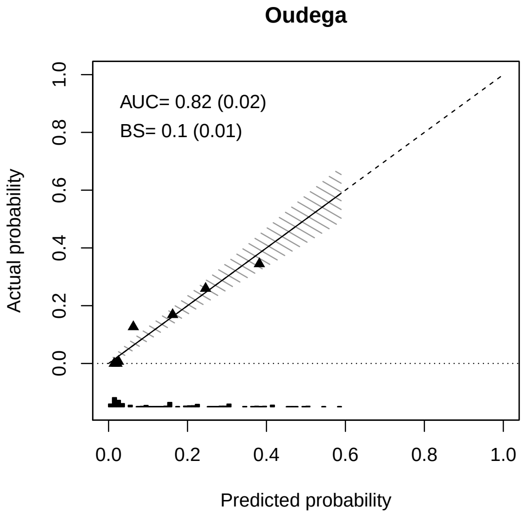
Figure 7.1: Calibration plots of the updated prediction models in the empirical example.



Calibration curves of the aggregated models in the validation sample and corresponding 95% confidence intervals (shaded area). The Area under the ROC curve (AUC) and the Brier score (BS) are presented with their standard error. The triangles indicate groups of observations with similar predicted probabilities and their corresponding outcome proportion.

Wells) in the validation sample. However, because this performance was only attained by extensive updating strategies (i.e. re-estimation of individual regression coefficients), the updated Wells models are likely overfitted to the validation sample. The Oudega and Gagne model involved less adjustments and also achieved good performance.

Figure 7.2: Calibration plots of the updated prediction models in the empirical example.



Calibration curves of the aggregated models in the validation sample and corresponding 95% confidence intervals (shaded area). The Area under the ROC curve (AUC) and the Brier score (BS) are presented with their standard error. The triangles indicate groups of observations with similar predicted probabilities and their corresponding outcome proportion.

Although updating strategies may effectively adjust literature models to local circumstances, their extensiveness is usually impeded by a lack of validation data. Furhermore, because the limited scope of updating methods does not accomodate the accumulation of other potentially useful models, updating methods may not always improve model performance. In the next section, we therefore describe how meta analaytical approaches or model aggregation can be augmented to external validation studies.

MODEL AGGREGATION

We here describe two aggregation approaches and consider the situation in which a literature search and a critical appraisal have been performed. These approaches extend the classical paradigm by aggregating all literature models into a meta-model that is optimized for the validation data at hand. The meta-model is then based upon a broader base of prior evidence and is likely to provide more insight into which predictive variables are truly informative.

Here, we consider that the validation sample is described by K independent predictors, a dichotomous outcome, and contains N subjects. For instance, in the empirical example we have $N = 1\,028$

and $K = 13$. Let $\boldsymbol{X}$ denote the $N \times K$ matrix of all the independent variables theorized to be predictors of outcome $\boldsymbol{y}$ based on the set of literature models. We assume that the validation sample captures all predictors included in the previously published prediction models, or at least represents good proxy variables [1, 75]. We denote the set of literature models as $\mathcal{M} = [\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_M]$, where M corresponds to the total amount of literature models that are being aggregated. Here, each prediction model $\mathcal{M}_j$ considers the set of predictors $\mathcal{C}_j$ and is parameterized by vector $\hat{\boldsymbol{\theta}}_{\boldsymbol{j}}$. For the empirical example, these parameters are presented in Table 7.1. Although technically any type of prediction model (such as regresson models, decision trees, neural networks, etc.) may form the basis of model aggregation, we here consider the case that all literature models were developed using logistic regression, such that:

$$\mathcal{M} = \left\{ \begin{array}{c} \mathcal{M}_1 \\ \mathcal{M}_2 \\ \vdots \\ \mathcal{M}_M \end{array} \right\} = \left\{ \begin{array}{c} \text{logit}^{-1}(\text{LP}_1) \\ \text{logit}^{-1}(\text{LP}_2) \\ \vdots \\ \text{logit}^{-1}(\text{LP}_M) \end{array} \right\} = \left\{ \begin{array}{c} \text{logit}^{-1}(\hat{\theta}_0 + \hat{\theta}_1 \boldsymbol{x}_1 + \ldots + \hat{\theta}_K \boldsymbol{x}_K)_1 \\ \text{logit}^{-1}(\hat{\theta}_0 + \hat{\theta}_1 \boldsymbol{x}_1 + \ldots + \hat{\theta}_K \boldsymbol{x}_K)_2 \\ \vdots \\ \text{logit}^{-1}(\hat{\theta}_0 + \hat{\theta}_1 \boldsymbol{x}_1 + \ldots + \hat{\theta}_K \boldsymbol{x}_K)_M \end{array} \right\} \quad (7.1)$$

where $\boldsymbol{x}_k \subset \boldsymbol{X}$ represents a vector with the observations of predictor k for all subjects and where it is possible that $(\hat{\theta}_k)_m = 0$ if model m does not include predictor k. Note that LP_m represents the linear predictor of model m.

Model Averaging

A straigthforward approach to aggregate models with a varying apparent performance is to create a weighted average. Because the resulting meta-model takes all available evidence into account, it tends to outperform each of the original models (given an appropriate choice of averaging weights). A well-known implementation of model averaging is Bayesian Model Averaging (BMA), where multiple models are developed from the same data and subsequently combined into a meta-model [119, 164]. Here, we adapt BMA to allow the aggregation of models that were developed from different samples and may be heterogeneous with the validation sample. The resulting strategy consists of three individual steps, described below.

In the first step, the literature models are updated in the validation sample to increase their mutual comparability. Hereto, strategies such as intercept update, model calibration or even model revision may be considered and selected by hand or closed-testing procedures [274]. Afterwards the updated models are applied in the validation sample to calculate a predicted outcome event $\hat{p}_{im} = \text{logit}^{-1}(\text{LP}_{im})$ for each subject (leading to $M \times N$ predicted probabilities). A model averaged prediction for subject i (where $i = 1, \ldots, N$) can then be written as:

$$\overline{p}_i = \sum_{m=1}^{M} w_m \hat{p}_{im} \tag{7.2}$$

where w_m represents the weight associated to model $\mathcal{M}_m$ (and all weights sum up to 1).

In the second step, appropriate model weights are estimated. A simple weight scheme may assign equal weight $w_m = 1/M$ to each model. As a result, $\overline{p}_i$ then represents the geometric mean of the predictions for subject i. An alternative approach to weight the predictions from the updated literature models is to rely on the concepts of BMA, where the predictions from each model are weighted by the posterior model probability [119]:

$$w_m = \frac{\exp(-0.5\ \mathrm{BIC}_m)}{\sum_{l=1}^{M} \exp(-0.5\ \mathrm{BIC}_l)} \tag{7.3}$$

where $\mathrm{BIC}_m = -2\ \ell_m + u_m \ln(N)$ and u_m represents the number of estimated parameters for updating literature model m. Typical values for u are 1 (intercept update), 2 (logistic calibration) or $K + 1$ (model revision). Finally, ℓ_m represents the log-likelihood of model m in the validation sample and is given as $\ell_m = \sum_{i=1}^{N} (y_i \ln(\hat{p}_{im}) + (1 - y_i) \ln(1 - \hat{p}_{im}))$. Note that ℓ_m takes lower values for decreasing model fit in the validation sample. Consequently, the likelihood of each updated literature model is penalized according to how much information from the validation sample was used to improve its external performance. In summary, the BIC ensures that models with good performance in the validation sample and not extensively updated will have a larger contribution in the summarized predictions. Conversely, models that perform poorly (such that $-2\,\ell_m$ increases) or were extensively updated (such that $u_m \ln(N)$ increases) receive lower weights (as BIC_m increases and w_m therefore decreases). Note that small differences in BIC usually lead to large differences in posterior model probabilities (i.e. resulting model weights) because the exponential transformation needs to be applied. This implies that some models may end up with weights very close to zero or one, particularly when few literature models are under consideration. As a consequence, Model Averaging may lead to model selection.

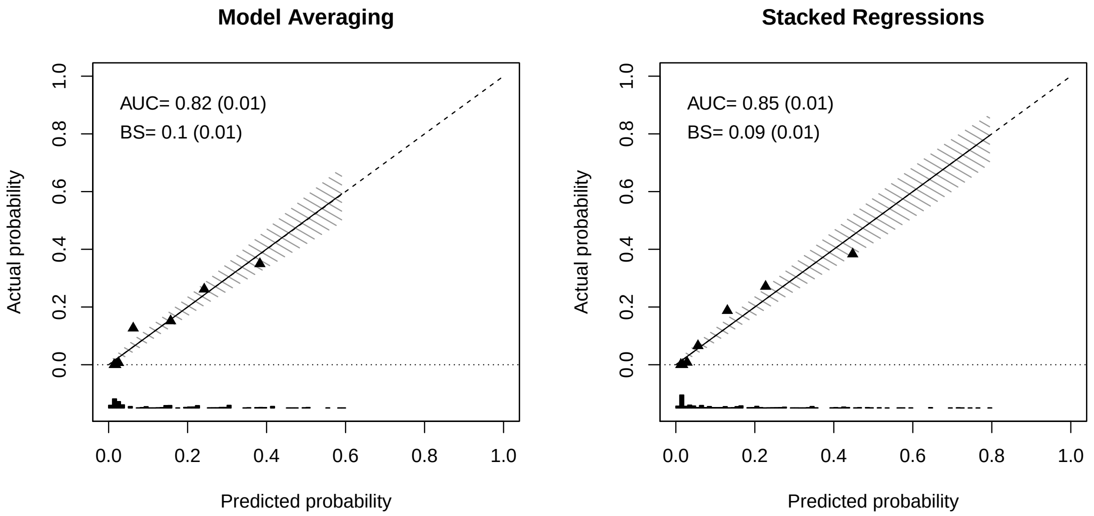
In the empirical example, the log-likelihood of the updated models in the validation sample is given as -360 (Hamilton), -318 (Gagne), -312 (Oudega), -307 (Wells) and -304 (Modified Wells). Although the Wells ($u = 10$) and Modified Wells ($u = 11$) models achieve the best fit in the validation sample, the Oudega and Gagne models involved less extensive updating strategies ($u = 2$). As a consequence, the BIC of of these models was much lower and resulted in non-zero weights (0.998 and 0.002 for the Oudega and Gagne model respectively).

Finally, in the third step the models's averaged predictions $\overline{p}_i$ are used to develop the meta-model. This is achieved by performing a linear regression analysis where the original predictor variables are used as independent variables, and the averaged predictions are used as dependent variable. Here, we apply the logistic transformation to ensure linearity of the dependent variable, the transformed outcome is then given as $z_i = \text{logit}\,(\overline{p}_i)$. Furthermore, we only include the predictor variables from the literature models for which $w \geq 0.0001$, such that the total amount of predictors reduces to K_{AVG}. The linear regression analysis is then given as:

$$\begin{aligned} z_i &= \beta_0 + \sum_{k=1}^{K_{\text{AVG}}} \beta_k x_{ik} + \epsilon_i \\ \epsilon_i &\sim \mathcal{N}(0, \sigma^2) \end{aligned} \tag{7.4}$$

The unknown parameters β_0 (model intercept), β_k (predictor effects) and σ^2 (error variance) can be estimated with maximum likelihood estimation. By allocating low weights to estimates from poorly fitting or extensively updated models we effectively push the weighted average towards the validation sample, maximizing its relevance for the corresponding population. The meta-model then represents a logistic regression model with intercept β_0 and predictor effects β_k.

Figure 7.3: Calibration plots of the aggregated prediction models in the empirical example.



Calibration curves of the aggregated models in the validation sample and corresponding 95% confidence intervals (shaded area). The Area under the ROC curve (AUC) and the Brier score (BS) are presented with their standard error. The triangles indicate groups of observations with similar predicted probabilities and their corresponding outcome proportion.

In the empirical example, we developed a meta-model using the predictors from the Oudega and Gagne models (since $w_{\text{Oudega}} = 0.998$ and $w_{\text{Gagne}} = 0.002$) and the averaged linear predictor as outcome. Since $w = 0$ for the Wells, Modified Wells and Hamilton model, the corresponding updated literature models are not included in the meta-model. The regression coefficients of the resulting meta-model are depicted in Table 7.1 and indicate that the meta-model is almost identical to the updated Oudega model. The calibration of the meta-model is depicted in Figure 7.3, where the AUC and Brier Score have slightly improved. Consequently, in this example Model Averaging has led to a selection of literature models, and yielded an aggregated model that captures the updated *Oudega* model and also includes two new predictors *altdiagn* and *histdvt* from the *Gagne* model.

Stacked Regressions

There are several concerns with the Model Averaging approach. Firstly, Model Averaging requires the user to update each literature model to ensure that it is adjusted to the validation population. It may be clear that such strategy uses relatively much information from the validation sample and may lead to overfitting. In addition, there is no formal procedure to choose an appropriate updating strategy. Secondly, the implementation of Model Averaging tends to produce extreme weights being assigned to models, leading to a degree of skewing towards stronger models. This is because Model Averaging operates under the assumption that only one of the literature models is correct and places too much weight on their maximum likelihood [163]. As a consequence, Model Averaging generally reduces to a selection procedure that accounts for model uncertainty [161, 163].

For this reason, we here propose a second approach that emphasizes model combination rather than informative model selection. This approach is based on Stacked Regressions [37] and relates to ensemble learning where the predictions of multiple models are combined into a weighted summary [37, 290]. However, instead of individually identifying and subsequently averaging the best updated literature models, Stacked Regressions simultaneously updates, discovers and estimates the best combination of literature models in the validation sample. This implies that the aforementioned steps of Model Averaging no longer need to be applied, and that the meta-model is developed from the original literature models forthwith.

In essence, Stacked Regressions treats the predictions of each literature model as a predictor variable of the meta-model and subsequently creates a linear combination of model predictions:

$$\begin{aligned} y_i &\sim \text{Bernoulli}(\pi_i) \\ \text{logit}(\pi_i) &= \alpha_0 + \sum_{m=1}^{M} \alpha_m \text{LP}_{im} \end{aligned} \tag{7.5}$$

under the constraint $\alpha_m \geq 0$ to ensure that models with a negative contribution on the combined prediction will effectively be discarded from the meta-model. The unkown parameters $\alpha_0, \alpha_1, \ldots, \alpha_M$ can be estimated by minimizing an error function using bound constrained optimization [47]. For instance, the original implementation of Stacked Regressions adopted a squared error loss function related to the Brier Score [37]:

$$\sum_{i=1}^{N} \left(y_i - \alpha_0 - \sum_{m=1}^{M} \alpha_m \hat{p}_{im} \right)^2 \tag{7.6}$$

Thus, the predictions of model m for subject i, i.e. $\hat{p}_{im} = \text{logit}^{-1}(\text{LP}_{im})$, are combined in the validation sample by means of a weighted sum that minimizes their decrepancy with the actually observed outcomes y_i. Hereto, the predictions of each model are weighted by an independent parameter α_m that emphasizes good prediction in overall, and penalizes models with poor performance or extreme predictions (similar to logistic calibration). The weight parameter α_0 is unrestricted and ensures that the baseline risk of the synthesis model is optimal for the validation sample (similar to intercept updating). We further adapted this minimization function to implement the Maximum Likelihood Estimator:

$$\begin{aligned} -\Bigg[\sum_{i=1}^{N} y_i \ln \left(1 + \exp \left(-\alpha_0 - \sum_{m=1}^{M} \alpha_m \text{LP}_{im} \right) \right) - (1 - y_i) \\ \ln \left(1 + \exp \left(\alpha_0 + \sum_{m=1}^{M} \alpha_m \text{LP}_{im} \right) \right) \Bigg] \end{aligned} \tag{7.7}$$

again under the constraint $\alpha_m \geq 0$.

In the empirical example, we used the original literature models and calculated their linear predictor in the validation sample (Table 7.1). For score charts such as the Wells model, we used the reported weights as coefficients in LP_{Wells}. We subsequently performed Stacked Regressions and obtained the following weight parameters: 0.50 (Gagne), 0.54 (Oudega) and 0 (Wells, Modified Wells, Hamilton). The residual weight parameter $\hat{\alpha}_0$ was 1.01. Consequently, by applying Stacked Regressions we effectively discarded 3 literature models that did not contribute to the performance of the synthesis model.

Finally, the regression coefficients of the aggregated model (with intercept term β_0 and regression coefficients $\beta_1, \ldots, \beta_K$) can be calculated as a weighted sum of the regression coefficients of the original models: $\beta_0 = \hat{\alpha}_0 + \sum_{m=1}^{M} \hat{\alpha}_m (\hat{\theta}_0)_m$ and $\beta_k = \sum_{m=1}^{M} \hat{\alpha}_m (\hat{\theta}_k)_m$ where $1 \leq k \leq K$. Note that some predictor variables may not be included in the linear predictor as a result of discarding models with α_m equal to zero. Consequently, variable reduction may occur when literature models include peculiar predictors and lead to poor predictions in the IPD. The meta-model is no longer dependent on the predictions from the individual literature models.

In the empirical example, the estimated regression coeffients from the meta-model are depicted in Table 7.1. Although Stacked Regressions identified the same predictors as Model Averaging, different parameter estimates were obtained by both approaches. For instance, the predictor effect of *ddimdich* decreased from 2.39 to 1.62, whereas the predictor effect of *malign* increased from 0.34 to 1.22. Results in Figure 7.3 further indicate that Stacked Regressions achieved the best AUC and BS. Additional validation studies are needed to evaluate whether aggregation has indeed improved external validity.

Concluding remarks

The Model Averaging and Stacked Regressions approaches synthesize previously published prediction models into a summary model that is adjusted by or to the validation data at hand. Model Averaging first updates each literature model by estimating $\sum_{m=1}^{M} u_m$ unknown parameters from the validation sample. Afterwards, M additional parameters are estimated to obtain appropriate model weights. Finally, synthesis is achieved using a linear weighting scheme with $1 + K_{\text{AVG}}$ unknown parameters that are again estimated from the validation sample. Stacked Regressions directly combines all original literature models into a meta-model by estimating $M+1$ (Stacked Regressions) unknown parameters from the validation sample. It is therefore the most parsimonious approach in terms of required degrees of freedom.

Chapter 7

APPLICATIONS

In order to illustrate the approaches, we describe two applications with previously collected datasets used for diagnosis of Deep Venous Thrombosis (DVT) and prognosis of Traumatic Brain Injury (TBI). In each application, we considered Model Averaging using weights based on the BIC and Stacked Regressions (using the Maximum Likelihood Estimator) to combine the previously published prediction models with the validation sample. We also evaluated model calibration (i.e. intercept and overall slope) of the literature model with the highest AUC. Finally we performed stepwise logistic regression using backward selection (based on AIC) and penalised maximum likelihood estimation [101, 168] as alternative approaches ignoring the models from the literature and developing a novel prediction model from the validation data. To evaluate the generalizability of the newly developed, updated and aggregated models, we employed a split-sample procedure to ensure external validation was applied in new subjects. Here, we measured the AUC and Brier Score as indication of model performance.

Deep Venous Thrombosis

As a first example, we performed a simulation study with previously collected clinical data for diagnosing DVT (Table A.1 in the Appendix). This simulation study is based on the data from 7 previously conducted studies ($N = 7\,116$) and 14 candidate predictors ($K = 14$), with a median event rate of 22% (range 13% to 39%). The data of each study were used to develop a prediction model according to stepwise logistic regression with backward selection (based on AIC). These models serve as source for selecting the literature models in the forthcoming analyses.

The simulation study consists of 7 analyses, i.e. one for each available study, and is based on the following procedure. For each study sample, we used a split-sample procedure for generating two validation datasets. This procedure samples N_{VAL1} subjects without replacement from the study sample for applying the described methods (redevelopment using backward selection, redevelopment using PMLE, model updating of intercept and common slope, Model Averaging, Stacked Regressions), and the remaining N_{VAL2} subjects for externally validating the resulting models. In this manner, we can develop and validate the aggregated prediction models in different but related subjects. Afterwards, we evaluated the performance of the developed prediction models by measuring their AUC and BS in the second validation sample. We repeated this process 30 times for each analysis, using two different sample sizes: $N_{\text{VAL1}} = 200$ and $N_{\text{VAL1}} = 500$. By evaluating different sample sizes for the former validation sample, it is possible to ascertain the effect of variable selection and overfitting, and thus expose the need for incorporating external evidence.

Table 7.2: Results from Application 1 based on 30 random split-sample validation samples (AUC and standard error).

	Study 1		**Study 2**		**Study 3**	
N	500 (528)	200 (828)	500 (314)	200 (614)	500 (1256)	200 (1556)
BWS	0.84 (0.02)	0.80 (0.04)	0.79 (0.02)	0.76 (0.02)	0.89 (0.01)	0.87 (0.02)
PMLE	0.85 (0.02)	0.82 (0.02)	0.79 (0.02)	0.77 (0.01)	0.89 (0.01)	0.88 (0.01)
MU†	0.83 (0.02)	0.83 (0.01)	0.78 (0.02)	0.78 (0.01)	0.89 (0.01)	0.89 (0.00)
MA†	0.84 (0.02)	0.84 (0.01)	0.79 (0.02)	0.79 (0.01)	0.89 (0.01)	0.89 (0.01)
SR	0.84 (0.01)	0.84 (0.01)	0.79 (0.02)	0.79 (0.01)	0.90 (0.01)	0.89 (0.01)

	Study 4		**Study 5**		**Study 6**	
N	500 (291)	200 (591)	500 (575)	200 (875)		200 (157)
BWS	0.74 (0.03)	0.72 (0.02)	0.86 (0.01)	0.85 (0.02)		0.75 (0.03)
PMLE	0.75 (0.03)	0.73 (0.03)	0.87 (0.01)	0.85 (0.02)		0.77 (0.03)
MU†	0.74 (0.02)	0.74 (0.02)	0.88 (0.01)	0.88 (0.01)		0.78 (0.02)
MA†	0.75 (0.02)	0.75 (0.01)	0.88 (0.01)	0.88 (0.01)		0.79 (0.02)
SR	0.75 (0.02)	0.75 (0.01)	0.88 (0.01)	0.88 (0.01)		0.79 (0.02)

	Study 7	
N	500 (795)	200 (1095)
BWS	0.77 (0.01)	0.75 (0.01)
PMLE	0.78 (0.01)	0.77 (0.01)
MU†	0.77 (0.02)	0.77 (0.01)
MA†	0.77 (0.01)	0.77 (0.01)
SR	0.77 (0.01)	0.77 (0.01)

BWS = Model redevelopment (backward selection), PMLE = Model redevelopment (penalized Maximum Likelihood Estimation), MU = Model updating, MA = Model averaging, SR = Stacked Regressions

There are 2 validation samples for each scenario. The first validation sample is used for model redevelopment, model updating or model aggregation. The resulting prediction models are then evaluated in the second validation sample (sample size indicated between brackets). The study datasets have median event rate of 22% (range 13% to 39%).

† Literature models are updated by re-estimating the intercept and common slope.

Results in Table 7.2 and 7.3 demonstrate that prediction models developed with stepwise reduction yield the poorest performance, particularly when few data ($N_{\text{VAL1}} = 200$) are available. Results in Table 7.4 illustrate that this approach achieves the poorest discrimination and calibration in 136 and, respectively, 162 scenarios of the 210 considered scenarios. Although the performance of prediction models considerably improves by applying penalization (PMLE), this only holds when relatively much data ($N_{\text{VAL1}} = 500$) are at hand. When few data are at hand ($N_{\text{VAL1}} = 200$), updating the literature models performed slightly better than redevelopment. For instance, updating the best literature model yielded the highest AUC in 40, versus 8 (stepwise reduction) or 24 (PMLE) of the 210 considered scenarios. However, we noticed that the best models were developed by means of aggregation. Results in Table 7.4 indicate that Model Averaging slightly outperforms Stacked Regressions across all repeated scenarios. Particularly, Model Averaging

Table 7.3: Results from Application 1 based on 30 random split-sample validation samples (Brier Score and standard error).

	Study 1		**Study 2**		**Study 3**	
N	500 (528)	200 (828)	500 (314)	200 (614)	500 (1256)	200 (1556)
BWS	0.09 (0.01)	0.10 (0.01)	0.18 (0.01)	0.19 (0.01)	0.11 (0.00)	0.12 (0.01)
PMLE	0.09 (0.01)	0.10 (0.01)	0.18 (0.01)	0.19 (0.01)	0.11 (0.00)	0.12 (0.01)
MU†	0.09 (0.01)	0.10 (0.00)	0.18 (0.01)	0.18 (0.00)	0.11 (0.00)	0.11 (0.00)
MA†	0.09 (0.01)	0.09 (0.00)	0.18 (0.01)	0.18 (0.00)	0.11 (0.00)	0.11 (0.00)
SR	0.09 (0.01)	0.09 (0.00)	0.18 (0.01)	0.18 (0.00)	0.11 (0.00)	0.11 (0.00)

	Study 4		**Study 5**		**Study 6**	
N	500 (291)	200 (591)	500 (575)	200 (875)		200 (157)
BWS	0.12 (0.01)	0.13 (0.01)	0.10 (0.01)	0.11 (0.01)		0.16 (0.01)
PMLE	0.12 (0.01)	0.13 (0.01)	0.10 (0.01)	0.11 (0.01)		0.16 (0.01)
MU†	0.12 (0.01)	0.12 (0.00)	0.10 (0.01)	0.10 (0.00)		0.15 (0.01)
MA†	0.12 (0.01)	0.12 (0.00)	0.10 (0.01)	0.10 (0.00)		0.15 (0.01)
SR	0.12 (0.01)	0.12 (0.00)	0.10 (0.01)	0.10 (0.00)		0.15 (0.01)

	Study 7	
N	500 (795)	200 (1095)
BWS	0.15 (0.00)	0.16 (0.01)
PMLE	0.15 (0.00)	0.15 (0.01)
MU†	0.15 (0.00)	0.15 (0.00)
MA†	0.15 (0.00)	0.15 (0.00)
SR	0.15 (0.00)	0.15 (0.00)

BWS = Model redevelopment (backward selection), PMLE = Model redevelopment (penalized Maximum Likelihood Estimation), MU = Model updating, MA = Model averaging, SR = Stacked Regressions

There are 2 validation samples for each scenario. The first validation sample is used for model redevelopment, model updating or model aggregation. The resulting prediction models are then evaluated in the second validation sample (sample size indicated between brackets). The study datasets have median event rate of 22% (range 13% to 39%).

† Literature models are updated by re-estimating the intercept and common slope.

yielded the best and second best AUC in 77 and respectively 90 of the 210 considered scenarios where $N_{\text{VAL1}} = 200$. Stacked Regressions, however, yielded the best overall performance (BS) in 83 and respectively 63 of the 210 considered scenarios.

Traumatic Brain Injury

In this second application, we performed a simulation study with previously collected clinical data predicting unfavorable outcome 6 months after Traumatic Brain Injury (TBI) (Table A.2 in the Appendix) [152]. This simulation study uses the same procedure as described in the previous application, and is based on the data from 10 studies ($N = 9\,149$) with 9 candidate predictors: *age*,

Table 7.4: Performance comparison of derived prediction models in Application 1 in terms of achieved ranks.

Achieved ranks based on AUC										
$N = 500$	Rank 1		Rank 2		Rank 3		Rank 4		Rank 5	
BWS	26	(14%)	42	(23%)	22	(12%)	33	(18%)	57	(32%)
PMLE	73	(41%)	30	(17%)	19	(11%)	36	(20%)	22	(12%)
MU[†]	30	(17%)	23	(13%)	28	(16%)	27	(15%)	72	(40%)
MA[†]	17	(9%)	50	(28%)	41	(23%)	56	(31%)	16	(9%)
SR	34	(19%)	35	(19%)	70	(39%)	28	(16%)	13	(7%)
$N = 200$										
BWS	8	(4%)	12	(6%)	9	(4%)	45	(21%)	136	(65%)
PMLE	24	(11%)	17	(8%)	30	(14%)	100	(48%)	39	(19%)
MU[†]	40	(19%)	21	(10%)	83	(40%)	38	(18%)	28	(13%)
MA[†]	77	(37%)	90	(43%)	29	(14%)	12	(6%)	2	(1%)
SR	61	(29%)	70	(33%)	59	(28%)	15	(7%)	5	(2%)
Achieved ranks based on Brier Scores										
$N = 500$	Rank 1		Rank 2		Rank 3		Rank 4		Rank 5	
BWS	23	(13%)	19	(11%)	34	(19%)	42	(23%)	62	(34%)
PMLE	47	(26%)	39	(22%)	24	(13%)	45	(25%)	25	(14%)
MU[†]	17	(9%)	24	(13%)	36	(20%)	25	(14%)	78	(43%)
MA[†]	28	(16%)	57	(32%)	37	(21%)	53	(29%)	5	(3%)
SR	65	(36%)	41	(23%)	49	(27%)	15	(8%)	10	(6%)
$N = 200$										
BWS	4	(2%)	4	(2%)	10	(5%)	30	(14%)	162	(77%)
PMLE	23	(11%)	10	(5%)	29	(14%)	128	(61%)	20	(10%)
MU[†]	22	(10%)	37	(18%)	94	(45%)	33	(16%)	24	(11%)
MA[†]	78	(37%)	96	(46%)	29	(14%)	5	(2%)	2	(1%)
SR	83	(40%)	63	(30%)	48	(23%)	14	(7%)	2	(1%)

BWS = Model redevelopment (backward selection), PMLE = Model redevelopment (penalized Maximum Likelihood Estimation), MU = Model updating, MA = Model averaging, SR = Stacked Regressions

Results are based on 300 scenarios (30 random split-sample validation samples for 10 studies), in which the performance of each model is ranked from best (rank 1) to worst (rank 5). The totals (and percentage) of achieved ranks are summarized for each approach. Note that percentages may not always sum up to 100% due to rounding.

[†] Literature models are updated by re-estimating the intercept and common slope.

Table 7.5: Results from Application 2 based on 30 random split-sample validation samples (AUC and standard error).

	Study 74		**Study 75**		**Study 77**	
N	500 (618)	200 (918)	500 (541)	200 (841)	500 (419)	200 (719)
BWS	0.73 (0.02)	0.70 (0.02)	0.74 (0.01)	0.72 (0.02)	0.69 (0.02)	0.67 (0.02)
PMLE	0.73 (0.01)	0.72 (0.01)	0.74 (0.01)	0.73 (0.01)	0.70 (0.03)	0.68 (0.01)
MU[†]	0.72 (0.02)	0.72 (0.01)	0.73 (0.01)	0.73 (0.01)	0.69 (0.02)	0.69 (0.01)
MA[†]	0.73 (0.02)	0.72 (0.01)	0.74 (0.01)	0.73 (0.01)	0.70 (0.02)	0.69 (0.01)
SR	0.73 (0.01)	0.73 (0.01)	0.74 (0.01)	0.73 (0.01)	0.70 (0.02)	0.70 (0.01)

	Study 79		**Study 81**		**Study 85**	
N	500 (1010)	200 (1310)	500 (291)	200 (591)	500 (322)	200 (622)
BWS	0.69 (0.01)	0.68 (0.02)	0.71 (0.02)	0.70 (0.02)	0.77 (0.02)	0.76 (0.02)
PMLE	0.70 (0.01)	0.68 (0.01)	0.71 (0.03)	0.71 (0.01)	0.77 (0.02)	0.76 (0.01)
MU[†]	0.69 (0.01)	0.69 (0.01)	0.72 (0.03)	0.71 (0.01)	0.77 (0.02)	0.77 (0.01)
MA[†]	0.70 (0.01)	0.70 (0.01)	0.72 (0.03)	0.72 (0.01)	0.78 (0.02)	0.77 (0.01)
SR	0.70 (0.01)	0.70 (0.01)	0.72 (0.03)	0.72 (0.01)	0.78 (0.02)	0.77 (0.01)

	Study 86		**Study 89**		**Study 90**	
N	500 (319)	200 (619)	500 (17)	200 (317)	500 (356)	200 (656)
BWS	0.68 (0.02)	0.65 (0.02)	0.56 (0.17)	0.63 (0.02)	0.64 (0.03)	0.61 (0.02)
PMLE	0.69 (0.02)	0.67 (0.02)	0.55 (0.18)	0.63 (0.02)	0.64 (0.03)	0.62 (0.01)
MU[†]	0.68 (0.02)	0.67 (0.02)	0.58 (0.18)	0.65 (0.02)	0.64 (0.02)	0.64 (0.01)
MA[†]	0.68 (0.02)	0.67 (0.01)	0.58 (0.18)	0.66 (0.02)	0.64 (0.02)	0.64 (0.01)
SR	0.68 (0.02)	0.67 (0.01)	0.58 (0.18)	0.66 (0.02)	0.64 (0.02)	0.64 (0.01)

	Study 91	
N	500 (256)	200 (556)
BWS	0.75 (0.02)	0.73 (0.02)
PMLE	0.75 (0.02)	0.73 (0.01)
MU[†]	0.75 (0.02)	0.74 (0.01)
MA[†]	0.75 (0.02)	0.75 (0.01)
SR	0.75 (0.02)	0.75 (0.01)

BWS = Model redevelopment (backward selection), PMLE = Model redevelopment (penalized Maximum Likelihood Estimation), MU = Model updating, MA = Model averaging, SR = Stacked Regressions
There are 2 validation samples for each scenario. The first validation sample is used for model redevelopment, model updating or model aggregation. The resulting prediction models are then evaluated in the second validation sample (sample size indicated between brackets). The study datasets have median event rate of 46% (range 38% to 65%).

[†] Literature models are updated by re-estimating the intercept and common slope.

Table 7.6: Results from Application 2 based on 30 random split-sample validation samples (Brier Score and standard error).

	Study 74		**Study 75**		**Study 77**	
N	500 (618)	200 (918)	500 (541)	200 (841)	500 (419)	200 (719)
BWS	0.21 (0.01)	0.22 (0.01)	0.20 (0.01)	0.21 (0.01)	0.22 (0.01)	0.23 (0.01)
PMLE	0.21 (0.01)	0.21 (0.00)	0.20 (0.00)	0.20 (0.01)	0.21 (0.01)	0.22 (0.00)
MU†	0.21 (0.01)	0.21 (0.01)	0.20 (0.00)	0.20 (0.01)	0.22 (0.01)	0.22 (0.00)
MA†	0.21 (0.01)	0.21 (0.01)	0.20 (0.00)	0.20 (0.01)	0.22 (0.01)	0.22 (0.00)
SR	0.21 (0.01)	0.21 (0.00)	0.20 (0.00)	0.20 (0.01)	0.22 (0.01)	0.22 (0.00)

	Study 79		**Study 81**		**Study 85**	
N	500 (1010)	200 (1310)	500 (291)	200 (591)	500 (322)	200 (622)
BWS	0.22 (0.00)	0.23 (0.01)	0.20 (0.01)	0.20 (0.01)	0.20 (0.01)	0.20 (0.01)
PMLE	0.22 (0.00)	0.23 (0.00)	0.20 (0.01)	0.20 (0.01)	0.20 (0.01)	0.20 (0.00)
MU†	0.22 (0.00)	0.22 (0.00)	0.20 (0.01)	0.20 (0.01)	0.19 (0.01)	0.20 (0.00)
MA†	0.22 (0.00)	0.22 (0.00)	0.20 (0.01)	0.20 (0.00)	0.19 (0.01)	0.19 (0.00)
SR	0.22 (0.00)	0.22 (0.00)	0.20 (0.01)	0.20 (0.01)	0.19 (0.01)	0.20 (0.00)

	Study 86		**Study 89**		**Study 90**	
N	500 (319)	200 (619)	500 (17)	200 (317)	500 (356)	200 (656)
BWS	0.22 (0.01)	0.23 (0.01)	0.24 (0.04)	0.23 (0.01)	0.24 (0.01)	0.25 (0.01)
PMLE	0.21 (0.01)	0.22 (0.01)	0.25 (0.04)	0.23 (0.01)	0.24 (0.01)	0.24 (0.00)
MU†	0.22 (0.01)	0.22 (0.01)	0.24 (0.04)	0.23 (0.01)	0.24 (0.01)	0.24 (0.00)
MA†	0.22 (0.01)	0.22 (0.00)	0.24 (0.04)	0.23 (0.01)	0.24 (0.01)	0.24 (0.00)
SR	0.22 (0.01)	0.22 (0.00)	0.24 (0.04)	0.23 (0.01)	0.24 (0.01)	0.24 (0.00)

	Study 91	
N	500 (256)	200 (556)
BWS	0.19 (0.01)	0.20 (0.01)
PMLE	0.19 (0.01)	0.20 (0.01)
MU†	0.19 (0.01)	0.20 (0.01)
MA†	0.19 (0.01)	0.20 (0.01)
SR	0.19 (0.01)	0.20 (0.01)

BWS = Model redevelopment (backward selection), PMLE = Model redevelopment (penalized Maximum Likelihood Estimation), MU = Model updating, MA = Model averaging, SR = Stacked Regressions

There are 2 validation samples for each scenario. The first validation sample is used for model redevelopment, model updating or model aggregation. The resulting prediction models are then evaluated in the second validation sample (sample size indicated between brackets). The study datasets have median event rate of 46% (range 38% to 65%).

† Literature models are updated by re-estimating the intercept and common slope.

Table 7.7: Performance comparison of derived prediction models in Application 2 in terms of achieved ranks.

Achieved ranks based on AUC										
$N = 500$	Rank 1		Rank 2		Rank 3		Rank 4		Rank 5	
BWS	58	(19%)	47	(16%)	50	(17%)	70	(23%)	75	(25%)
PMLE	75	(25%)	64	(21%)	28	(9%)	63	(21%)	70	(23%)
MU†	37	(12%)	41	(14%)	74	(25%)	44	(15%)	104	(35%)
MA†	51	(17%)	79	(26%)	67	(22%)	85	(28%)	18	(6%)
SR	79	(26%)	69	(23%)	81	(27%)	38	(13%)	33	(11%)
$N = 200$										
BWS	21	(7%)	15	(5%)	13	(4%)	82	(27%)	169	(56%)
PMLE	42	(14%)	22	(7%)	37	(12%)	124	(41%)	75	(25%)
MU†	44	(15%)	47	(16%)	109	(36%)	53	(18%)	47	(16%)
MA†	102	(34%)	117	(39%)	61	(20%)	17	(6%)	3	(1%)
SR	91	(30%)	99	(33%)	80	(27%)	24	(8%)	6	(2%)
Achieved ranks based on Brier Scores										
$N = 500$	Rank 1		Rank 2		Rank 3		Rank 4		Rank 5	
BWS	45	(15%)	54	(18%)	41	(14%)	53	(18%)	107	(36%)
PMLE	83	(28%)	55	(18%)	27	(9%)	90	(30%)	45	(15%)
MU†	28	(9%)	42	(14%)	75	(25%)	49	(16%)	106	(35%)
MA†	43	(14%)	94	(31%)	72	(24%)	75	(25%)	16	(5%)
SR	101	(34%)	55	(18%)	85	(28%)	33	(11%)	26	(9%)
$N = 200$										
BWS	17	(6%)	6	(2%)	14	(5%)	60	(20%)	203	(68%)
PMLE	46	(15%)	26	(9%)	33	(11%)	151	(50%)	44	(15%)
MU†	34	(11%)	62	(21%)	104	(35%)	57	(19%)	43	(14%)
MA†	98	(33%)	115	(38%)	67	(22%)	16	(5%)	4	(1%)
SR	105	(35%)	91	(30%)	82	(27%)	16	(5%)	6	(2%)

BWS = Model redevelopment (backward selection), PMLE = Model redevelopment (penalized Maximum Likelihood Estimation), MU = Model updating, MA = Model averaging, SR = Stacked Regressions

Results are based on 300 scenarios (30 random split-sample validation samples for 10 studies), in which the performance of each model is ranked from best (rank 1) to worst (rank 5). The totals (and percentage) of achieved ranks are summarized for each approach. Note that percentages may not always sum up to 100% due to rounding.

† Literature models are updated by re-estimating the intercept and common slope.

EDH (0 = no epidural hematoma, 1 = epidural hematoma), *i_tsah* (0 = no traumatic subarachnoid hemorrhage, 1 = traumatic subarachnoid hemorrhage), *i_hypoxia* (0 = no hypoxia, 1 = hypoxia), *i_hypots* (0 = no hypotension, 1 = hypotension), *i_dsysbp* (systolic blood pressure), *i_hb* (hemoglobin level), *i_glucos* (glucose level) and *i_sodium* (sodium level). The included datasets have median event rate of 46% (range 38% to 65%). Results in Table 7.5, 7.6 and 7.7 again demonstrate that prediction models developed with stepwise reduction have the poorest discriminative ability and calibration, and that Model Averaging and Stacked Regression consistently yield superior prediction models.

SIMULATION STUDY

Finally, we perform a simulation study to evaluate the performance of the aggregation methods in the validation population. Particularly, we investigate the influence of the validation sample size and the presence of heterogeneity between the populations of the literature models and the validation sample. We use a logistic regression model that serves as reference to generate data for the validation samples. Here, the outcome y_i for subject i with characteristics $(x_1)_i, \ldots, (x_{10})_i$ is given as:

$$\begin{aligned} y_i \sim & \,\text{Bernoulli}(\pi_i) \\ \text{logit}(\pi_i) = & \,\beta_0 + \beta_1(x_1)_i + \beta_2(x_2)_i + \beta_3(x_3)_i + \beta_4(x_4)_i + \beta_5(x_5)_i + \beta_6(x_6)_i \\ & + \beta_7(x_7)_i + \beta_8(x_8)_i + \beta_9(x_9)_i + \beta_{10}(x_{10})_i \end{aligned} \tag{Model A}$$

where $\beta_0 = -3$ and $\beta_1 = \beta_2 = \ldots = \beta_6 = 1$ and $\beta_7 = \beta_8 = \beta_9 = \beta_{10} = 0$. The 10 covariates are independent and each taking the values -1, 0 and 1 with probability 1/3. The corresponding outcome prevalence is 16%.

For each scenario, we generate 5 literature models and two validation samples. The literature models are developed using logistic regression with backward selection from a literature sample that contains 20 events. To ensure the generation of stable literature models, we draw literature samples until model convergence is reached and the error variance of estimated literature coefficients is within acceptable ranges ($<$ 50 for the model intercept and $<$ 10 for the predictor effects). Furthermore, to ensure that the populations of the literature models are different but related to the validation population, we use different reference models for generating the literature samples (that are based on the original reference model; see scenario A, B and C). Note that it is possible that generated literature models $j = 1, \ldots, 5$ may have $(\beta_7)_j, \ldots, (\beta_{10})_j \neq 0$ due to small sample bias or confounding (scenario C), or $(\beta_1)_j, \ldots, (\beta_6)_j = 0$ due to selection procedures. All validation samples are generated directly from the original reference model. The first validation sample (with 14–114 events) is then used to apply the previously described methods (model redevelopment using backward selection, model redevelopment using PMLE, model updating using logistic calibration, Model Averaging, Stacked Regressions using the Maximum Likelihood Estimator). Conversely, the second validation sample (with 1 000 events) is used for assessing the performance of the developed, updated and aggregated models in the original patient population (as defined by the reference model).

Heterogeneous baseline risk across literature models (Scenario A)

We first consider the scenario in which the literature models have a heterogeneous baseline risk. This is achieved by generating the literature samples from a deviation of the reference model, where the intercept takes a random value per study from a normal distribution with mean β_0 and standard deviation 0.50. This implies that the outcome y_{ij} for subject i in study j is generated according to:

$$\begin{aligned} y_{ij} &\sim \text{Bernoulli}(\pi_{ij}) \\ \pi_{ij} &= \text{logit}^{-1}\left((\beta_0)_j + (x_1)_i + (x_2)_i + (x_3)_i + (x_4)_i + (x_5)_i + (x_6)_i\right) \\ (\beta_0)_j &\sim \mathcal{N}(-3, 0.50) \end{aligned} \tag{Model B}$$

The resulting interquartile range (IQR) for the distribution of the intercept term of the literature models $(\beta_0)_j$ is -3.35 to -2.66.

Heterogeneous baseline risk and predictor effects across literature models (Scenario B)

In this second scenario, we consider the situation in which heterogeneity occurs in the baseline risk and common slope of the literature models. To this purpose, the slopes from Model A are multiplied by a factor H which is sampled from a Gamma distribution (to ensure that the reference coefficients remain positive) with scale 0.5 and rate 0.5. This implies that the outcome y_{ij} for subject i in study j is generated according to:

$$
\begin{aligned}
y_{ij} &\sim \text{Bernoulli}(\pi_{ij}) \\
\pi_{ij} &= \text{logit}^{-1}\left((\beta_0)_j + H_j\left[(x_1)_i + (x_2)_i + (x_3)_i + (x_4)_i + (x_5)_i + (x_6)_i\right]\right) \\
(\beta_0)_j &\sim \mathcal{N}(-3, 0.50) \\
H_j &\sim \Gamma(0.5, 0.5)
\end{aligned}
\tag{Model C}
$$

The resulting IQR for the distribution of the predictor heterogeneity factor H_j is 0.67 to 1.27. This implies that the overall strength of the regression coefficients in the population of study j will be too strong ($H_j > 1$) or too weak ($H_j < 1$) in comparison to the validation sample.

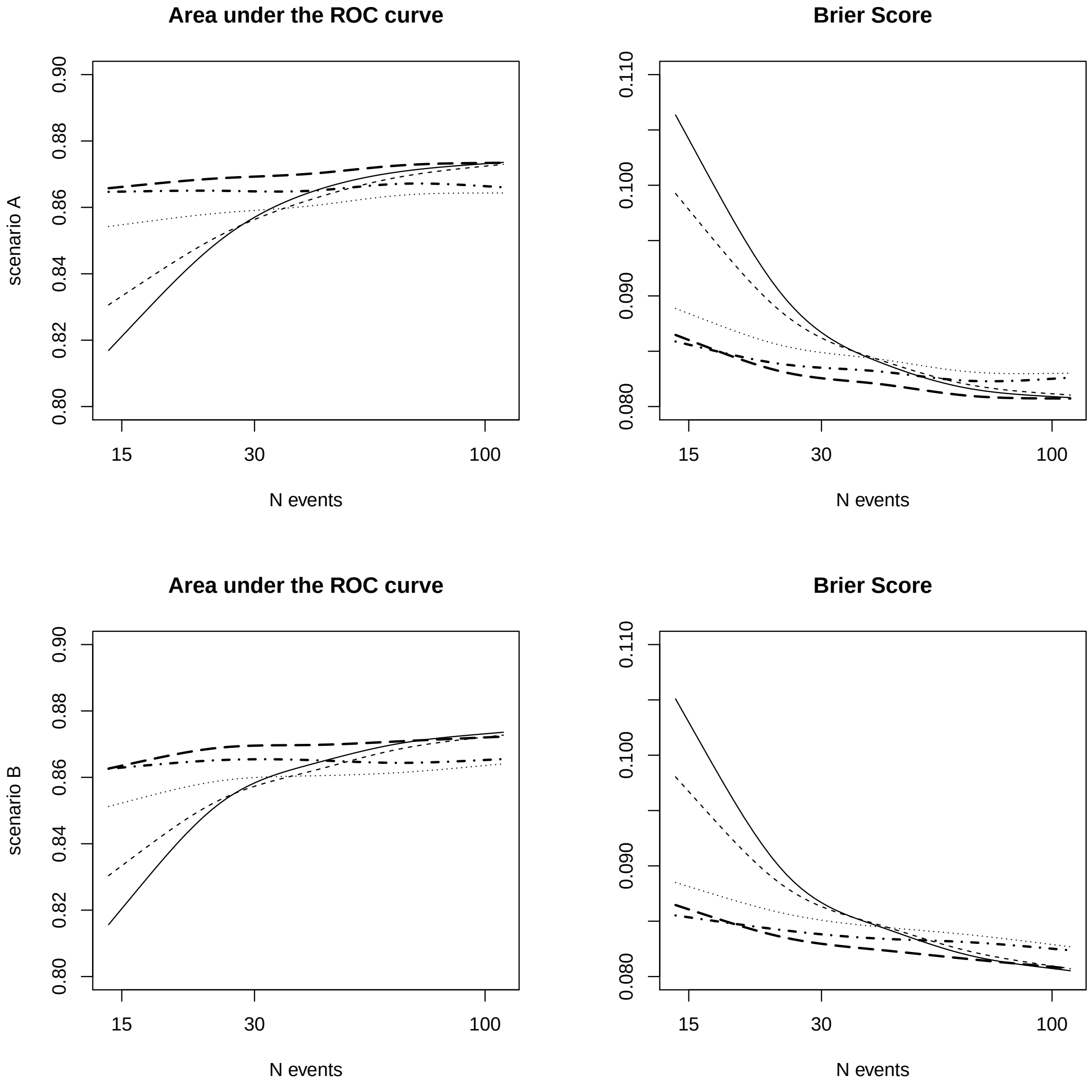
Non-accomodated heterogeneity (Scenario C)

In this third scenario, we consider the situation in which heterogeneity occurs in the literature models that is not accomodated for by the updating strategies. To this purpose, we extend the previous scenarios by confounding the literature models with two extraneous variables x_7 and x_8. Whereas in previous scenarios the generated literature models would, on average, yield similar regression coefficients as the reference model, confounding allows them to be systematically heterogeneous with the validation samples. Here, we introduce moderate confounding by amending Model C with two additional factors β_7 and β_8 that are different from zero.

$$
\begin{aligned}
y_{ij} &\sim \text{Bernoulli}(\pi_{ij}) \\
\pi_{ij} &= \text{logit}^{-1}\Big((\beta_0)_j + H_j\left[(x_1)_i + (x_2)_i + (x_3)_i + (x_4)_i + (x_5)_i + (x_6)_i\right] \\
&\qquad + (\beta_7)_j\,(x_7)_i + (\beta_8)_j\,(x_8)_i\Big) \\
(\beta_0)_j &\sim \mathcal{N}(-3, 0.50) \\
(\beta_7)_j, (\beta_8)_j &\sim \mathcal{N}(0.5, 0.2) \\
H_j &\sim \Gamma(0.5, 0.5)
\end{aligned}
\tag{Model D}
$$

In this model, each literature model $j = 1, \ldots, 5$ is affected by two confounders $(\beta_7)_j$ and $(\beta_8)_j$. Note that it is possible to accomodate for this heterogeneity by adjusting the updating strategy to re-estimate β_7 and β_8 individually (in addition to re-estimating the model's intercept and common slope). However, we here assume that the confounding effect goes unnoticed and evaluate whether the aggregation approaches are robust against the presence of unsuspected heterogeneity.

Figure 7.4: Results from the simulation study (Scenario A and B).



Results from the simulation study for 2 scenarios: (A) heterogeneous baseline risk across literature models and (B) heterogeneous baseline risk and predictor effects across literature models (common predictor variables). The following approaches are evaluated: redevelopment using backward selection (**solid line**), redevelopment using penalised maximum likelihood (**dash**), model updating of intercept and common slope (**dot**), Model Averaging (**dash-dot**) and Stacked Regressions (**long dash**).

Results

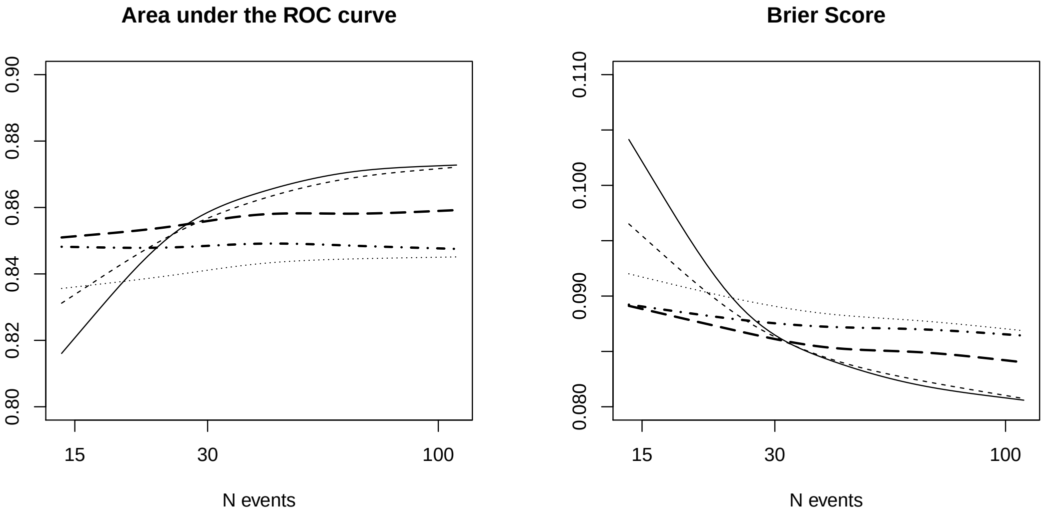
Results in Figure 7.4 indicate that model aggregation (i.e. Model Averaging and Stacked Regressions) outperforms traditional modeling techniques in validation samples (i.e. model redevelopment and model updating) particularly when few data are at hand. For instance, when literature models have a heterogeneous baseline risk (scenario A as described in section 7) and the validation sample contains 15 outcome events, stacked Regressions achieved an AUC of 0.87, versus 0.82 (model redevelopment using backward selection), 0.83 (model redevelopment using penalization) and 0.86 (model updating). When relatively much data were at hand (100 events or more), model redevelopment performed similar to Stacked Regressions and slightly outperformed Model Averaging and model updating (in terms of AUC and BS). Similarly, when literature models have a heterogeneous baseline risk and predictor effects (scenario B as described in section 7), Stacked Regressions achieved the best AUC and BS irrespective of the validation sample size. For instance, when the validation sample contains 15 events this approach achieved an AUC of 0.86 versus 0.82 (model redevelopment using backward selection), 0.83 (model redevelopment using penalization) and 0.85 (model updating). Again, model redevelopment achieved optimal performance for large sample sizes (100 events or more), where it slightly outperformed Model Averging and model updating. Finally, when literature models have a heterogeneous baseline risk plus overall slope, and effect modification occurs (scenario C as described in section 7), aggregation or updating of literature models is only advantageous for small sample sizes (Figure 7.5). Particularly, when the validation sample contains 30 events or less, Stacked Regressions and Model Averaging outperformed traditional modeling techniques in terms of AUC and BS respectively. For larger sample sizes, we noticed that model redevelopment techniques perform similarly and clearly outperform aggregation and updating approaches in terms of AUC and BS.

In general, our simulation studies indicate that model aggregation always outperforms model updating, and that model redevelopment is only useful when literature models are too heterogeneous with the validation sample to combine (i.e. differences beyond intercept and common slope) and sufficient data are available.

DISCUSSION

Here we have shown that a novel model validation and updating paradigm involving aggregation or meta-analysis of existing evidence effectively adjusts the resultant models to new circumstances, especially in situations of few validation data. Because this paradigm augments newly collected participant data with relevant evidence from published prediction models, it is likely that resulting meta-models are less prone to over-optimism and more generalizable towards new patient populations or settings. Aggregation can therefore help resolve the tendency towards development and reporting of numerous prediction models with similar goals or for the same clinical problem, and improve the cost-effectiveness of prediction research by making better use of prior evidence.

Figure 7.5: Results from the simulation study (Scenario C).



Results from the simulation study for scenario (C) heterogeneous baseline risk and predictor effects across literature models (different predictor variables). The following approaches are evaluated: redevelopment using backward selection (**solid line**), redevelopment using penalised maximum likelihood (**dash**), model updating of intercept and common slope (**dot**), Model Averaging (**dash-dot**) and Stacked Regressions (**long dash**).

Methods to integrate prior evidence in prediction modelling are currently lacking, and the few available (updating) strategies merely adjust previously published prediction models to new (validation) data [15, 16, 132, 167, 169, 239, 270]. We posited that aggregation over all available evidence, rather than selective updating, may further improve the performance of existing or even novel developed prediction models. For this reason, we explored two aggregation techniques: Model Averaging and Stacked Regressions, that combine the predictions from updated literature models by means of a weighted average. Whereas Model Averaging achieves aggregation in three consecutive stages (i.e. updating the literature models, calculating the model weights and estimating the single aggregated model), Stacked Regressions updates, weights and estimates the meta-model simultaneously. We further demonstrated that Stacked Regressions is the most efficient approach for developing a meta-model, as it involves a minimal amount of unknown parameters whilst accomodating for potential between-study heterogeneity and avoiding over-optimism. Its underlying mechanisms are similar to logistic calibration, and allow the disposition of existing models with poor performance. In contrast to previously proposed techniques [66], the aggregation methods do not require existing models to have similar predictors and can be applied when relatively few data are at hand.

Results from example and simulation studies applying these approaches demonstrate that model aggregation consistently outperforms traditional modeling techniques (such as model redevelopment or model updating) when few data are available. Model redevelopment was only useful when literature models were too heterogeneous with the local circumstances and sufficient patient data were at hand. Furthermore, model updating was almost always outperformed by model aggregation or model redevelopment. These findings are in line with previous research [132, 168, 239, 276], and further expose the advantages of incorporating prior evidence when few data is available [65, 66, 109, 205, 243, 245].Although it is beyond the remit of this paper, these techniques could conceivably also be used to design smaller prediction modeling studies when previous literature models are available. Further research would be necessary before these approaches could be used in the study design phase.

Similarly, future research investigating how a quality appraisal may identify candidate prediction models from the literature, and how variable selection and shrinkage may be achieved during model aggregation would further improve upon this work. Alternative weighting schemes (for Model Averaging) or Bayesian estimation procedures (for Stacked Regressions) may also further improve the predictive performance of aggregated meta-models. It is, for instance, possible to weight literature models based on corresponding similarities between the development and validation samples (similar to risk of bias tools in the meta-analysis of RCTs). The implementation of Model Averaging here tends to produce extreme weights being assigned to models, leading to a degree of skewing towards stronger models. Although we penalized stronger models when they involved extensive updating strategies, extreme weights remained present for most considered scenarios. These results confirm the conclusions of recent studies positing that Model Averaging essentially represents an informative strategy of model selection [163]. More promising approaches may therefore employ Bayesian Model Combination techniques where linear (or even non-linear) model combinations are estimated [71, 163]. This strategy is adopted here by Stacked Regressions, where linear combinations of model predictions are considered in a frequentist framework. Further research is needed to investigate how estimation of such combinations may account for updating complexity and penalize the contribution of models that perform well in the validation sample because they are strongly adjusted towards these data.

The aggregation of existing models may not be desirable when strong heterogeneity is expected [66]. Although it is possible to overcome this challenge by implementing more advanced updating procedures (e.g. model revision) prior to (Model Averaging) or during (Stacked Regressions) aggregation, such approach evidently requires additional validation data to test the so developed meta-model.Critically appraising prior evidence during the systematic review process can help to expose this and guide the selection of appropriate updating and/or aggregation approaches [34]. The complexity of the updating procedure could also explicitly be accounted for through penalization of the contribution of literature models when they were updated in the validation sample at hand. Aggregation remains problematic when the validation sample does not contain (important) predictor variables that are included in the literature models. Possible solutions may ignore the missing predictor or re-estimate the remaining predictors in the validation sample at hand. Whilst this was evidently not the case in the case studies and simulations we presented here, these strategies could further improve upon the approaches.

In conclusion, model aggregation is a promising approach to combine previously published prediction models and adjust them to a new patient population. The resulting meta-models show better performance than those generated through entire new model redevelopment or model updating strategies. We recommend the use of aggregation techniques when validation samples are relatively small and sufficient useful models (as identified from a critical appraisal of the literature) are available. In scenarios where large amounts of data are at hand and patient populations from the literature models are too heterogeneous with the validation population, developing a novel model may still be the best strategy.

ACKNOWLEDGEMENTS

We would like to thank Stan Buckens for his input and comments during the preparation of the manuscript. We also thank those researchers who agreed to share their individual participant data from the International Mission for Prognosis and Analysis of Clinical Trials (IMPACT) project to facilitate this article. We further gratefully acknowledge the following investigators for sharing of individual participant data from the deep vein thrombosis (DVT) studies: A.J. Ten-Cate-Hoek, R. Oudega, K.G.M. Moons, R.E.G. Schutgens, R.A. Kraaijenhagen, D.B. Toll, D.R. Anderson, P.S. Wells and J.L. Elf.

Part IV

General Discussion

8

A framework to interpret external validation results of clinical prediction models

Submitted

Thomas P. A. Debray
Yvonne Vergouwe
Hendrik Koffijberg
Daan Nieboer
Ewout W. Steyerberg
Karel G. M. Moons

Chapter 8

"An idea is always a generalization, and generalization is a property of thinking. To generalize means to think."

– Georg Wilhelm Friedrich Hegel

Clinical prediction models are commonly developed to facilitate prognostic or diagnostic probability estimations in daily medical practice [101, 170, 237, 247]. Such models are typically developed by (statistically) associating multiple predictors with outcome data from a so-called derivation or development sample [101, 170, 171, 237]. Well known examples are the Wells models for diagnosing Deep Venous Thrombosis [283], the Gail model for prediction of breast cancer incidence [82] and the Framingham risk scores [289]. As almost every prediction model is developed to be applied to new individuals, the value of a prediction model depends on its performance outside the development sample [15, 16, 136, 169, 192]. This implies that its predictive mechanisms should remain sufficiently accurate across new samples from the same target population (rather reflecting a model's reproducibility) or even from different but related target populations (rather reflecting a model's transportability) [136, 169, 273]. Roughly speaking, good model performance in the same or different target populations may both reflect a model's generalizability.

The generalizability of prediction models is commonly assessed in so-called (external) validation samples or studies [15, 16, 101, 136, 167, 169, 171, 237, 247]. Such studies aim to quantify the predictive performance of a previously developed model in individuals that were not used to develop the model [15, 16, 136, 169]. In practice, validation studies may range from temporal, to geographical, to validations across different medical settings or domains with increasingly different case mix and discrepancies in predictor and outcome definitions, and thus, in case of good model performance in the validation sample, increasing potential of transportability of the model [15, 16, 169, 259].

It is debatable to what extent a model should be transportable across different but related patient populations. For instance, it may not be desirable to transport a prediction model to a population where outcomes are defined differently, unless the model can be adjusted fairly straightforward. Transportability therefore relates to which extent differences in target populations do not meaningfully affect model performance. Because validation studies typically differ from development samples, some quantification of their relatedness may be useful to better interpret model validation results and to identify the actual population in which a prediction model can successfully be implemented.

Unfortunately, explicit criteria for inferring on differences in case mix between the development and validation sample remain ill-defined, let alone that there are methods to formally quantify these differences. As a consequence, it is often unclear to what extent good performance in the validation sample implies rather reproducibility of the model across samples of the same target population of the development sample, or truly transportability of the model across different but related populations [273]. This, in our view, impedes transparent interpretation of results from external validation studies [167].

Based on previous recommendations [15, 16, 136, 169, 192, 247], we here describe a framework of methodological steps and address statistical methods for analyzing and interpreting the results of external validation studies. We illustrate this framework with an empirical example of prediction models for the diagnosis of Deep Venous Thrombosis developed by logistic regression modeling. The framework may, however, mutatis mutandis be applied to prediction models developed by e.g. survival modeling or even linear regression modeling. Consequently, we aim to improve the inference making of studies aimed at testing of prediction models in samples of new individuals and thus inferences of model generalizability. The framework is intended to ultimately facilitate faster and wider implementation of genuinely useful models and allow a speedier identification of models that are of limited value [228].

EMPIRICAL EXAMPLE DATA

Deep Vein Thrombosis (DVT) is a blood clot that forms in a vein in the body and may lead to blockage in the lungs, preventing oxygenation of the blood and potentially causing death. Multivariable diagnostic prediction models have been developed during the past decades to safely exclude DVT without having to refer for further testing. Physicians may doubt whether or when to use such a diagnostic prediction model if their patient(s) represent a clinically relevant subgroup, such as the elderly or patients with a history of DVT and/or pulmonary embolism[261]. For this paper, we hypothesize that it is yet unclear to what extent these models are generalizable across samples of the same or different (but related) target populations, because the performance of a prediction model may change according to characteristics of the patients or clinical setting (e.g. primary or secondary care).

To illustrate our framework, we used the individual participant data (IPD) from four datasets (Table 8.1 and Table A.1 in the Appendix) to develop and externally test (validate) a multivariable diagnostic model for predicting the presence of DVT. Specifically, we used one dataset ($n = 1\,295$) to develop a logistic regression model with 7 patient characteristics and the D-dimer test result (Table 8.2). Afterwards, we externally validated this model in the three remaining datasets ($n_1 = 791$, $n_2 = 1\,028$ and $n_3 = 1\,756$).

Table 8.1: Baseline Table for 4 Primary Care DVT Datasets.

	Development	Val. 1	Val. 2	Val. 3
Line of care	**primary**	primary	primary	secondary
N	**1 295**	791	1 028	1 756
Incidence DVT	**22%**	16%	13%	23%
Male gender	**36%**	38%	37%	37%
Oral contraceptive use	**10%**	10%	10%	5%
Presence of malignancy	**6%**	5%	5%	13%
Recent surgery	**14%**	13%	8%	11%
Absence of leg trauma	**85%**	82%	72%	85%
Vein distension	**20%**	20%	15%	16%
Calf difference $\geq$ 3cm	**43%**	41%	30%	24%
D-dimer abnormal	**70%**	72%	46%	52%

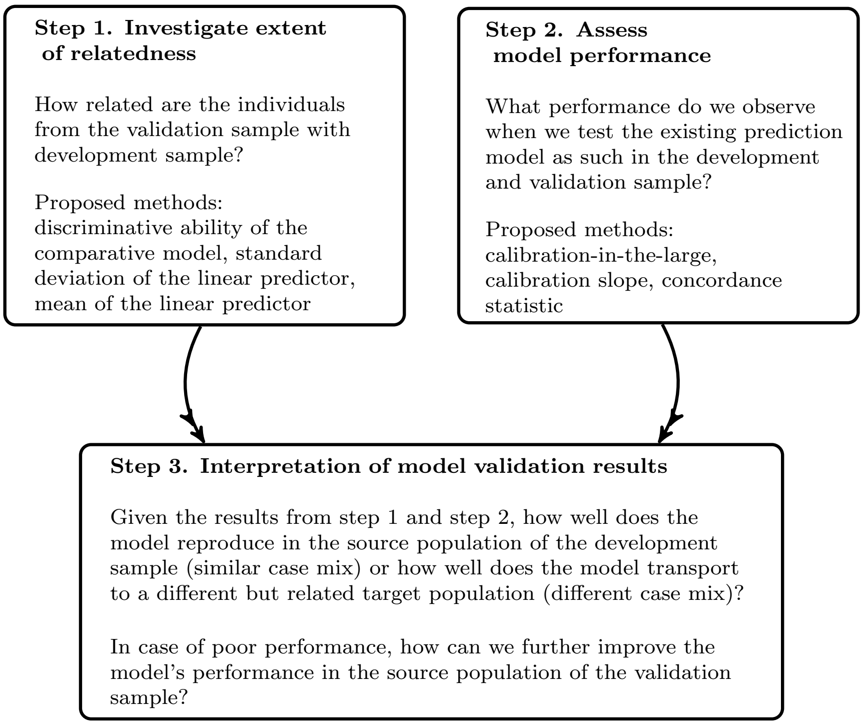
METHODS

Below, we describe a framework of three methodological steps (Figure8.1) and address statistical methods for analyzing and interpreting the results of external validation studies. In the first step, we quantify to what extent the development and validation sample are related. The second step assesses the model's predictive accuracy in the development and validation sample to identify the extent to which its predictive mechanisms differ or remain accurate and valid. The final step interprets the model's performance in the validation sample in terms of reproducibility or transportability. This step also indicates what type of revisions to the model, based on the validation sample, may be necessary. We describe a straightforward analytical and judgmental implementation for each step, and illustrate the approach using the diagnostic prediction model from our case study.

Step 1: Investigate relatedness of development and validation sample

This first step aims to quantify to what extent the development and validation sample are related or different. Two samples can have any degree of relatedness ranging from 'identical' to 'not related at all' [15, 16, 136]. Different but related samples are located between these extremes, and determination of their (relative) position is essential for interpreting the results of a model validation study and make inferences on the generalizability of the model. Typically, two (or more) samples increasingly differ when their subject characteristics (case mix), outcome occurrence or the predictor effects (regression coefficients) differ across the corresponding populations [15, 68,

Figure 8.1: Proposed approach to external validation studies



Step 1. Investigate extent of relatedness

How related are the individuals from the validation sample with development sample?

Proposed methods: discriminative ability of the comparative model, standard deviation of the linear predictor, mean of the linear predictor

Step 2. Assess model performance

What performance do we observe when we test the existing prediction model as such in the development and validation sample?

Proposed methods: calibration-in-the-large, calibration slope, concordance statistic

Step 3. Interpretation of model validation results

Given the results from step 1 and step 2, how well does the model reproduce in the source population of the development sample (similar case mix) or how well does the model transport to a different but related target population (different case mix)?

In case of poor performance, how can we further improve the model's performance in the source population of the validation sample?

Typical validation studies are restricted to step 2: 'Assess model performance'.

259, 273, 275]. Consequently, it seems useful to evaluate the extent to which the development and validation sample have (1) a similar case mix and (2) share common predictor effects.

A straightforward approach for evaluating relatedness of case mix may evaluate differences between the subject characteristics of the development and validation sample separately, and compare their ranges [15, 16, 54, 192]. Although this approach is useful for comparing the overall case mixes, it does not take the interrelation of subject characteristics (per sample) into account. For instance, Table 8.1 reveals that the development sample and Validation Study 1 have a very similar case mix of predictor variables, but a different outcome occurrence (22% versus 16%). In Validation Study 3, however, the outcome occurrences are similar but the case mix considerably differs. It is not directly clear which of the validation samples is now more similar to the development sample, and would lead to small or larger change in the predictive performance of the model as compared to the performance found in the development set.

The heterogeneity of predictor-outcome associations between the development and validation sample can also be evaluated by, for example, refitting the original model in the validation sample (Table8.2). Unfortunately, also for this approach it is not directly clear how to summarize differences in estimated regression coefficients (or corresponding adjusted odds ratios) as heterogeneity between the underlying target populations.

Given the pitfalls of the aforementioned direct comparisons, we here propose two statistical approaches that calculate an overall measure of (dis)similarity between the development and validation sample. The first approach calculates a summary measure of relatedness based on how well individuals from both samples can be distinguished. Conversely, the second approach assesses to which extent the risk distributions of the development and validation sample diverge.

Table 8.2: Estimated Regression Coefficients (SE) for 4 Primary Care DVT Datasets.

	Dvl.	Val. 1	Val. 2	Val. 3
N	**1 295**	791	1 028	1 756
Constant (model intercept)	**-5.0 (0.4)**	-6.7 (1.1)	-4.7 (0.4)	-4.5 (0.3)
Male gender	**0.7 (0.2)**	0.4 (0.2)	0.6 (0.2)	0.5 (0.1)
Oral contraceptive use	**0.8 (0.3)**	0.5 (0.4)	-7.0 (58.6)	0.5 (0.3)
Presence of malignancy	**0.5 (0.3)**	-0.1 (0.4)	0.7 (0.4)	0.3 (0.2)
Recent surgery	**0.4 (0.2)**	0.6 (0.3)	0.0 (0.4)	0.5 (0.2)
Absence of leg trauma	**0.7 (0.2)**	0.81 (0.3)	0.6 (0.3)	0.3 (0.2)
Vein distension	**0.5 (0.2)**	0.2 (0.3)	0.2 (0.3)	0.6 (0.2)
Calf difference $\geq$ 3cm	**1.2 (0.2)**	0.9 (0.2)	0.9 (0.2)	1.4 (0.1)
D-dimer abnormal	**2.4 (0.3)**	4.0 (1.0)	2.4 (0.3)	3.0 (0.2)

The linear predictor for a subject (given by the model from the development sample) is as follows -5.02 + 0.71×male gender + 0.76×OC use + 0.50×presence of malignancy + 0.42×recent surgery + 0.67×absence of leg trauma + 0.53×vein distension + 1.15×calf difference $\geq$ 3cm + 2.43×abnormal D-dimer. The probability (or risk) of Deep Vein Thrombosis for the same subject is given by $1/\left(1+\exp(-\text{linear predictor})\right)$

Comparative model. The relatedness between two samples is typically tested by assuming an underlying distribution of subject characteristics. For instance, we may assume that individuals from the development sample follow a multivariate normal distribution. This strategy is, however, often undesirable because it cannot adequately account for dichotomous or non-linear variables. For this reason, we propose to evaluate the extent to which individuals from the development and validation sample can be distinguished. This approach is a generalization of Hotelling's T^2 which is related to discriminant analysis and the Mahalanobis distance metric [180]. This can be achieved fairly straightforward by estimating a binary logistic regression model, also labelled comparative model, to predict the probability that an individual belongs to the development sample. The comparative model should at least consider the predictor and outcome variables of the prediction model that is being validated. It may be clear that if the comparative model discriminates poorly (or well), both samples are strongly (or not much) related in terms of the considered predictor variables and outcome status. The discriminative ability can be quantified using measures such as the concordance (c) statistic.

Distribution of the linear predictor. It is also possible to directly compare the distribution of the model's predicted risks in the development and validation sample [237, 273]. This can be achieved by calculating the spread, here defined as the standard deviation (SD), and the mean of the linear predictor (LP) of the original model in the development and validation sample. Because the LP is the logit transformation of the predicted risks in logistic regression, its interpretation is fairly straightforward. In general, an increased (or decreased) population variability of the LP indicates more (or less) heterogeneity of case mix. As the case mix heterogeneity increases, individuals have a larger variety of patient characteristics and the model tends to discriminate better [273]. Specifically, the discriminative ability may improve (or deteriorate) when the SD of the LP increases (or decreases) because individual risk estimates become more (or less) separable. Conversely, differences in mean of the LP between the development and validation sample reflect the difference in overall (predicted) outcome frequency – i.e. in fact a reflection of case mix severity – and may therefore reveal the model's calibration-in-the-large in the validation sample [96].

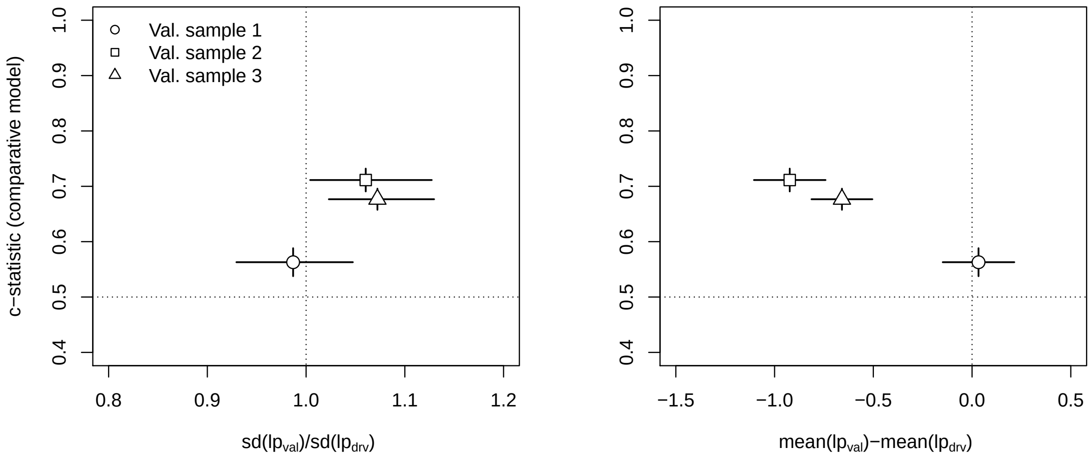
Empirical example. Results from the empirical example (Figure 8.2) demonstrate that the comparative model and the distribution of the linear predictor generally lead to similar conclusions. Specifically, we found that it was difficult to distinguish between individuals from the development sample and Validation Study 1 (c = 0.56 with 95% CI of 0.54; 0.59). These samples also had a similar SD of the LP (1.45 vs. 1.47) and similar mean of the LP (-1.72 vs. -1.75). These results indicate that the development sample and Validation Study 1 had a similar distribution of case mix, and we can expect similar model performance in both samples.

For Validation Study 2 and 3, we found an increased spread of the LP and a decreased average of the LP. The comparative models indicated that individuals from the development and validation sample could be distinguished more easily and that their case mix was indeed much less related to the case mix of the development sample ($c = 0.71$ and $c = 0.68$ respectively).

Step 2: Assessment of the model's performance in the validation study

In this second step we evaluate the originally developed model's performance in the validation sample. This is typically quantified in terms of calibration and discrimination [83, 101, 237]. Calibration reflects the extent to which the predicted probabilities and actual probabilities agree,

Figure 8.2: Results from step 1 in the empirical example.



Results of analyzing the validation sample (median with 95% CI) and validating the prediction model. The Y-axis reflects the extent to which the validation sample is different but related to the development sample (as indicated by the *c*-statistic of the comparative model). In the left graph, the X-axis reflects the potential for good performance indicated by the relative difference in SD of the linear predictor. In the right graph, the X-axis reflects the difference between the means of the linear predictors.

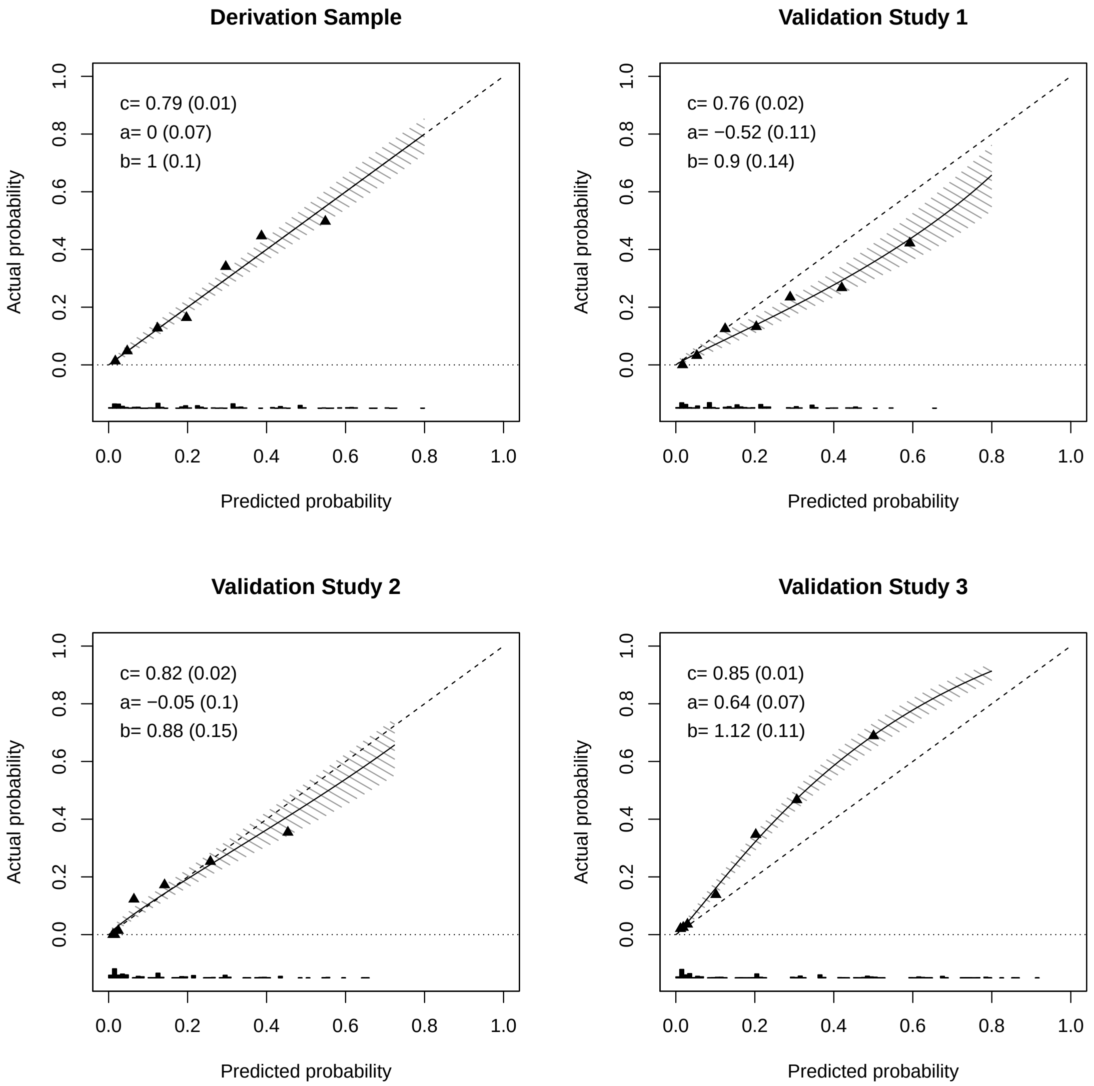
whereas discrimination is the ability to distinguish high-risk patients from low-risk patients. Here, we focus on the calibration-in-the-large plus calibration slope, and the concordance (c)-statistic as summary measures of calibration and discrimination respectively [57, 77, 116, 160, 237, 249]. These statistics are ideally calculated in validation samples obtained from (prospective) cohort studies [170], although the calibration slope and c-statistic remain useful summary measures in retrospective study designs (e.g. case-control). The calibration slope can be used as a statistic for evaluating to which extent the model's predictive mechanisms remain valid in the validation sample. Finally, we recommend visual inspection of the calibration plot, where groups of predicted probabilities are plotted against actually observed outcomes and perfect predictions should be on the 45-degree line [249].

Calibration-in-the-large. This statistic quantifies whether the average of predictions corresponds with the average outcome frequency, and ideally equals 0. Values below (or above) this value indicate that the model overestimate (or respectively underestimate) the outcome presence. By definition, the calibration-in-the-large is always optimal (0) in the development sample of the prediction model. Consequently, it is a useful statistic for identifying whether unexplained differences exist in the outcome frequency of the validation sample, e.g. due to mechanisms not captured by the included predictors [16, 96, 279]. Note that calculation of the calibration-in-the-large statistic is only meaningful when observed outcome frequencies are representative for the target population. This implies that validation samples should ideally be obtained from (prospective) cohort studies [170].

Calibration slope. The calibration slope, denoted as b_{overall}, reflects whether predicted risks are appropriately scaled with respect to each other over the entire range of predicted probabilities ($b_{\text{overall}} = 1$) [60, 160]. Typically, $b_{\text{overall}} > 1$ occurs when predicted probabilities do not vary enough (e.g. predicted risks are systematically too low) and $0 < b_{\text{overall}} < 1$ occurs when they vary too much (e.g. predicted risks are too low for low outcome risks, and too high for high outcome risks). A poor calibration slope ($0 < b_{\text{overall}} < 1$) usually reflects overfitting of the model in the development sample, but may also indicate inconsistency of predictor effects between the development and validation sample [132, 239, 249, 270, 275]. It is therefore a useful measure of external validity.

Concordance statistic. The *c-statistic* represents the probability that individuals with the outcome receive a higher predicted probability than those without [57]. It corresponds to the area under the receiver operating characteristic (ROC) curve for binary outcomes, and can range from 0.5 (no discrimination) to 1.0 (perfect discrimination). Because the c-statistic reveals to what extent the prediction model can rank order the individuals according to the outcome in the validation sample, it is a useful tool for evaluating its discriminative value.

Figure 8.3: Results from step 2 in the empirical example.



Calibration plots with 95% confidence intervals of the developed multivariable prediction model when applied in the development and three validation samples. Perfect calibration is represented by the dotted line through the origin with slope equal to 1. We generated 7 quantile groups predicted probabilities, and illustrated their corresponding outcome proportion with a triangle. The following statistics are presented with their standard error: c = c-statistic, a = calibration-in-the-large, b = calibration slope

Empirical example. In Validation Study 1, we found that the discriminative ability of the developed model slightly decreased (Figure 8.3). Predicted risks were systematically too high (calibration-in-the-large=-0.52 with $P < 0.0001$) but remained proportionally accurate (calibration slope: 0.90). For Validation Study 2 and 3, we found an increased discriminative ability in the validation sample. This increase was expected from step 1 due to an increased spread of the linear predictor. Although the achieved calibration-in-the-large and calibration slope was reasonable for Validation Study 2, predicted risks were systematically too low and did not vary enough in Validation Study 3 (calibration slope: 1.12 with $P < 0.0001$).

Step 3: Interpretation of model validation results

In this final step, we describe how the model's predictive accuracy in the validation sample (step 2) can be interpreted by combining the results from step 1. We also indicate what may be done to further improve the model's performance in the source population of the validation sample in case of poor performance.

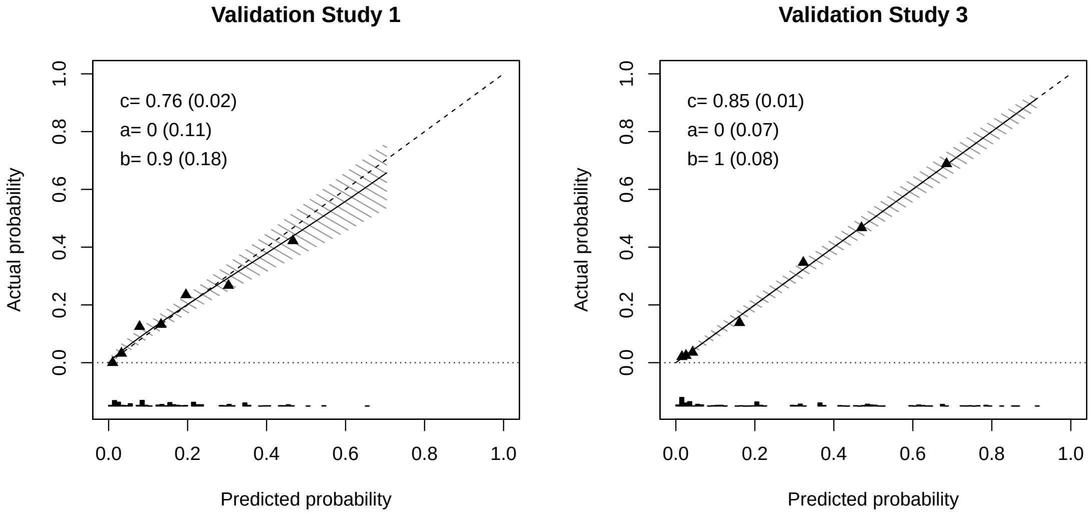
In step 1 we identified whether the reproducibility (similar case mix) or transportability (different case mix) of the prediction model is assessed by testing its performance in the validation sample. Step 1 also indicates whether the discriminative ability of the prediction model – as estimated in step 2 – changes due to differences in case mix heterogeneity, and whether the calibration-in-the-large deteriorates due to differences in case mix severity [273]. Step 2 directly indicates whether differences in case mix between the development and validation sample actually affect model performance in the latter. Consequently, the combined results from step 1 and step 2 indicate to what extent the model indeed seems generalizable (i.e. either in terms of reproducibility or transportability).

In case of poor predictive performance, several methods may improve the model's accuracy in the validation sample at hand. These updating methods may range from an intercept update, to adjustment or even re-estimation of individual regression coefficients or adding predictors [132, 133, 169, 237, 270]. Specifically, a poor calibration-in-the-large may be overcome by re-estimating its intercept in the validation sample (intercept update) [132, 133, 169, 237, 270]. Similarly, a poor calibration slope (e.g. due to overfitting) may be corrected by applying logistic calibration (i.e. overall adjustment of the calibration slope). On the other hand, when predictor effects are heterogeneous between the development and validation sample, and calibration plots show inconsistent predictions across the whole range of predicted probabilities, updating strategies becomes more difficult and may require the re-estimation and inclusion of additional predictors. In those scenarios, the validation study indicates that the model's predictive mechanisms may no longer be valid and a larger model revision or updating is needed.

Empirical example. In Validation Study 1, we can conclude that the model reproduces well because the development and validation sample were strongly related and model performance was concordantly adequate in the validation sample. The model may, however, be improved by an intercept update as predicted risks were systematically too high (Figure 8.4).

For Validation Study 2 and 3, we found substantial differences in the case mix between the development and validation sample. Because the model's discrimination improved in the validation sample and its calibration remained fairly adequate, its transportability to the target populations of the validation sample(s) appears reasonable. For Validation Study 3, however, some mis-calibration occurred such that the prediction model should be revised, e.g. by updating its intercept and common calibration slope in the validation sample (Figure 8.4).

Figure 8.4: Results from step 3 in the empirical example: calibration plots after recalibration in the validation sample.



Calibration plot of the multivariable prediction model, revised in validation study 1 (update of intercept) and validation study 3 (update of intercept and common calibration slope).
The following statistics are presented with their standard error: c = *c*-statistic, a = calibration-in-the-large, b = calibration slope

Chapter 8

DISCUSSION

Studies to quantify the performance of developed (existing) prediction models in other individuals (external validation studies) are important to assess the model's performance outside the development setting and to evaluate its extent of generalizability and usefulness [15, 16, 101, 136, 167, 169, 171, 228, 237, 247]. Moreover, such studies guide us in deciding upon updating or re-development strategies.[169, 237, 239, 270, 272] It is often unclear how model validation study results relate to the general concept of generalizability of the prediction model, and how researchers should interpret good or poor model performance in the validation sample. We presented a framework to better determine whether the external validation study rather assesses a model's reproducibility or its transportability. This framework extends the framework originally proposed by Altman and Royston[15], and Justice *et al*[136]. The latter already proposed to distinguish generalizability in reproducibility to the overall target population and transportability to different but related target populations.

We pose that external validation studies with stronger case mix differences as compared to the development sample increasingly address the model's transportability across populations. It is, however, possible that some well-developed models are not reproducible or transportable and may first require to varying extents model updating prior to actual implementation in another target population.

Some considerations have to be made. Firstly, development data may not always be available. This implies that accurate calculation of differences in case mix (step 1) may not directly be possible and impedes the interpretation of validation study results in step 3. Although it remains possible to evaluate case mix differences between subject characteristics on the average level by relying on published baseline tables, this approach does not take the interrelation of subject characteristics into account.

Secondly, we proposed using a comparative model (based on a generalization of Hotelling's T^2) and the distribution of the linear predictor to evaluate case mix differences between the development and validation sample in a single dimension. Although the comparative model explicitly accounts for differences in subject characteristics, outcome occurrence and their interrelation, and the distribution of the linear predictor merely compares risk distributions, both approaches tend to yield similar conclusions. The linear predictor may, however, be less useful for evaluating case-mix differences in survival data as it does not account for baseline survival. Conversely, the usefulness of the comparative model strongly depends upon its included variables and may be prone to overfitting. Other metrics for quantifying the relatedness between samples – such as the overlap coefficient [53, 162] or extensions of the Mahalanobis distance metric [64] – have not been evaluated here and may lead to different conclusions.

Thirdly, we noted that differences in characteristics outside the model, such as important predictors that were not recorded, may substantially influence a model's generalizability, as we found in Validation Study 3. These missed predictors could for instance explain differences in baseline risk, or interact with included predictor effects.[68, 96, 169, 279] As a consequence, interpretation of differences in case mix is not always straightforward and clinical expertise remains paramount in interpreting the results of a validation study.

Fourthly, we used the calibration-in-the-large, the calibration slope and the *c*-statistic as summary statistics for assessing model performance and interpreting generalizability. Other measures such as the case-mix corrected *c*-index may provide additional insights into model performance [273], and have not been evaluated here. Furthermore, by focusing on summary statistics of model calibration, precipitate conclusions about external validity may be reached. For instance, it is possible that the prediction model shows good calibration as a whole, but yields inaccurate predictions in specific risk categories. This, in turn, may affect the model's generalizability towards these risk categories. We therefore emphasize the graphing of calibration plots and visual inspection of these plots in addition to calculation of the calibration slope [249].

Finally, it is important to recognize that good performance of a model in another validation sample does not always correlate with its clinical usefulness. This is because external validity merely requires the model's predictive mechanisms to remain accurate across different samples, whereas clinical usefulness also refers to cost-effectiveness, clinical judgment and strongly depends on the context [15, 16, 167, 169, 249, 293].

Conclusions. The proposed methodological framework for prediction model validation studies may enhance the judgment to what extent the individuals in the validation study were different from the development sample, how it can be placed in view of other validation studies of the same model, and to what extent the generalizability of the model is studied.

ACKNOWLEDGEMENTS

We would like to thank Stan Buckens for his input and comments during the preparation of the manuscript. We also gratefully acknowledge the following investigators for sharing of individual participant data from the deep vein thrombosis (DVT) studies: A.J. Ten Cate-Hoek, R. Oudega, K.G.M. Moons, R.A. Kraaijenhagen and D.B. Toll.

9

Perspectives for future research

Chapter 9

“Take away paradox from the thinker and you have a professor.”

– Soren Kierkegaard

The methods presented in this thesis enable the principles of meta-analysis to be applied in clinical prediction research and represent improvements over existing meta-analytical approaches (developed for therapeutic research), tailored to the problem of prediction, either diagnostic and prognostic. These methods were previously lacking, and, as we have demonstrated in this thesis, may considerably improve the performance of previously developed or novel prediction models. Although we believe that meta-analysis deserves a fundamental role in clinical prediction research as well, several challenges still need to be addressed before it can successfully be implemented.

First, the presence of between-study heterogeneity in predictor-outcome associations strongly affects the validity and usefulness of a synthesis, meta-analytical model. Meta-analysis of any kind, so also for prediction (modeling) studies, may not be desirable when studies are too distantly related. This situation may arise when study populations differ too much and predictor-outcome associations strongly vary across studies. Consequently, it is important to investigate under which conditions data can be meaningfully combined, and how similarity of notably the predictor-outcome associations can be promoted. This thesis proposed several methods for quantifying the presence of between-study heterogeneity in predictor-outcome associations across studies (**Chapter 2**) and measuring the relatedness between two study samples (**Chapter 8**). It may therefore be valuable to investigate how predictor selection algorithms may be implemented when using these methods to identify a useful set of homogeneous predictor-outcome associations across the studies. In addition, some guidelines are needed to decide upon when to conduct a meta-analysis, and upon the implementation of a particular synthesis method in view of the quantified between-study heterogeneity in predictor-outcome associations.

Second, meta-analysis of previously published prediction models using only aggregate, reported data or results may not be desirable when no IPD is at hand. Particularly, we demonstrated that the availability of an IPD set allows the synthesis model to be updated towards a specific target population (**Chapter 6** and **Chapter 7**). Without any IPD set, any meta-analysis will produce a weighted average of regression coefficients that are not longer applicable to any of the original development populations. As a consequence, unfocused evidence synthesis, i.e. meta-analysis without some form of adjustment towards a specific target population, is likely to yield meaningless synthesis models that do not perform well when applied in any local circumstances. Fortunately, basic evidence such as the local outcome prevalence may already suffice to adjust synthesis models to a particular patient population when the model's predictor effects are homogeneous across studies (**Chapter 3**). Future research may investigate how synthesis models with heterogeneous predictor effects can effectively be adjusted for a specific population.

Third, the meta-analysis of aggregate data produced by models from different statistical families or with different structures is not straightforward. For instance, aggregate data for survival times may be derived from proportional hazard models (such as Poisson, Cox or exponential regression) or accelerated failure time models (such as Weibull regression), each leading to a different interpretation of the estimated baseline hazard and regression coefficients of predictors. Other, less related, families are also widely available and may consist of artificial neural networks, decision trees or even support vector machines. In the latter scenarios, pooling of estimated regression coefficients (**Chapter 6**) may no longer be possible and more advanced methods (**Chapter 7**) are required to effectively implement all existing models and combine their output after evaluation. It is therefore worthwhile to investigate how differently specified models can effectively be combined into a single explicit synthesis model that is straightforward to implement, again of course with the use of at least one IPD set.

Fourth, when implementing a clinical prediction model in a clinical decision support system (CDSS), subject characteristics and final diagnoses can continuously be recorded. This makes it feasible to gradually adjust a prediction model towards the local circumstances. Unfortunately, it is not clear how updating strategies should be implemented when datasets dynamically increase over time. It has been suggested that parsimonious updating strategies are preferred when relatively few data are at hand [239]. Although extensive updating strategies may further improve the model's performance in local circumstances, their implementation (usually) requires a substantial amount of data. A framework is therefore needed to identify whether early adaptation of a prediction model is justified, and to decide upon the extensiveness of updating methods. It may further be helpful to identify when a prediction model has sufficiently been validated, and subsequent updating is no longer required.

Finally, we demonstrated that the presence of missing data may considerably hamper the development of a meta-analytical prediction model. For instance, the imputation of systematically missing predictor variables is not straightforward as the choice of imputation method may considerably affect the degree of between-study heterogeneity in predictor-outcome associations (**Chapter 4**). Traditional imputation strategies ignore the clustering of participants within studies and may therefore support a self-fulfilling prophecy when assessing the degree of homogeneity in imputed predictor variables. More sophisticated imputation methods are required to appropriately account for missing data in an IPD meta-analysis and to evaluate the presence of between-study heterogeneity when predictor variables are systematically missing.

Appendix

Table A.1: Overview of Deep Venous Thrombosis datasets.

Study	**Line of care**	**Size**	**Chapter**							**Ref.**
			1	2	3*	4	6	7	8	
AMUSE-1	primary	1 028		✓	✓	✓	✓	✓	✓	[44]
AIDA	secondary	814		✓	✓	✓		✓		[223]
EDIT	secondary	153		✓	✓	✓				[18]
Kraaijenhagen	secondary	1 756	✓	✓	✓	✓		✓	✓	[140]
Toll	primary	532	✓	✓	✓	✓	✓	✓	✓	[260]
	primary	259	✓	✓	✓	✓	✓	✓	✓	[261]
EDITED	secondary	1 075	✓	✓	✓	✓		✓		[17]
Kearon (2001)	secondary	429		✓	✓	✓				[137]
Elf	secondary	325		✓	✓	✓		✓		[73]
Oudega	primary	1 295		✓	✓	✓		✓	✓	[182]
Stevens	secondary	436	✓	✓	✓	✓				[232]
DIEM	secondary	541	✓	✓	✓	✓				[282]
Bates	secondary	550		✓	✓	✓				[24]
Kearon (2005)	secondary	809		✓	✓	✓				[138]

* The Elf study in Chapter 5 contains 357 subjects. In addition, the Kearon and Bates studies were combined into a single dataset of 1 768 subjects.

Table A.2: Overview of Traumatic Brain Injury datasets.

Study	Country	Size	Chapter		Ref.
			6	7	
HIT I Nimodipine Trial	UK, FI	350	✓		[21]
UK Four Centre Study	UK	791	✓	✓	[174]
Traumatic Coma Data Bank	USA	603	✓		[80]
EBIC Survey	NL, BE, FR, SP, I, SZ, UK, DK, SW, G, FI, Y	822	✓	✓	[173]
HIT II Nimodipine Trial	NL, BE, FR, H, I, SZ, Y, AS, UK, SW, NW, G, FI	819	✓	✓	[256]
NABIS	USA	385	✓		-
PHARMOS	USA, NL, BE, FR, SP, IS, FI, AS, UK, DK, PL, G AU, TK, I	856	✓	✓	-
Tirilazad International Trial	NL, BE, FR, SP, I, SZ, UK, G, DK, SW, NW, AU, PR, FI, IS	1 118	✓	✓	[154]
Tirilazad Domestic Trial	USA, C	1 041	✓	✓	-
Selfotel International Trial	C, NL, BE, FR, SP,I, UK, DK, SW, NW, G, AG, AU	409	✓		[172]
Saphir/C-CPP-ene Trial	NL, BE, SP, I, SZ, UK, DK, SW, G, FI	919	✓	✓	-
PEGSOD Trial	USA, C, K, HK, IS	1 510	✓	✓	[294]
Bradycor/CP-0127 Trial	USA, C	126	✓		[153]
CSTAT	-	517	✓	✓	-
APOE	UK	756	✓	✓	-

AG = Argentina; AS = Austria; AU = Australia; BE = Belgium; C = Canada; DK = Denmark, FI = Finland; FR = France; G = Germany; H = Hungary; HK = Hong Kong; I = Italy; IS = Israel; K = Korea; NL = The Netherlands; NW = Norway; PL = Poland; PR = Portugal; SP = Spain; SW = Sweden; SZ = Switzerland; TK = Turkey; UK = United Kingdom; USA = United States of America; Y = Yugoslavia
An overview of these datasets is given in [152].

Bibliography

Bibliography

"The secret to creativity is knowing how to hide your sources."

– Albert Einstein

[1] Abbasi A, Peelen LM, Corpeleijn E, et al. Prediction models for risk of developing type 2 diabetes: systematic literature search and independent external validation study. *British Medical Journal.* 2012;345:e5900. doi:10.1136/bmj.e5900.

[2] Abo-Zaid GMA, Guo B, Deeks JJ, et al. Individual participant data meta-analyses should not ignore clustering. *Journal of Clinical Epidemiology.* 2013;66(8):865–873. doi:10.1016/j.jclinepi.2012.12.017.

[3] Abo-Zaid GMA, Sauerbrei W, Riley RD. Individual participant data meta-analysis of prognostic factor studies: state of the art? *BMC Medical Research Methodology.* 2012;12(1):56. doi:10.1186/1471-2288-12-56.

[4] Abu-Hanna A, Lucas PJ. Prognostic models in medicine. AI and statistical approaches. *Methods of Information in Medicine.* 2001;40(1):1–5.

[5] Adams ST, Leveson SH. Clinical prediction rules. *British Medical Journal.* 2012;344:d8312. doi:10.1136/bmj.d8312.

[6] Ahmed I, Debray TPA, Moons KGM, Riley RD. Developing and validating risk prediction models in an individual participant data meta-analysis. *Submitted.* 2013;.

[7] Ahmed I, Sutton AJ, Riley RD. Assessment of publication bias, selection bias, and unavailable data in meta-analyses using individual participant data: a database survey. *British Medical Journal.* 2012;344:d7762. doi:10.1136/bmj.d7762.

[8] Al Khalaf MM, Thalib L, Doi SAR. Combining heterogenous studies using the random-effects model is a mistake and leads to inconclusive meta-analyses. *Journal of Clinical Epidemiology.* 2011;64(2):119–123. doi:10.1016/j.jclinepi.2010.01.009.

[9] Albert A, Anderson J. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika.* 1984;71(1):1–10.

[10] Altman DG. Systematic reviews of evaluations of prognostic variables. *British Medical Journal.* 2001;323(7306). doi:10.1136/bmj.323.7306.224.

[11] 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.

[12] Altman DG, Bland JM. How to obtain the confidence interval from a P value. *British Medical Journal.* 2011;343:d2090. doi:10.1136/bmj.d2090.

[13] Altman DG, Deeks JJ. Meta-analysis, Simpson's paradox, and the number needed to treat. *BMC Medical Research Methodology.* 2002;2:3. doi:10.1186/1471-2288-2-3.

[14] Altman DG, Riley RD. Primer: an evidence-based approach to prognostic markers. *Nature Clinical Practice Oncology.* 2005;2(9):466–472. doi:10.1038/ncponc0287.

[15] Altman DG, Royston P. What do we mean by validating a prognostic model? *Statistics in Medicine*. 2000;19(4):453–473.

[16] Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. *British Medical Journal*. 2009;338:b605. doi:10.1136/bmj.b605.

[17] Anderson DR, Kovacs MJ, Kovacs G, et al. Combined use of clinical assessment and D-dimer to improve the management of patients presenting to the emergency department with suspected deep vein thrombosis (the EDITED Study). *Journal of Thrombosis and Haemostasis*. 2003;1(4):645–651.

[18] Anderson DR, Wells PS, Stiell I, et al. Management of patients with suspected deep vein thrombosis in the emergency department: combining use of a clinical diagnosis model with D-dimer testing. *The Journal of Emergency Medicine*. 2000;19(3). doi:10.1016/S0736-4679(00)00225-0.

[19] Assmann G, Cullen P, Schulte H. Simple scoring scheme for calculating the risk of acute coronary events based on the 10-year follow-up of the prospective cardiovascular Münster (PROCAM) study. *Circulation*. 2002;105(3):310–315. doi:10.1161/hc0302.102575.

[20] Austin PC. Estimating multilevel logistic regression models when the number of clusters is low: a comparison of different statistical software procedures. *International Journal of Biostatistics*. 2010;6(1):Article 16. doi:10.2202/1557-4679.1195.

[21] Bailey I, Bell A, Gray J, et al. A trial of the effect of nimodipine on outcome after head injury. *Acta Neurochir (Wien)*. 1991;110(3-4):97–105.

[22] 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.

[23] Baltagi BH, Bresson G, Griffin JM, Pirotte A. Homogeneous, heterogeneous or shrinkage estimators? Some empirical evidence from French regional gasoline consumption. *Empirical Economics*. 2003;28:795–811. doi:10.1007/s00181-003-0161-9.

[24] Bates SM, Kearon C, Crowther M, et al. A diagnostic strategy involving a quantitative latex D-dimer assay reliably excludes deep venous thrombosis. *Annals of Internal Medicine*. 2003;138(10):787–794.

[25] Becker B, Wu M. The Synthesis of Regression Slopes in Meta-Analysis. *Statistical Science*. 2007;22:414–429. doi:10.1214/07-STS243.

[26] Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*. 2005;24(11):1713–1723. doi:10.1002/sim.2059.

[27] Bennett DA. Review of analytical methods for prospective cohort studies using time to event data: single studies and implications for meta-analysis. *Statistical methods in medical research*. 2003;12(4):297–319. doi:10.1191/0962280203sm319ra.

[28] Berlin JA, Santanna J, Schmid CH, Szczech LA, Feldman HI, Anti-lymphocyte antibody induction therapy study group. Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head. *Statistics in Medicine*. 2002;21(3):371–387. doi:10.1002/sim.1023.

[29] Bland JM. Cluster randomised trials in the medical literature: two bibliometric surveys. *BMC Medical Research Methodology*. 2004;4:21. doi:10.1186/1471-2288-4-21.

[30] Bleeker SE, Moll HA, Steyerberg EW, et al. 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.

[31] Blettner M, Sauerbrei W, Schlehofer B, Scheuchenpflug T, Friedenreich C. Traditional reviews, meta-analyses and pooled analyses in epidemiology. *International Journal of Epidemiology*. 1999;28(1):1–9. doi:10.1093/ije/28.1.1.

[32] Böhning D, Sarol Jr J. Estimating risk difference in multicenter studies under baseline-risk heterogeneity. *Biometrics*. 2000;56(1):304–308. doi:10.1111/j.0006-341X.2000.00304.x.

[33] Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*. 2010;1(2):97–111. doi:10.1002/jrsm.12.

[34] Bouwmeester W, Zuithoff NPA, Mallett S, et al. Reporting and methods in clinical prediction research: a systematic review. *PLoS Medicine*. 2012;9(5):1–12.

[35] Bradburn MJ, Deeks JJ, Berlin JA, Russell Localio A. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. *Statistics in Medicine*. 2007;26(1):53–77. doi:10.1002/sim.2528.

[36] Braitman LE, Davidoff F. Predicting clinical states in individual patients. *Annals of Internal Medicine*. 1996;125(5):406–412.

[37] Breiman L. Stacked Regressions. *Machine Learning*. 1996;24(1):49–64. doi:10.1007/BF00117832.

[38] Breslow NE, Day NE. Statistical methods in cancer research. Volume II – The design and analysis of cohort studies. *IARC scientific publications*. 1987;(82):1–406.

[39] Brier GW. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*. 1950;78(1):1–3.

[40] Brockwell SE, Gordon IR. A comparison of statistical methods for meta-analysis. *Statistics in Medicine*. 2001;20(6):825–840. doi:10.1002/sim.650.

[41] Broström G, Holmberg H. Generalized linear models with clustered data: Fixed and random effects models. *Computational Statistics & Data Analysis*. 2011;55(12):3123–3134. doi:10.1016/j.csda.2011.06.011.

[42] Brotman DJ, Walker E, Lauer MS, O'Brien RG. In search of fewer independent risk factors. *Archives of Internal Medicine*. 2005;165(2):138–145. doi:10.1001/archinte.165.2.138.

[43] Browne WJ, Draper D. A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*. 2006;1(3):473–514. doi:10.1214/06-BA117.

[44] Büller HR, Ten Cate-Hoek AJ, Hoes AW, et al. Safely ruling out deep venous thrombosis in primary care. *Annals of Internal Medicine*. 2009;150(4):229–235.

[45] Burgess S, White IR, Resche-Rigon M, Wood AM. Combining multiple imputation and meta-analysis with individual participant data. *Statistics in Medicine*. 2013;doi:10.1002/sim.5844.

[46] Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. *Statistics in Medicine.* 2006;25(24):4279–4292. doi:10.1002/sim.2673.

[47] Byrd RH, Lu P, Nocedal J, Zhu C. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing.* 1995;16(5):1190–1208. doi:10.1137/0916069.

[48] Callas PW, Pastides H, Hosmer DW. Empirical comparisons of proportional hazards, poisson, and logistic regression modeling of occupational cohort data. *American Journal Of Indusrial Medicine.* 1998;33(1):33–47. doi:10.1002/(SICI)1097-0274(199801)33:1⟨33::AID-AJIM5⟩3.0.CO;2-X.

[49] Carlsson CA, von Essen C, Löfgren J. Factors affecting the clinical corse of patients with severe head injuries. 1. Influence of biological factors. 2. Significance of posttraumatic coma. *Journal of Neurosurgery.* 1968;29(3):242–251. doi:10.3171/jns.1968.29.3.0242.

[50] Chen H, Manning AK, Dupuis J. A Method of Moments Estimator for Random Effect Multivariate Meta-Analysis. *Biometrics.* 2012;Accepted for publication. doi:10.1111/j.1541-0420.2012.01761.x.

[51] Chu H, Cole SR. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. *Journal of Clinical Epidemiology.* 2006;59(12):1331–1332. doi:10.1016/j.jclinepi.2006.06.011.

[52] Clarke M. Doing new research? Don't forget the old. *PLoS Medicine.* 2004;1(2):e35. doi:10.1371/journal.pmed.0010035.

[53] Clemons TE, Bradley Jr EL. A nonparametric measure of the overlapping coefficient. *Computational Statistics & Data Analysis.* 2000;34(1):51–61. doi:10.1016/S0167-9473(99)00074-2.

[54] Collins GS, Altman DG. Identifying patients with undetected colorectal cancer: an independent validation of QCancer (Colorectal). *British Journal of Cancer.* 2012;107(2):260–265. doi:10.1038/bjc.2012.266.

[55] Collins GS, Omar O, Shanyinde M, Yu LM. A systematic review finds prediction models for chronic kidney disease were poorly reported and often developed using inappropriate methods. *Journal of Clinical Epidemiology.* 2013;66(3):268–277. doi:10.1016/j.jclinepi.2012.06.020.

[56] Conroy RM, Pyörälä K, Fitzgerald AP, et al. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. *European Heart Journal.* 2003;24(11):987–1003. doi:10.1016/S0195-668X(03)00114-3.

[57] 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.

[58] Cook NR. Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. *Clinical Chemistry.* 2008;54(1):17–23. doi:10.1373/clinchem.2007.096529.

[59] Copas JB. Using regression models for prediction: shrinkage and regression to the mean. *Statistical Methods in Medical Research.* 1997;6(2):167–183. doi:10.1177/096228029700600206.

[60] Cox DR. Two further applications of a model for binary regression. *Biometrika.* 1958;45(3):562–565.

[61] Crowther MJ, Riley RD, Staessen JA, Wang J, Gueyffier F, Lambert PC. Individual patient data meta-analysis of survival data using Poisson regression models. *BMC Medical Research Methodology.* 2012;12:34. doi:10.1186/1471-2288-12-34.

[62] Davison A, Hinkley D. Number 1 in Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge. 1997.

[63] de Leeuw C, Klugkist I. Augmenting data with published results in Bayesian linear regression. *Multivariate Behavioral Research.* 2012;47(3):369–391. doi:10.1080/00273171.2012.673957.

[64] de Leon AR, Carriere KC. A generalized Mahalanobis distance for mixed data. *Journal of Multivariate Analysis.* 2005;92:174–185. doi:10.1016/j.jmva.2003.08.006.

[65] Debray TPA, Koffijberg H, Lu D, Vergouwe Y, Steyerberg EW, Moons KG. Incorporating published univariable associations in diagnostic and prognostic modeling. *BMC Medical Research Methodology.* 2012;12(1):121. doi:10.1186/1471-2288-12-121.

[66] Debray TPA, Koffijberg H, Vergouwe Y, Moons KG, Steyerberg EW. Aggregating published prediction models with individual participant data: a comparison of different approaches. *Statistics in Medicine.* 2012;31(23):2697–2712. doi:10.1002/sim.5412.

[67] Debray TPA, Moons KGM, Abo-Zaid GMA, Koffijberg H, Riley RD. Individual participant data meta-analysis for a binary outcome: one-stage or two-stage? *PLoS ONE.* 2013; 8(4):e60650. doi:10.1371/journal.pone.0060650.

[68] Debray TPA, Moons KGM, Ahmed I, Koffijberg H, Riley RD. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. *Statistics in Medicine.* 2013;32(18):3158–3180. doi:10.1002/sim.5732.

[69] Deeks JJ. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. *Statistics in Medicine.* 2002;21(11):1575–1600. doi:10.1002/sim.1188.

[70] DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials.* 1986; 7(3):177–188. doi:10.1016/0197-2456(86)90046-2.

[71] Domingos P. Bayesian Averaging of Classifiers and the Overfitting Problem. In *In Proceedings of the 17th international conference on machine learning.* Morgan Kaufmann. 2000; pages 223–230.

[72] Dutton MT. *Individual Patient-Level Data Meta-Analysis: A Comparison of Methods For The Diverse Populations Collaboration Data Set.* Ph.D. thesis, Florida State University. 2010.

[73] Elf JL, Strandberg K, Nilsson C, Svensson PJ. Clinical probability assessment and D-dimer determination in patients with suspected deep vein thrombosis, a prospective multicenter management study. *Thrombosis Research.* 2009;123(4):612–616. doi:10.1016/j.thromres.2008.04.007.

[74] Ettema RGA, Peelen LM, Kalkman CJ, Nierich AP, Moons KGM, Schuurmans MJ. Predicting prolonged intensive care unit stays in older cardiac surgery patients: a validation study. *Intensive Care Medicine.* 2011;37(9):1480–1487.

[75] Ettema RGA, Peelen LM, Schuurmans MJ, Nierich AP, Kalkman CJ, Moons KGM. Prediction models for prolonged intensive care unit stay after cardiac surgery: systematic review and validation study. *Circulation.* 2010;122(7):682–689.

[76] Falagas ME. The increasing body of research data in clinical medicine has led to the need for evidence synthesis studies. Preface. *Infectious disease clinics of North America*. 2009; 23(2):xiii. doi:10.1016/j.idc.2009.02.002.

[77] Fawcett T. ROC graphs: Notes and practical considerations for researchers. Technical Report HPL-2003-4, HP Labs Tech Report. 2004.

[78] Fibrinogen Studies Collaboration, Jackson D, White I, et al. Systematically missing confounders in individual participant data meta-analysis of observational cohort studies. *Statistics in Medicine*. 2009;28(8):1218–1237. doi:10.1002/sim.3540.

[79] Firth D. Bias reduction of maximum likelihood estimates. *Biometrika*. 1993;80(1):27–38. doi:10.1093/biomet/80.1.27.

[80] Foulkes MA, Eisenberg HM, Jane JA, Marmarou A, Marshall LF, the Traumatic Coma Data Bank Research Group. The Traumatic Coma Data Bank: design, methods, and baseline characteristics. *Journal of Neurosurgery*. 1991;75(1):S8–S13.

[81] 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.

[82] Gail MH, Brinton LA, Byar DP, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *Journal of the National Cancer Institute*. 1989;81(24):1879–1886. doi:10.1093/jnci/81.24.1879.

[83] Gail MH, Pfeiffer RM. On criteria for evaluating models of absolute risk. *Biostatistics*. 2005; 6(2):227–239. doi:10.1093/biostatistics/kxi005.

[84] Geersing GJ. *Strategies in suspected venous thrombo-embolism in primary care*. Ph.D. thesis, Utrecht University, Utrecht, The Netherlands. 2011.

[85] Geersing GJ, Bouwmeester W, Zuithoff P, Spijker R, Leeflang M, Moons KGM. Search filters for finding prognostic and diagnostic prediction studies in Medline to enhance systematic reviews. *PLoS One*. 2012;7(2):e32844.

[86] Geersing GJ, Janssen KJ, Oudega R, et al. 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.

[87] Gelman A, Jakulin A, Pittau MG, Su YS. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*. 2008;2(4):1360–1383. doi:10.1214/08-AOAS191.

[88] Gerds TA, Cai T, Schumacher M. The Performance of Risk Prediction Models. *Biometrical Journal*. 2008;50(4):457–479. doi:10.1002/bimj.200810443.

[89] Glidden DV, Vittinghoff E. Modelling clustered survival data from multicentre clinical trials. *Statistics in Medicine*. 2004;23(3):369–388. doi:10.1002/sim.1599.

[90] Graham JW. *Multiple Imputation and Analysis with Multilevel (Cluster) Data*, chapter 6, pages 133–150. Statistics for Social and Behavioral Sciences. Springer New York, New York. 2012;doi:10.1007/978-1-4614-4018-5_6.

[91] Greenland S. Quantitative methods in the review of epidemiologic literature. *Epidemiologic Reviews.* 1987;9(1):1–30.

[92] Greenland S. Invited commentary: a critical look at some popular meta-analytic methods. *American Journal of Epidemiology.* 1994;140(3):290–296.

[93] Greenland S. Principles of multilevel modelling. *International Journal of Epidemiology.* 2000; 29(1):158–167. doi:10.1093/ije/29.1.158.

[94] Greenland S, Mickey RM. Closed Form and Dually Consistent Methods for Inference on Strict Collapsibility in 2 x 2 x K and 2 x J x K Tables. *Journal of the Royal Statistical Society Series C (Applied Statistics).* 1988;37(3):335–343.

[95] Guang G, Hongxin Z. Multilevel modelling for binary data. *Annual Review of Sociology.* 2000;26:441–462. doi:10.1146/annurev.soc.26.1.441.

[96] Hailpern SM, Visintainer PF. Odds ratios and logistic regression: further examples of their use and interpretation. *The Stata Journal.* 2003;3(3):213–225.

[97] Hall PA, Going JJ. Predicting the future: a critical appraisal of cancer prognosis studies. *Histopathology.* 1999;35:489–494. doi:10.1046/j.1365-2559.1999.00862.x.

[98] Hamza TH, van Houwelingen HC, Stijnen T. The binomial distribution of meta-analysis was preferred to model within-study variability. *Journal of Clinical Epidemiology.* 2008; 61(1):41–51. doi:10.1016/j.jclinepi.2007.03.016.

[99] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology.* 1982;143(1):29–36.

[100] Hardy RJ, Thompson SG. A likelihood approach to meta-analysis with random effects. *Statistics in Medicine.* 1996;15:619–629.

[101] Harrell Jr FE. *Regression Modeling Strategies with applications to Linear Models, Logistic Regression and Survival Analysis.* Springer Series in Statistics. Springer, New York. 2001.

[102] Harrell Jr FE, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. *Statistics in Medicine.* 1984;3(2):143–152. doi:10.1002/sim.4780030207.

[103] Hastie T, Tibshirani R, Friedman J. Basis Expansions and Regularization. In *The Elements of Statistical Learning*, Springer Series in Statistics, chapter Basis Expansions and Regularization, pages 139–189. Springer New York, New York, NY. 2009;doi:10.1007/978-0-387-84858-7_5.

[104] Hayden JA, Côté P, Bombardier C. Evaluation of the quality of prognosis studies in systematic reviews. *Annals of Internal Medicine.* 2006;144(6):427–437.

[105] Haynes RB, McKibbon KA, Wilczynski NL, Walter SD, Werre SR, . Optimal search strategies for retrieving scientifically strong studies of treatment from Medline: analytical survey. *British Medical Journal.* 2005;330(7501):1179. doi:10.1136/bmj.38446.498542.8F.

[106] Hedges LV, Vevea JL. Fixed- and Random-Effects Models in Meta-Analysis. *Psychological Methods.* 1998;3(4):486–504.

[107] Heinze G, Schemper M. A solution to the problem of separation in logistic regression. *Statistics in Medicine.* 2002;21(16):2409–2419. doi:10.1002/sim.1047.

[108] Hemingway H, Croft P, Perel P, et al. Prognosis Research Strategy (PROGRESS) 1: a framework for researching clinical outcomes. *British Medical Journal.* 2012;in press.

[109] Hemingway H, Riley RD, Altman DG. Ten steps towards improving prognosis research. *British Medical Journal.* 2009;339:b4184. doi:10.1136/bmj.b4184.

[110] Hernández AV, Steyerberg EW, Habbema JDF. Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. *Journal of Clinical Epidemiology.* 2004;57(5):454–460. doi:10.1016/j.jclinepi.2003.09.014.

[111] Higgins JPT, Green S. *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0.* The Cochrane Collaboration. 2011.

[112] Higgins JPT, 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.

[113] 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.

[114] Higgins JPT, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society Series A (Statistics in Society).* 2009; 172(1):137–159. doi:10.1111/j.1467-985X.2008.00552.x.

[115] Higgins JPT, Whitehead A, Turner RM, Omar RZ, Thompson SG. Meta-analysis of continuous outcome data from individual patients. *Statistics in Medicine.* 2001;20(15):2219–2241. doi:10.1002/sim.918.

[116] Hilden J. The area under the ROC curve and its competitors. *Medical Decision Making.* 1991;11(2):95–101. doi:10.1177/0272989X9101100204.

[117] Hlatky MA, Boothroyd DB, Bravata DM, et al. Coronary artery bypass surgery compared with percutaneous coronary interventions for multivessel disease: a collaborative analysis of individual patient data from ten randomised trials. *Lancet.* 2009;373(9670):1190–1197. doi:10.1016/S0140-6736(09)60552-3.

[118] Hlatky MA, Greenland P, Arnett DK, et al. Criteria for evaluation of novel markers of cardiovascular risk: a scientific statement from the American Heart Association. *Circulation.* 2009;119(17):2408–2416. doi:10.1161/CIRCULATIONAHA.109.192278.

[119] Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian Model Averaging: A Tutorial. *Statistical Science.* 1999;14(4):382–417.

[120] Hukkelhoven CWPM, Steyerberg EW, Rampen AJJ, et al. Patient age and outcome following severe traumatic brain injury: an analysis of 5600 patients. *Journal of Neurosurgery.* 2003; 99(4):666–673. doi:10.3171/jns.2003.99.4.0666.

[121] Hunter JE, Schmidt FL. Fixed Effects vs. Random Effects Meta-Analysis Models: Implications for Cumulative Research Knowledge. *Implications for cumulative research knowledge.* 2000;8(4):275–292. doi:10.1111/1468-2389.00156.

[122] 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.

[123] Ingui BJ, Rogers MA. Searching for clinical prediction rules in MEDLINE. *Journal of the American Medical Informatics Association.* 2001;8(4):391–397. doi:10.1136/jamia.2001.0080391.

[124] Ioannidis JP, Cappelleri JC, Schmid CH, Lau J. Impact of epidemic and individual heterogeneity on the population distribution of disease progression rates. An example from patient populations in trials of human immunodeficiency virus infection. *American Journal of Epidemiology.* 1996;144(11):1074–1085.

[125] Ioannidis JP, Lau J. Heterogeneity of the baseline risk within patient populations of clinical trials: a proposed evaluation algorithm. *American Journal of Epidemiology.* 1998;148(11):1117–1126.

[126] Ioannidis JPA, Rosenberg PS, Goedert JJ, O'Brien TR, . Commentary: meta-analysis of individual participants' data in genetic epidemiology. *American Journal of Epidemiology.* 2002;156(3):204–210. doi:10.1093/aje/kwf031.

[127] 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.

[128] Jackson D, Bowden J, Baker R. How does the DerSimonian and Laird procedure for random effects meta-analysis compare with its more efficient but harder to compute counterparts? *Journal of Statistical Planning and Inference.* 2010;140(4):961–970. doi:10.1016/j.jspi.2009.09.017.

[129] 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.

[130] 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.

[131] Janssen KJM, Donders ART, Harrell FEJ, et al. Missing covariate data in medical research: to impute is better than to ignore. *Journal of Clinical Epidemiology.* 2010;63(7):721–727.

[132] Janssen KJM, Moons KGM, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. *Journal of Clinical Epidemiology.* 2008;61(1):76–86. doi:10.1016/j.jclinepi.2007.04.018.

[133] 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.

[134] 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.

[135] Jones AP, Riley RD, Williamson PR, Whitehead A. Meta-analysis of individual patient data versus aggregate data from longitudinal clinical trials. *Clinical Trials.* 2009;6(1):16–27. doi:10.1177/1740774508100984.

[136] Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Annals of Internal Medicine*. 1999;130(6):515–524.

[137] Kearon C, Ginsberg JS, Douketis J, et al. Management of suspected deep venous thrombosis in outpatients by using clinical assessment and D-dimer testing. *Annals of Internal Medicine*. 2001;135(2):108–111.

[138] Kearon C, Ginsberg JS, Douketis J, et al. A randomized trial of diagnostic strategies after normal proximal vein ultrasonography for suspected deep venous thrombosis: D-dimer testing compared with repeated ultrasonography. *Annals of Internal Medicine*. 2005;142(7):490–496.

[139] Koopman L, van der Heijden GJMG, Grobbee DE, Rovers MM. Comparison of methods of handling missing data in individual patient data meta-analyses: an empirical example on antibiotics in children with acute otitis media. *American Journal of Epidemiology*. 2008; 167(5):540–545. doi:10.1093/aje/kwm341.

[140] Kraaijenhagen RA, Piovella F, Bernardi E, et al. Simplification of the diagnostic management of suspected deep vein thrombosis. *Archives of Internal Medicine*. 2002;162(8):907–911.

[141] Krishnapuram B, Carin L, Figueiredo MAT, Hartemink AJ. Sparse Multinomial Logistic Regression: Fast Algorithms and Generalization Bounds. *IEEE Transactions on Pattern Analysis and Machine Learning*. 2005;27(6):957–968. doi:10.1109/TPAMI.2005.127.

[142] Kyzas PA, Denaxa-Kyza D, Ioannidis JPA. Almost all articles on cancer prognostic markers report statistically significant results. *European Journal of Cancer*. 2007;43(17):2559–2579. doi:10.1016/j.ejca.2007.08.030.

[143] Lee KJ, Thompson SG. The use of random effects models to allow for clustering in individually randomized trials. *Clinical Trials*. 2005;2(2):163–173. doi:10.1191/1740774505cn082oa.

[144] Lee KJ, Thompson SG. Flexible parametric models for random-effects distributions. *Statistics in Medicine*. 2008;27(3):418–434. doi:10.1002/sim.2897.

[145] Lesaffre E, Albert A. Partial separation in Logistic Discrimination. *Journal of the Royal Statistical Society Series B (Methodological)*. 1989;51(1):109–116. doi:10.2307/2345845.

[146] Lesaffre E, Spiessens B. On the effect of the number of quadrature points in a logistic random-effects model: an example. *Journal of the Royal Statistical Society Series C (Applied Statistics)*. 2001;50(3):325–335. doi:10.1111/1467-9876.00237.

[147] 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.

[148] Ma R, Krewski D, Burnett RT. Random Effects Cox Models: A Poisson Modelling Approach. *Biometrika*. 2003;90(1):157–169.

[149] Maddala GS, Li H, Srivastava VK. A Comparative Study of Different Shrinkage Estimators for Panel Data Models. *Annals of Economics and Finance*. 2001;2(1):1–30.

[150] Maengseok N, Lee Y. REML estimation for binary data in GLMMs. *Journal of Multivariate Analysis*. 2007;98:896–915. doi:10.1016/j.jmva.2006.11.009.

[151] Mantel N, Haensel W. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*. 1959;22:719–748.

[152] Marmarou A, Lu J, Butcher I, et al. IMPACT database of traumatic brain injury: design and description. *Journal of Neurotrauma.* 2007;24(2):239–250. doi:10.1089/neu.2006.0036.

[153] Marmarou A, Nichols J, Burgess J, et al. Effects of the bradykinin antagonist Bradycor (deltibant, CP-1027) in severe Traumatic Brain Injury: results of a multi-center, randomized, placebo-controlled trial. *Journal of Neurotrauma.* 1999;16(6):431–344.

[154] Marshall LF, Maas AI, Marshall SB, et al. A multicenter trial on the efficacy of using tirilazad mesylate in cases of head injury. *Journal of Neurosurgery.* 1998;89(4):519–525.

[155] Mason C, Perreault W. Collinearity, Power, and Interpretation of Multiple Regression Analysis. *Journal of Marketing Research.* 1991;28(3):268–280. doi:10.2307/3172863.

[156] Mathew T, Nordström K. On the equivalence of meta-analyis using literature and using individual patient data. *Biometrics.* 1999;55(4):1221–1223. doi:10.1111/j.0006-341X.1999.01221.x.

[157] Mathew T, Nordström K. Comparison of one-step and two-step meta-analysis models using individual patient data. *Biometrical Journal.* 2010;52(2):271–287. doi:10.1002/bimj.200900143.

[158] Mavridis D, Salanti G. A practical introduction to multivariate meta-analysis. *Statistical Methods in Medical Research.* 2012;doi:10.1177/0962280211432219.

[159] 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;132:1–13. doi:10.1007/s10549-011-1818-2.

[160] Miller ME, Langefeld CD, Tierney WM, Hui SL, McDonald CJ. Validation of probabilistic predictions. *Medical Decision Making.* 1993;13(1):49–58.

[161] Minka T. Bayesian model averaging is not model combination. Technical report, MIT. 2000.

[162] Mizuno S, Yamaguchi T, Fukushima A, Matsuyama Y, Ohashi Y. Overlap coefficient for assessing the similarity of pharmacokinetic data between ethnically different populations. *Clinical Trials.* 2005;2(2):174–181. doi:10.1191/1740774505cn077oa.

[163] Monteith K, Carroll JL, SEppi K, Martinez T. Turning Bayesian model averaging into Bayesian model combination. In *Proceedings of the 2011 International Joint Conference on Neural Networks.* San Jose, California. 2011; pages 2657–2663.

[164] Montgomery JM, Nyhan B. Bayesian Model Averaging: Theoretical Developments and Practical Applications. *Political Analysis.* 2010;18(2):245–270. doi:10.1093/pan/mpq001.

[165] Montori VM, Swiontkowski MF, Cook DJ. Methodologic Issues in Systematic Reviews and Meta-Analyses. *Clinical Orthopaedics and Related Research.* 2003;413:43–54. doi:10.1097/01.blo.0000079322.41006.5b.

[166] Moons KGM. Criteria for scientific evaluation of novel markers: a perspective. *Clinical Chemistry.* 2010;56(4):537–541. doi:10.1373/clinchem.2009.134155.

[167] 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. doi:10.1136/bmj.b606.

[168] Moons KGM, Donders ART, Steyerberg EW, Harrell FE. Penalized maximum likelihood estimation to directly adjust diagnostic and prognostic prediction models for overoptimism: a clinical example. *Journal of Clinical Epidemiology.* 2004;57(12):1262–1270. doi:10.1016/j.jclinepi.2004.01.020.

[169] 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. doi:10.1136/heartjnl-2011-301247.

[170] Moons KGM, Kengne AP, Woodward M, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart.* 2012;98(9):683–690. doi:10.1136/heartjnl-2011-301246.

[171] Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? *British Medical Journal.* 2009;338. doi:10.1136/bmj.b375.

[172] Morris GF, Bullock R, Marshall SB, Marmarou A, Maas A, Marshall LF. Failure of the competitive N-methyl-D-aspartate antagonist Selfotel (CGS 19755) in the treatment of severe head injury: results of two Phase III clinical trials. *Journal of Neurosurgery.* 1999;91(5):737–743.

[173] Murray GD, Teasdale GM, Braakman R, et al. The European Brain Injury Consortium survey of head injuries. *Acta Neurochir (Wien).* 1999;141(3):223–236.

[174] Murray LS, Teasdale GM, Murray GD, Miller DJ, Pickard JD, Shaw MD. Head injuries in four British neurosurgical centres. *British Journal of Neurosurgery.* 1999;13(6):564–549.

[175] Mushkudiani NA, Hukkelhoven CWPM, Hernández AV, et al. 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.

[176] Nashef SA, Roques F, Michel P, Gauducheau E, Lemeshow S, Salamon R. European system for cardiac operative risk evaluation (EuroSCORE). *European Journal of Cardiothorac Surgery.* 1999;16(1):9–13. doi:10.1016/S1010-7940(99)00134-7.

[177] Nia VP. 8th Iranian Statistics Conference. In *Gauss-Hermite quadrature: numerical or statistical method?* 2006; .

[178] Nielsen FS. Proper and Improper Multiple Imputation. *International Statistical Review.* 2003;71(3):593–627. doi:10.1111/j.1751-5823.2003.tb00214.x.

[179] Normand SL. Meta-analysis: formulating, evaluating, combining, and reporting. *Statistics in Medicine.* 1999;18(3):321–359. doi:10.1002/sim.3602.

[180] O'Brien PC. Comparing Two Samples: Extensions of the t, Rank-Sum, and Log-Rank Tests. *Journal of the American Statistical Association.* 1988;83(401):52–61.

[181] Olkin I, Sampson A. Comparison of meta-analysis versus analysis of variance of individual patient data. *Biometrics.* 1998;54(1):317–322.

[182] 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.

[183] Park S, Lake ET. Multilevel modeling of a clustered continuous outcome: nurses' work hours and burnout. *Nursing Research.* 2005;54(6):406–413.

[184] Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology.* 1996;49(12):1373–1379. doi:10.1016/S0895-4356(96)00236-3.

[185] Pencina MJ, D'Agostino Sr RB, D'Agostino Jr RB, 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.

[186] 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.

[187] Peters TJ, Richards SH, Bankhead CR, Ades AE, Sterne JAC. Comparison of methods for analysing cluster randomized trials: an example involving a factorial design. *International Journal of Epidemiology.* 2003;32(5):840–846. doi:10.1093/ije/dyg228.

[188] Phillips RS, Sutton AJ, Riley RD, et al. Predicting infectious complications in neutropenic children and young people with cancer (IPD protocol). *Systematic Reviews.* 2012;1(1):8. doi:10.1186/2046-4053-1-8.

[189] Pinheiro JC, Bates DM. Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model. *Journal of Computational and Graphical Statistics.* 1995;4(1):12–35.

[190] R Development Core Team. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing Vienna Austria.* 2011;.

[191] Rabe-Hesketh S, Skrondal A. Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal.* 2002;2(1):1–21.

[192] Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. *Annals of Internal Medicine.* 2006;144(3):201–209.

[193] 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.

[194] Rice VH, Stead LF. Nursing interventions for smoking cessation. *Cochrane Database of Systematic Reviews.* 2001;(3):CD001188. doi:10.1002/14651858.CD001188.

[195] Riley RD. Multivariate meta-analysis: the effect of ignoring within-study correlation. *Journal of the Royal Statistical Society Series A (Statistics in Society).* 2009;172(4):789–811. doi:10.1111/j.1467-985X.2008.00593.x.

[196] Riley RD, Abrams KR, Lambert P, Sutton A, Altman D. Where next for evidence synthesis of prognostic marker studies? Improving the quality and reporting of primary studies to facilitate clinically relevant evidence-based results. In NMG Auget Jean-Louis;Balakrishnan, editor, *Advances in Statistical Methods for the Health Sciences*, Statistics for Industry and Technology, pages 39–58. Birkhäuser Boston. 2007;.

[197] Riley RD, Abrams KR, Lambert PC, Sutton AJ, Thompson JR. An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. *Statistics in Medicine*. 2007;26(1):78–97. doi:10.1002/sim.2524.

[198] 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.

[199] Riley RD, Abrams KR, Sutton AJ, et al. Reporting of prognostic markers: current problems and development of guidelines for evidence based practice in the future. *British Journal of Cancer*. 2003;88:1191–1198. doi:10.1038/sj.bjc.6600886.

[200] Riley RD, Dodd SR, Craig JV, Thompson JR, Williamson PR. Meta-analysis of diagnostic test studies using individual patient data and aggregate data. *Statistics in Medicine*. 2008; 27(29):6111–6136. doi:10.1002/sim.3441.

[201] Riley RD, Hayden JA, Steyerberg EW, et al. Prognosis Research Strategy (PROGRESS) 2: Prognostic Factor Research. *PLoS Medicine*. 2013;10(2):e1001380. doi:10.1371/journal.pmed.1001380.

[202] Riley RD, Higgins JPT, Deeks JJ. Interpretation of random effects meta-analyses. *British Medical Journal*. 2011;342:d549. doi:10.1136/bmj.d549.

[203] Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. *British Medical Journal*. 2010;340:c221. doi:10.1136/bmj.c221.

[204] Riley RD, Lambert PC, Staessen JA, et al. 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.

[205] Riley RD, Ridley G, Williams K, Altman DG, Hayden J, de Vet HCW. Prognosis research: toward evidence-based results and a Cochrane methods group. *Journal of Clinical Epidemiology*. 2007;60(8):863–5; author reply 865–6. doi:10.1016/j.jclinepi.2007.02.004.

[206] Riley RD, Sauerbrei W, Altman DG. Prognostic markers in cancer: the evolution of evidence from single studies to meta-analysis, and beyond. *British Journal of Cancer*. 2009; 100(8):1219–1229. doi:10.1038/sj.bjc.6604999.

[207] 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.

[208] Riley RD, Steyerberg EW. Meta-analysis of a binary outcome using individual participant data and aggregate data. *Research Synthesis Methods*. 2010;1(1):2–19. doi:10.1002/jrsm.4.

[209] Robinson LD, Jewell NP. Some Surprising Results about Covariate Adjustment in Logistic Regression Models. *International Statistical Review / Revue Internationale de Statistique*. 1991;59(2):227–240.

[210] Royston P. Estimating a smooth baseline hazard function for the Cox model. Technical report, University College London. 2011.

[211] Royston P, Altman DG. Visualizing and assessing discrimination in the logistic regression model. *Statistics in Medicine*. 2010;29(24):2508–2520. doi:10.1002/sim.3994.

[212] Royston P, Altman DG. External validation of a Cox prognostic model: principles and methods. *BMC Medical Research Methodology.* 2013;13(33). doi:10.1186/1471-2288-13-33.

[213] Royston P, Moons KGM, Altman DG, Vergouwe Y. Prognosis and prognostic research: Developing a prognostic model. *British Medical Journal.* 2009;338:b604. doi:10.1136/bmj.b604.

[214] Royston P, Parmar MKB. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine.* 2002;21(15):2175–2197. doi:10.1002/sim.1203.

[215] Royston P, Parmar MKB, Sylvester R. Construction and validation of a prognostic model across several studies, with an application in superficial bladder cancer. *Statistics in Medicine.* 2004;23(6):907–926. doi:10.1002/sim.1691.

[216] Rubin D. *Multiple Imputation for Nonresponse in Surveys.* John Wiley and Sons, New York. 1987.

[217] Rücker G, Schwarzer G, Carpenter JR, Schumacher M. Undue reliance on I(2) in assessing heterogeneity may mislead. *BMC Medical Research Methodology.* 2008;8:79. doi:10.1186/1471-2288-8-79.

[218] Sauerbrei W. Prognostic factors. Confusion caused by bad quality design, analysis and reporting of many studies. *Advances in oto-rhino-laryngology.* 2005;62:184–200. doi:10.1159/000082508.

[219] Sauerbrei W, Holländer N, Riley R, Altman D. Evidence-Based Assessment and Application of Prognostic Markers: The Long Way from Single Studies to Meta-Analysis. *Communications in Statistics - Theory and Methods.* 2006;35(7). doi:10.1080/03610920600629666.

[220] Schafer JL. *Analysis of Incomplete Multivariate Data.* Chapman & Hall, London. 1997.

[221] Schmid CH, Griffith JL. Multivariate Classification Rules: Calibration and Discrimination. *Encyclopedia of Biostatistics.* 2005;5(2):3491–3497. doi:10.1002/0470011815.b2a13049.

[222] Schuit E, Kwee A, Westerhuis MEMH, et al. A clinical prediction model to assess the risk of operative delivery. *BJOG : an international journal of obstetrics and gynaecology.* 2012;119(8):915–923. doi:10.1111/j.1471-0528.2012.03334.x.

[223] Schutgens REG, Ackermark P, Haas FJLM, et al. Combination of a normal D-dimer concentration and a non-high pretest clinical probability score is a safe strategy to exclude deep venous thrombosis. *Circulation.* 2003;107(4):593–597.

[224] Selvarajah S, Fong AYY, Selvaraj G, Haniff J, Uiterwaal CSPM, Bots ML. An Asian validation of the TIMI risk score for ST-segment elevation myocardial infarction. *PLoS One.* 2012;7(7):e40249. doi:10.1371/journal.pone.0040249.

[225] Shrier I, Boivin JF, Platt RW, et al. The interpretation of systematic reviews with meta-analyses: an objective or subjective process? *BMC Medical Informatics and Decision Making.* 2008;8:19. doi:10.1186/1472-6947-8-19.

[226] Simmonds MC, Higgins JPT, Stewart LA, Tierney JF, Clarke MJ, Thompson SG. Meta-analysis of individual patient data from randomized trials: a review of methods used in practice. *Clinical Trials.* 2005;2(3):209–217. doi:10.1191/1740774505cn087oa.

[227] Simon R, Altman DG. Statistical aspects of prognostic factor studies in oncology. *British Journal of Cancer*. 1994;69:979985. doi:10.1038/bjc.1994.192.

[228] Smith-Spangler CM. Transparency and Reproducible Research in Modeling. *Medical Decision Making*. 2012;32(5):663–666. doi:10.1177/0272989X12458977.

[229] Spiegelhalter DJ. Probabilistic prediction in patient management and clinical trials. *Statistics in Medicine*. 1986;5(5):421–433.

[230] Steenbergen MR, Jones BS. Modeling multilevel data structures. *American Journal of Political Science*. 2002;46(1):218–237.

[231] Sterne JAC. *Meta-analysis in Stata: an updated collection from the Stata Journal.* Stata Press. 2009.

[232] Stevens SM, Elliott CG, Chan KJ, Egger MJ, Ahmed KM. Withholding anticoagulation after a negative result on duplex ultrasonography for suspected symptomatic deep venous thrombosis. *Annals of Internal Medicine*. 2004;140(12):985–991.

[233] Stewart GB, Altman DD, Askie L, Duley L, Simmonds M, Stewart LA. Statistical analysis of individual participant data meta-analyses: a comparison of methods and recommendations for practice. *PLoS One*. 2012;7(10):e46042. doi:10.1371/journal.pone.0046042.

[234] Stewart LA, Clarke MJ. Practical methodology of meta-analyses (overviews) using updated individual patient data. *Statistics in Medicine*. 1995;14(19):2057–2079. doi:10.1002/sim.4780141902.

[235] Stewart LA, Parmar MK. Meta-analysis of the literature or of individual patient data: is there a difference? *Lancet*. 1993;341(8842):418–422. doi:10.1016/0140-6736(93)93004-K.

[236] Stewart LA, Tierney JF. To IPD or not to IPD? Advantages and disadvantages of systematic reviews using individual patient data. *Evaluation & the health professions*. 2002;25(1):76–97. doi:10.1177/0163278702025001006.

[237] Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating.* Springer, New York. 2009.

[238] 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.

[239] 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.

[240] Steyerberg EW, Bossuyt PM, Lee KL. Clinical trials in acute myocardial infarction: should we adjust for baseline characteristics? *American Heart Journal*. 2000;139(5):745–751. doi:10.1016/S0002-8703(00)90001-2.

[241] 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.

[242] Steyerberg EW, Eijkemans MJ, Harrell FEJ, Habbema JD. Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. *Medical Decision Making*. 2001;21(1):45–56. doi:10.1177/0272989X0102100106.

[243] 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.

[244] Steyerberg EW, Eijkemans MJC, Habbema JDF. Application of Shrinkage Techniques in Logistic Regression Analysis: A Case Study. *Statistica Neerlandica*. 2001;55(1):76–88. doi:10.1111/1467-9574.00157.

[245] Steyerberg EW, Eijkemans MJC, Harrell Jr FE, Habbema JDF. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. *Statistics in Medicine*. 2000;19(8):1059–1079.

[246] Steyerberg EW, Kievit J, de Mol Van Otterloo JC, van Bockel JH, Eijkemans MJ, Habbema JD. Perioperative mortality of elective abdominal aortic aneurysm surgery. A clinical prediction rule based on literature and individual patient data. *Archives of Internal Medicine*. 1995;155(18):1998–2004. doi:10.1001/archinte.1995.00430180108012.

[247] Steyerberg EW, Moons KGM, van der Windt DA, et al. Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. *PLoS Medicine*. 2013;10(2):e1001381. doi:10.1371/journal.pmed.1001381.

[248] Steyerberg EW, Mushkudiani N, Perel P, 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.

[249] Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010;21(1):128–138. doi:10.1097/EDE.0b013e3181c30fb2.

[250] Stijnen T, Hamza TH, Özdemir P. Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data. *Statistics in Medicine*. 2010;29(29):3046–3067. doi:10.1002/sim.4040.

[251] 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.

[252] Sutton AJ, Cooper NJ, Jones DR. Evidence synthesis as the key to more coherent and efficient research. *BMC Medical Research Methodology*. 2009;9:29. doi:10.1186/1471-2288-9-29.

[253] Sutton AJ, Higgins JPT. Recent developments in meta-analysis. *Statistics in Medicine*. 2008;27(5):625–650. doi:10.1002/sim.2934.

[254] Sutton AJ, Kendrick D, Coupland C. Meta-analysis of individual- and aggregate-level data. *Statistics in Medicine*. 2008;27(5):651–669. doi:10.1002/sim.2916.

[255] Sweeting MJ, Sutton AJ, Lambert PC. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. *Statistics in Medicine*. 2004;23(9):1351–1375. doi:10.1002/sim.1761.

[256] 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. doi:10.3171/jns.1994.80.5.0797.

[257] Thompson S, Kaptoge S, White I, et al. Statistical methods for the time-to-event analysis of individual participant data from multiple epidemiological studies. *International Journal of Epidemiology.* 2010;39(5):1345–1359. doi:10.1093/ije/dyq063.

[258] Toll D. *Excluding Deep Vein Thrombosis in primary care: Validation, updating, and implementation of a diagnostic rule.* Ph.D. thesis, Utrecht University. 2008.

[259] 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.

[260] Toll DB, Oudega R, Bulten RJ, Hoes AW, Moons KGM. Excluding deep vein thrombosis safely in primary care. *The Journal of Family Practice.* 2006;55(7):613–618.

[261] 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). doi:10.1093/fampra/cmm075.

[262] Trikalinos TA, Ioannidis JP. Predictive modeling and heterogeneity of baseline risk in meta-analysis of individual patient data. *Journal of Clinical Epidemiology.* 2001;54(3):245–252. doi:10.1016/j.cct.2007.08.002.

[263] Tudur Smith C, Williamson PR. A comparison of methods for fixed effects meta-analysis of individual patient data with time to event outcomes. *Clinical Trials.* 2007;4(6):621–630. doi:10.1177/1740774507085276.

[264] Tudur Smith C, Williamson PR, Marson AG. Investigating heterogeneity in an individual patient data meta-analysis of time to event outcomes. *Statistics in Medicine.* 2005; 24(9):1307–1319. doi:10.1002/sim.2050.

[265] Turner EL, Perel P, Clayton T, et al. Covariate adjustment increased power in randomized controlled trials: an example in traumatic brain injury. *Journal of Clinical Epidemiology.* 2012;65(5):474–481. doi:10.1016/j.jclinepi.2011.08.012.

[266] Turner RM, Omar RZ, Yang M, Goldstein H, Thompson SG. A multilevel model framework for meta-analysis of clinical trials with binary outcomes. *Statistics in Medicine.* 2000; 19(24):3417–3432. doi:10.1002/1097-0258(20001230)19:24⟨3417::AID-SIM614⟩3.0.CO;2-L.

[267] van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research.* 2007;16(3):219–242.

[268] van Buuren S, Brand JPL, Groothuis-Oudshoorn CGM, Rubin DB. Fully conditional specications in multivariate imputation. *Journal of Statistical Computation and Simulation.* 2006; 72(12):1049–1064. doi:10.1080/10629360600810434.

[269] van Dieren S, Beulens JWJ, Kengne AP, et al. Prediction models for the risk of cardiovascular disease in patients with type 2 diabetes: a systematic review. *Heart.* 2012;98(5):360–369. doi: 10.1136/heartjnl-2011-300734.

[270] van Houwelingen HC. Validation, calibration, revision and combination of prognostic survival models. *Statistics in Medicine.* 2000;19(24):3401–3415. doi:10.1002/1097-0258(20001230)19:24⟨3401::AID-SIM554⟩3.0.CO;2-2.

[271] 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.

[272] 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.

[273] Vergouwe Y, Moons KGM, Steyerberg EW. External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *American Journal of Epidemiology.* 2010;172(8):971–980. doi:10.1093/aje/kwq223.

[274] Vergouwe Y, Nieboer D, Debray TPA, et al. Updating of risk models: a closed testing procedure for model selection. *To be submitted.* 2013;.

[275] Vergouwe Y, Steyerberg EW, Eijkemans MJC, Habbema JDF. Validity of prognostic models: when is a model clinically useful? *Seminars in Urologic Oncology.* 2002;20(2):96–107.

[276] Vergouwe Y, Steyerberg EW, Eijkemans MJC, Habbema JDF. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *Journal of Clinical Epidemiology.* 2005;58(5):475–483. doi:10.1016/j.jclinepi.2004.06.017.

[277] Vickers AJ, Cronin AM. Traditional statistical methods for evaluating prediction models are uninformative as to clinical value: towards a decision analytic framework. *Seminars in Oncology.* 2010;37(1):31–38. doi:10.1053/j.seminoncol.2009.12.004.

[278] Walker E, Hernandez AV, Kattan MW. Meta-analysis: Its strengths and limitations. *Cleveland Clinic Journal of Medicine.* 2008;75(6):431–439. doi:10.3949/ccjm.75.6.431.

[279] Walter SD. Variation in baseline risk as an explanation of heterogeneity in meta-analysis. *Statistics in Medicine.* 1997;16(24):2883–2900. doi:10.1002/(SICI)1097-0258(19971230)16:24⟨2883::AID-SIM825⟩3.0.CO;2-B.

[280] Wasson JH, Sox HC, Neff RK, Goldman L. Clinical prediction rules. Applications and methodological standards. *The New England Journal of Medicine.* 1985;313(13):793–799.

[281] Wells PS, Anderson DR, Bormanis J, et al. Value of assessment of pretest probability of deep-vein thrombosis in clinical management. *Lancet.* 1997;350(9094):1795–1798.

[282] Wells PS, Anderson DR, Rodger M, et al. 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.

[283] Wells PS, Ginsberg JS, Anderson DR, Wang D, Kearon C, Wells G. A simple clinical model to categorize pretest probability in patients with suspected pulmonary embolism. *Blood.* 1999;90:424a.

[284] White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in Medicine.* 2010;29(28):2920–2931. doi:10.1002/sim.3944.

[285] White IR, Higgins JPT, Wood AM. Allowing for uncertainty due to missing data in meta-analysis–part 1: two-stage methods. *Statistics in Medicine.* 2008;27(5):711–727. doi:10.1002/sim.3008.

[286] White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine.* 2011;30(4):377–399. doi:10.1002/sim.4067.

[287] White IR, Welton NJ, Wood AM, Ades AE, Higgins JPT. Allowing for uncertainty due to missing data in meta-analysis–part 2: hierarchical models. *Statistics in Medicine.* 2008; 27(5):728–745. doi:10.1002/sim.3007.

[288] Whitehead A, Omar RZ, Higgins JP, Savaluny E, Turner RM, Thompson SG. Meta-analysis of ordinal outcomes using individual patient data. *Statistics in Medicine.* 2001;20(15):2243–2260. doi:10.1002/sim.919.

[289] Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. *Circulation.* 1998;97(18):1837–1847. doi:10.1161/01.CIR.97.18.1837.

[290] Wolpert D. Stacked generalization. *Neural Networks.* 1992;5:241–259.

[291] Wong SSL, Wilczynski NL, Haynes RB, Ramkissoonsingh R, . Developing optimal search strategies for detecting sound clinical prediction studies in MEDLINE. *AMIA Annual Symposium Proceedings Archive.* 2003;pages 728–732.

[292] Woodward M, Brindle P, Tunstall-Pedoe H, SIGN group on risk estimation. Adding social deprivation and family history to cardiovascular risk assessment: the ASSIGN score from the Scottish Heart Health Extended Cohort (SHHEC). *Heart.* 2007;93(2):172–176. doi:10.1136/hrt.2006.108167.

[293] Wyatt JC, Altman DG. Prognostic models: clinically useful or quickly forgotten? *British Medical Journal.* 1995;311:1539. doi:10.1136/bmj.311.7019.1539.

[294] Young B, Runge JW, Waxman KS, et al. Effects of pegorgotein on neurologic outcome of patients with severe head injury. A multicenter, randomized controlled trial. *Journal of the American Medical Association.* 1996;276(7):538–543. doi:10.1001/jama.1996.03540070034027.

[295] Yusuf S, Peto R, Lewis J, Collins R, Sleight P. Beta blockade during and after myocardial infarction: an overview of the randomized trials. *Progress in Cardiovascular Diseases.* 1985; 27(5):335–371.

Summary

"Scientific knowledge is a body of statements of varying degrees of certainty – some most unsure, some nearly sure, none absolutely certain."

– Richard Phillips Feynman

THE goal of this thesis was to introduce and evaluate novel statistical methods for performing meta-analysis in clinical prediction research. Hereto, two scenarios were considered in which individual participant data (IPD) or aggregate data (AD) were available from multiple studies. As evidence synthesis requires some degree of similarity between the study populations considered for inclusion into a novel prediction model, specific attention was put on the identification and estimation of between-study heterogeneity. A brief summary of the chapters presented in this thesis is given below.

Section II: *Meta-analysis when individual participant data (IPD) is available from multiple studies.*

Chapter 1 evaluates whether meta-analyses should account for clustering of participants within studies. Logistic regression was applied in real and simulated examples. Results indicated that estimated associations tend to weaken and that statistical significance could disappear when clustering was ignored. Simulations showed that models accounting for clustering performed consistently, whilst downwardly biased effect estimates and low coverage occurred when clustering was ignored. It was concluded that routinely accounting for clustering in IPD meta-analyses would improve the validity and generalizability of estimated predictor-outcome associations, whilst ignoring clustering creates a risk of misleading effect estimates and spurious conclusions.

Chapter 2 describes and compares several random-effects meta-analysis models for estimating factor-outcome associations from multiple risk-factor or predictor finding studies with a binary outcome. These models account for heterogeneity across studies by applying a one-stage or a two-stage method. One-stage methods use the IPD of each study and meta-analyze using the exact binomial distribution, whereas two-stage methods first reduce evidence to the aggregated level (e.g. odds ratios) and then meta-analyse assuming approximate normality. Although one-stage and two-stage methods are generally thought to produce similar results, there is a dearth of previous empirical comparisons. In this chapter, an empirical dataset was used to compare several one-stage and two-stage methods for obtaining unadjusted and adjusted risk-factor estimates. Results showed that one-stage and two-stage methods occasionally provided different parameter estimates or suffered from estimation issues. In particular, two-stage models yielded unstable estimates when zero cell counts occurred and one-stage models did not always converge. It was therefore advised that the choice and implementation type (e.g. univariable or multivariable) of a one-stage or two-stage method should be prespecified in the study protocol, as they may lead to different conclusions regarding statistical significance.

Chapter 3 describes the use of IPD from multiple studies to develop a multivariable risk prediction model. It considers several approaches for developing a multivariable logistic regression model from an IPD meta-analysis (IPD-MA) with potential between-study heterogeneity. As prediction models are usually validated or applied in new individuals or study populations, this chapter also proposes strategies for choosing a valid model intercept and evaluating model generalizability. In an empirical evaluation, results showed how stratified estimation allowed the estimation of study-specific model intercepts, and demonstrated how focused intercept choice facilitated the model's implementation in new individuals or study populations. This chapter also illustrated how internal-external cross-validation allowed the evaluation of a single, integrated prediction model. Finally, it was concluded that a model's performance is likely to be more consistent across studies if there is little or no heterogeneity in the effects of the included predictors.

Chapter 4 evaluates several approaches for developing a prediction model from an IPD meta-analysis when some predictor variables are systematically missing. This situation may arise, for instance, when predictor variables have not all been measured in each study. As a consequence, some within-study predictor effects are no longer identifiable, and predictor selection can no longer unequivocally be achieved. Results from an empirical example demonstrated that traditional imputation strategies sufficed for developing prediction models with adequate performance, and can be implemented fairly straightforward. As such strategies may, however, lead to a self-fulfilling prophecy regarding homogeneity of predictor effects, it was concluded that more sophisticated imputation strategies are needed to guarantee model generalizability.

Section III: *Meta-analysis when aggregate data (AD) is available from multiple studies.*

Chapter 5 proposes a novel adaptation method to incorporate previously published univariable associations in the construction of a novel prediction model. A case study and a simulation study were performed to compare the novel adaptation method with established approaches. Results demonstrated that performance of estimated multivariable associations considerably improved for small datasets where external evidence was included. Although the error of estimated associations decreased with increasing amount of IPD, it did not disappear completely, even in very large datasets. It was concluded that the novel adaptation method outperforms established approaches and is especially worthwhile when relatively limited IPD are available.

Chapter 6 evaluates three approaches for aggregating previously published prediction models with new data. Hereto it considers the situation where models are reported in the literature with predictors similar to those available in an IPD dataset. The three approaches were applied to more than 10 prediction models for predicting the outcome after Traumatic Brain Injury, and to 5 previously published models for predicting the presence of Deep Venous Thrombosis. Results

from the case studies demonstrated that aggregation yielded prediction models with an improved discrimination and calibration (common measures of model performance) in a vast majority of scenarios, and resulted in equivalent performance (compared to the standard approach) in a small minority of scenarios. The proposed aggregation approaches were particularly useful when few IPD were at hand. It was concluded that 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.

Chapter 7 evaluates two approaches, Model Averaging and Stacked Regressions, for aggregating previously published prediction models when a validation dataset is at hand. In contrast to Chapter 8, these approaches no longer require models from the literature to have similar predictors, and yield user-friendly stand-alone models that are adjusted for the new validation data. Results from two clinical datasets and a simulation study demonstrated that aggregation yielded prediction models with an improved discrimination and calibration in a vast majority of scenarios, and resulted in equivalent performance (compared to developing a novel model from scratch) when validation samples were relatively large. It was concluded that model aggregation is a promising extension of model updating when several models are available from the literature, and a validation dataset is at hand. Furthermore, the aggregation methods do not require existing models to have similar predictors and can be applied when relatively few data are at hand.

Chapter 8 proposes a sequence of steps for a proper interpretation of validation results from a clinical prediction model. These steps involve quantifying the differences in case mix between the development and validation sample, and assessing the model's corresponding performance. This information was then used to interpret estimates of model discrimination and calibration in terms of case-mix differences. To enhance the interpretation of independent validation studies, it was proposed to distinguish between a model's reproducibility and transportability. Finally, it was described how inadequate model performance can be improved by applying updating strategies.

Samenvatting

"Denkt aleer gij doende zijt / en doende, denk dan nog."

– Guido Gezelle

HET is gebruikelijk dat patiënten pas behandeld worden na het stellen van een diagnose (wat is de onderliggende ziekte?) of prognose (wat zal de uitkomst van deze ziekte zijn?). Deze inschatting berust meestal op de (klinische) interpretatie van verschillende patiëntgegevens en/of testuitslagen, maar ook steeds vaker op de voorspellingen van een statistische (predictie)model. Predictiemodellen worden geïmplementeerd zodra voldoende bewezen is dat ze accurate voorspellingen leveren in nieuwe patiëntenpopulaties. Helaas worden predictiemodellen meestal ontwikkeld uit relatief kleine datasets, die niet altijd even betrouwbaar of representatief zijn. Bovendien houden predictiemodellen vaak geen rekening met verschillen tussen populaties, en wordt hun voorspellend vermogen zelden getest in nieuwe patiënten. De prestaties van predictiemodellen zijn daarom vaak onduidelijk of teleurstellend, waardoor onderzoekers meestal een volledig nieuw model ontwikkelen voor 'eigen' gebruik. Deze gang van zaken heeft de afgelopen decennia geleid tot een overvloed aan gelijksoortige predictiemodellen. Zo zijn er meer dan 60 modellen ontwikkeld die de kans op overlijden van borstkankerpatiënten inschatten.

Dit proefschrift behandelt meta-analytische technieken ter ondersteuning van klinisch predictieonderzoek. Deze technieken laten toe om de wetenschappelijke bevindingen uit eerdere studies te integreren bij de ontwikkeling van een nieuw predictiemodel. In dit proefschrift wordt onderscheid gemaakt tussen de beschikking over ruwe data, b.v. individuele patiëntgegevens (IPD), en geaggregeerde data (AD). Omdat IPD en AD vaak verzameld worden in populaties met een variërende mate van vergelijkbaarheid, besteedt dit proefschrift een bijzondere aandacht aan de identificatie en schatting van heterogeneiteit tussen studies.

Section II: *Meta-analyse wanneer ruwe data (IPD) beschikbaar is uit verschillende studies.*

Hoofdstuk 1 onderzoekt of meta-analyses rekening moeten houden met de samenhang (clustering) van deelnemers binnen studies. Bij het toepassen van logistische regressie in echte en gesimuleerde voorbeelden leidde het negeren van clustering tot een onderschatting van associaties en het verdwijnen van statistische significantie. Simulaties toonden verder aan dat modellen die rekening houden met clustering consistente prestaties leveren, en dat modellen die clustering negeren leiden tot een neerwaartse vertekening van associaties en een verminderde dekking van betrouwbaarheidsintervallen. Er werd geconcludeerd dat meta-analyses die routinematig rekening houden met clustering om de validiteit en generaliseerbaarheid van geschatte associaties te verbeteren.

Hoofdstuk 2 beschrijft en vergelijkt verschillende random-effects modellen voor het meta-analyseren van risico-factor associaties met een binaire uitkomst. Random-effects modellen houden rekening met heterogeniteit tussen studies door het toepassen van een een- of tweestaps methode. Eenstaps methoden gebruiken rechtstreeks de IPD uit elke studie en meta-analyseren deze gegevens met een exacte binomiale verdeling. Tweestaps methoden, daarentegen, reduceren de IPD eerst tot geaggregeerde data (zoals odds ratio's) en combineren de resulterende schattingen daarna uitgaande van normaliteit. Hoewel het algemeen aangenomen wordt dat een- en tweestaps methoden vergelijkbare resultaten opleveren, ontbreken empirische vergelijkingen. In dit hoofdstuk werd een empirische dataset gebruikt om verschillende een- en tweestaps methoden te vergelijken die uni- en multivariabele associaties schatten. Resultaten toonden aan dat een- en tweestaps methoden soms verschillende parameterschattingen opleveren of lijden aan schattingsproblemen. Met name tweestaps modellen leverden instabiele schattingen op bij het optreden van lege cellen, terwijl eenstapsmodellen niet altijd convergeerden. Er werd daarom geadviseerd dat de keuze en type (bv. univariabele of multivariabele) van een een- of tweestaps methode vooraf gespecificeerd moet worden in het studieprotocol.

Hoofdstuk 3 beschrijft hoe de IPD uit meerdere studies gebruikt kan worden tijdens de ontwikkeling van een multivariabel predictiemodel. Hiertoe werden verschillende technieken onderzocht om een logistisch regressiemodel te ontwikkelen die rekening houdt met heterogeniteit tussen studies. Daarnaast werden ook strategieën beschouwd die een geldig intercept opleveren wanneer het predictiemodel toepast of gevalideerd wordt in een nieuwe populatie. In een empirische studie werd aangetoond dat het waardevol is om een afzonderlijke intercept te schatten voor elke studie uit de meta-analyse. Dit laat immers toe om vervolgens een geschatte intercept te selecteren die gebruikt kan worden in een nieuwe populatie. Tot slot werd beschreven hoe de generaliseerbaarheid van het meta-analytisch predictiemodel geëvalueerd kan worden via interne-externe cross-validatie. Er werd geconcludeerd dat predictiemodellen betere prestaties leveren in nieuwe populaties wanneer hun associaties weinig of geen heterogeniteit tussen studies vertonen.

Hoofdstuk 4 onderzoekt hoe een predictiemodel ontwikkeld kan worden uit een IPD meta-analyse wanneer een aantal voorspellende variabelen systematisch ontbreken. Deze situatie kan optreden wanneer niet alle studies dezelfde variabelen hebben verzameld (zoals de resultaten van een biomarker test). Als gevolg hiervan zijn sommige associaties niet meer identificeerbaar in elke studie, en is het niet langer duidelijk hoe variabelen voor het predictiemodel geselecteerd kunnen worden. Een empirisch voorbeeld toonde aan dat imputatie van missende waarden een eenvoudige oplossing biedt, en rekening kan houden met heterogeniteit tussen studies.

Section III: *Meta-analyse wanneer geaggregeerde data (AD) beschikbaar is uit verschillende studies.*

Hoofdstuk 5 beschrijft een nieuwe adaptatiemethode om gepubliceerde univariabele (i.e. ongecorrigeerde) associaties te incorporeren tijdens de ontwikkeling van een multivariabel predictiemodel.

Deze adaptatiemethode werd vergeleken met bestaande technieken in een klinisch voorbeeld en een simulatiestudie. Resultaten toonden aan dat het incorporeren van gepubliceerde univariabele associaties leidde tot een aanzienlijk betere kwaliteit van geschatte multivariabele associaties in kleine datasets. Hoewel het negeren van van gepubliceerde associaties weinig impact had in grote datasets, verdween de error in geschatte associaties niet volledig. Er werd geconcludeerd dat de adaptatiemethode beter presteert dan bestaande technieken en vooral zinvol is wanneer relatief weinig IPD beschikbaar is.

Hoofdstuk 6 evalueert drie methoden die bestaande predictiemodellen kunnen integreren bij de ontwikkeling van een nieuw predictiemodel. Hiertoe werd de situatie beschouwd waarin een model ontwikkelingsdataset beschikbaar was, en dat de bestaande modellen (deels) dezelfde predictoren bevatten. De drie methoden werden toegepast op ruim tien predictiemodellen die de uitkomst na traumatisch hersenletsel voorspellen, en vijf gepubliceerde modellen die de aanwezigheid van diepe veneuze trombose voorspellen. Resultaten toonden aan dat aggregatie meestal leidde tot predictiemodellen met een betere discriminatie en kalibratie, en soms resulteerde in vergelijkbare prestaties (in vergelijking met de standaard methode). De beschreven aggregatie methoden waren met name waardevol in relatief kleine datasets. Er werd geconcludeerd dat het evalueren van de mate van heterogeniteit tussen de beschikbare data en de gepubliceerde predictiemodellen een cruciale rol speelt tijdens het bepalen van de optimale aggregatiemethode.

Hoofdstuk 7 evalueert twee methoden, *Model Averaging* en *Stacked Regressions*, die gepubliceerde predictiemodellen aggregeren wanneer een validatie dataset beschikbaar is. In tegenstelling tot Hoofdstuk 8, vereisen deze methoden niet dat de gepubliceerde modellen dezelfde predictoren bevatten. Bovendien leveren ze gebruiksvriendelijke stand-alone modellen die aangepast zijn voor de populatie van de validatieset. Resultaten uit twee klinische datasets en een simulatiestudie gaven aan dat aggregatie meestal leidde tot betere discriminatie en kalibratie, en resulteerde in vergelijkbare prestaties bij grotere validatie datasets. Er werd geconcludeerd dat model aggregatie een veelbelovende uitbreiding van model updating is wanneer meerdere gepubliceerde predictiemodellen en een validatie dataset beschikbaar zijn.

Hoofdstuk 8 stelt een stappenplan voor om de resultaten van een model validatiestudie beter te interpreteren. Deze stappen omvatten het kwantificeren van de case mix verschillen tussen de ontwikkelings- en validatie-sample, en het beoordelen van de resulterende model prestaties. Deze informatie wordt vervolgens gebruikt om schattingen van model discriminatie en kalibratie te interpreteren in termen van case-mix verschillen. Om de interpretatie van validatiestudies te verbeteren, werd voorgesteld om onderscheid te maken tussen de reproduceerbaarheid en transporteerbaarheid van een predictiemodel. Tenslotte werd beschreven hoe teleurstellende modelprestaties verbeterd kunnen worden door middel van updating strategieën.

Publication List

Publications in this thesis:

- Abo-Zaid G, Guo B, Deeks JJ, **Debray TPA**, Steyerberg EW, Moons KGM, Riley RD. (2013) Individual participant data meta-analyses should not ignore clustering. *Journal of Clinical Epidemiology*, 66(8): 865–873.
- **Debray TPA**, Koffijberg H, Lu D, Vergouwe Y, Steyerberg EW, Moons KGM. (2012) Incorporating published univariable associations in diagnostic and prognostic modeling. *BMC Medical Research Methodology*, 12(1): 121.
- **Debray TPA**, Koffijberg H, Vergouwe Y, Moons KGM, Steyerberg EW. (2012) Aggregating published prediction models with individual patient data: a comparison of different approaches. *Statistics in Medicine*, 31(23): 2697–2712.
- **Debray TPA**, Koffijberg H, Vergouwe Y, Nieboer D, Steyerberg EW, Moons KGM. (2013) A framework for developing, implementing and evaluating clinical prediction models in an individual participant data meta-analysis. *Statistics in Medicine* 2013, 32(18): 3158–3180.
- **Debray TPA**, Moons KGM, Abo-Zaid G, Koffijberg H, Riley RD. (2013) Individual participant data meta-analysis for a binary outcome: one-stage or two-stage?. *PLoS ONE*, 8(4): e60650.

Other publications:

- Janssen KJ, Siccama I, Vergouwe Y, Koffijberg H, **Debray TPA**, Keijzer M, Grobbee DE, Moons KG. Development and validation of clinical prediction models: Marginal differences between logistic regression, penalized maximum likelihood estimation, and genetic programming. (2012) *Journal of Clinical Epidemiology*, 65, 404–412.

Acknowledgements/Dankwoord

"We must find time to stop and thank the people who make a difference in our lives."

– John F. Kennedy

This thesis could not have been achieved without the help from many friends, colleagues and family. The following pages are dedicated as a redolent sign of gratitude towards them.

Prof. dr. K.G.M. Moons, beste Carl. Het is inmiddels een hele tijd geleden dat we in Utrecht kennis maakten en mijn sollicitatie uitvoerig bespraken. Toen ik een aantal maanden later vernam dat ik effectief als promovendus aan de slag mocht, was ik in de wolken. Hoewel mijn achtergrond in *machine learning* een aanzienlijk voordeel opleverde, was ik niet vertrouwd met vele epidemiologische begrippen en strategieën. Ik had dan ook enige moeite om mijn taalgebruik aan te passen en mezelf te integreren in een volledig nieuwe omgeving. Met name dankzij jouw inzicht, inzet en leiderschap ben ik de afgelopen jaren kunnen ontplooien tot een wetenschapper met een internationaal curicculum. Je maakte ondermeer tijd om me regelmatig bij te sturen, en regelde financiering voor buitenlandse projecten en conferenties. Ik waardeer je bijdragen ten zeerste, en hoop dat ik een waardevolle aanwinst voor het Julius Centrum zal betekenen.

Dr. ir. H. Koffijberg, beste Erik. Uiteraard heb ook jij een grote invloed gespeeld gedurende mijn promotietraject. Je kan erg goed verbanden leggen tussen verschillende onderzoeksdomeinen, dit heeft me erg geholpen tijdens het positioneren van mijn kennis in het Julius Centrum. Daarnaast richtte je je aandacht op vele technische hindernissen, en bood je me inzicht in het bedenken van pragmatische oplossingen. Tot slot betrok je me bij verschillende projecten, zoals de oprichting van een High Performance Computing omgeving. Kortom, jouw sleutelrol heeft mijn wetenschappelijke ontwikkeling erg gesteund en leidde tot vele succesvolle publicaties.

Dr. R.D. Riley, dear Richard. I am very grateful for the opportunity you provided me to collaborate in Birmingham. Within four months, we drafted several articles and attended numerous meetings. You teached me the skills to perform a multivariate meta-analysis, and introduced me to many other prominent scientists. In addition, you familiarized me with British icons such as Doctor Who and Manchester United. It was a great pleasure to meet your family near Macclesfield, and to attend an international football match in Manchester. I hope we can continue sharing wonderful experiences, both profesional and amical.

Prof. dr. E.W. Steyerberg, beste Ewout. Je aarzelde nooit om je expertise in predictie-onderzoek te delen wanneer ik een vraag had. Je kritische houding heeft me erg vooruit geholpen en meer inzicht verleend in het opzetten van simulatiestudies en het beschrijven van technische procedures.

Dr. Y. Vergouwe, beste Yvonne. Je doordachte suggesties hebben de kwaliteit van onze artikels ongetwijfeld erg verbeterd. Daarnaast woonde je vele presentaties bij, en reviseerde je talloze artikels. Bedankt voor je inzet en betrokkenheid!!

Dr. W.M. van der Wal, beste Willem. Ik was al snel gefascineerd door je kennis over R, en jouw ervaringen hebben me erg geïnspireerd. Bedankt voor je tips en de gezellige lunches!

Dr. M.M.G. Leeflang, beste Mariska. Je hebt me veel inzicht verleend in de uitvoering van een meta-analyse, en was ook nauw betrokken bij verschillende conferenties en projecten. Bedankt voor al je hulp!

Dear co-authors: prof. dr. K.G.M. Moons, prof. dr. E.W. Steyerberg, prof. dr. S. van Buuren, prof. dr. J.J. Deeks, dr. ir. H. Koffijberg, dr. Y. Vergouwe, dr. R.D. Riley, dr. I. Klugkist, dr. G.M.A. Abo-Zaid, dr. S. Jolani, dr. K.J.M. Janssen, D. Nieboer, D. Lu, C. Rietbergen, I. Ahmed and B. Guo. Many thanks for your input and valuable contributions!

Beste leden van de beoordelingscommissie, prof. dr. K.C.B. Roes, prof. dr. M.M. Rovers, prof. dr. S. van Buuren en prof. dr. T. Stijnen. Bedankt voor het lezen en beoordelen van mijn manuscript!

Lieve kamergenootjes, dear roommates. Gerdien, Victor, Putri, Liselotte, Stan, Paulien, Madelief, Michelle, Carla, Floriaan, Nanne, Suzanne, Laura, Henk, Joris, Lisette, Sjoukje, Tannie en Peter, ik ben de afgelopen jaren vaak met jullie opgetrokken en dat is me prima bevallen. Vele nieuwtjes gingen niet ongemerkt voorbij, talloze citaten kregen een plekje toegewezen. Tafelvoetbal werd een lokale sport, kilometers koffie werden succesvol volbracht en menige maaltijden werden het middelpunt van de dag. Ik wens jullie veel succes in de toekomst!!

Beste methodologie-collega's: Ewoud, Janneke, Henrieke, Sabrina, Pushpa, Stan, Hans, Linda, Rolf, Erik, Carl, Maarten, Anoukh, Christiana, Peter, Loes, Frederieke, Sanne, Susan, Walter, Kristel, Geert-Jan, Sjoukje, Liselotte, Wouter en vele anderen. Menige maandagen kwamen wij bij elkaar om theoretische onderwerpen te bespreken. Ik heb erg veel van jullie geleerd, en hopelijk ook mijn steentje kunnen bijdragen als toekomstig methodoloog.

Beste Maarten, Floriaan, John, Anoukh, Anneke, Julien, Stavros, Suzette, Sarah, Noor, Willemijn, Nienke, Ewoud, Ruud, Marie, Christiana, Charles, Sanne, Carla, Irene en alle andere promovendi. Bedankt voor de JOB borrels, etentjes, BBQs, housewarming parties, promovenski, vakantie-uitjes, Tivoli en de spannende tafelvoetbal potjes. Mede dankzij jullie waren de afgelopen jaren geweldig!

Beste Annina, Heinie, Coby, Monique, Henk, Wouter, Janet, Petra, Cor, Annet, Sabine en alle andere collega's van het Julius Centrum. Bedankt voor jullie ondersteuning!

Beste Charlotte, Shahab, Maryam, Jesper, Joran, Carel, Herbert en Irene. Bedankt voor jullie gastvrijheid toen ik op jullie afdeling een review uitvoerdde. Ik zal de gezellige lunches niet snel vergeten!

Beste Frits, ik herinner me nog goed toen we na vele jaren terug kennis maakten in Maastricht. De hechte vriendschap die daaruit volgde heeft me een andere kijk op vele dingen in het leven gebracht, en je was de eerste die me in mijn kwaliteiten deed geloven. Zonder jouw steun was ik wellicht nooit aan een doctoraat begonnen!

Beste gym-buddies: Stan, Todor en Elmer. Het was steeds erg gezellig tijdens de kracht- en circuitsessies, waar jullie me vaak tot het uiterste wisten te drijven. Hoewel Jean-Claude Van Damme nog weinig gelijkenissen vertoont, is m'n uithoudings- en incasseringsvermogen de laatste jaren erg verbeterd. Tot binnenkort!

Beste paranimfen Stan en Wouter. Jullie vriendschap heeft me de afgelopen jaren erg gesteund, en jullie waren ook telkens bereid me te helpen wanneer ik een probleem had. Zonder jullie zouden niet alleen de afgelopen vier jaar, maar ook de komende jaren er erg anders uitgezien hebben! Ik heb jullie betrokkenheid erg gewaardeerd!!

Dear Visa, Patama, Amanja, Patrick, Nivedita, Nayara, Dayanne, Gül, Sharmini, Cherwin, Sally, Karyanti, Tim, Ng, Peter, Shahab, Maryam, Judith, Scott and Gudrun. Many thanks for your hospitality and wonderful dinners. I much enjoyed experiencing your cuisines: Dutch, Finnish, Turkish, Persian, Indian, Thai, Indonesian, Malaysian, Chinese, Canadian, Brazilian, Antillean and Icelandic. I also relished our travels across Europe, and camping trips in the Dutch fields. I now fully appreciate the merits of exploring the world!

Dear Nayara, since we met during one of the ING dinners we became good friends and traveled together at several occasions. I will never forget your kindness, and hope to visit you one day in Brazil.

Dear Shima, Deb, Havva, Gudrun, Magda, Nayara, Scott, Maryam, Shahab, Sandrine and all other friends from ING. I much enjoyed our discussions, drinks, dinners, parties and travels. Your friendship means a lot to me and I hope we can stay close in the future.

Dear friends from the Lange Nieuwstraat: Amin, Ela, Joana, Wouter, Susana, Roberto, Saskia, Susanne, Magda, Natasha, Marie, Pradeep, Isa, Anibal, Amila, Vinicius, Anup and Khoi. Over the past few years, we have become a family with a strong relationship. We shared (and still continue to) numerous experiences including dinners, travels, hikes, parties, birthday celebrations and debates about the difference between *liefde* and *verliefdheid* or the boundaries of conscience. Although future demands from us to pursue a distinct path, there is no distance that cannot be traveled and no happiness that cannot be shared (but you may have to find someone else for putting out the glass). I say we meet again very soon!

To all my jitsu friends from Birmingham, including Bethan, Marisa, Syed, Emily, Thibault, Luke, James, Ole, Mat, Maria, Karen, Niall, Theo, Harry, Tom, Jim, Lewis, Kaji, Calvin, Andy, Charles, Phil, Peter, Chris and Hannah. I had a great time with you during the training sessions, the (pre-)gradings and the nationals in Sheffield. I am still very proud of my first achievements in martial arts, and hope to remember my locks and throwing techniques. Dear Bethan, Marisa, Emily and Rose, I enjoyed our journeys, parties and chess games. Thanks for sharing your friendship!! Thibault, it is a great pleasure to find you back here in Utrecht, and to meet your friends Julien and Anna. I look forward to future evenings at Café België, karaoke and other parties!

Dear roommates from the hub: Ikhlaaq, Kym, Danielle, George, Ghada, Helen, Kinga, Mousumi, Piers and Christina. It was a pleasure working with you and learning about your experiences in research. I hope we meet again in the future! I am also very grateful to Jon, Karen, Charlotte, Jennifer, Karla, Lavinia, Lucinda, Hema, Neil, Sue, Hui and all other colleagues from Birmingham University for their support and hospitality. Finally, I much enjoyed the drinks, lunches, dinners and boardgames with Rosanna, Christina, Ruth and Phil. Thanks for your hospitality during my stay in Birmingham!!

To all my other friends including Hendrik, Maarten, Christoph, Pieter, Guillaume, Matthieu, Sven, Machteld, Ineke, Mieke, Sarah, Nathalie, Pascal, Sebastiaan, Stefan, Peter, Marjolein, Robbie en Ben. Thanks for your making time to have fun, to share important phases in your life or even to visit me all the way in the Netherlands!!

Lieve mama, papa, Laura, Sarah, Bram en David. Hoewel ik jullie sinds mijn verhuis naar Nederland niet zo vaak meer zie, word ik steeds warm onthaald wanneer ik me nog eens in het land begeef. Ik heb genoten van de vele feestjes en uitstapjes, maar ook van de lekkere maaltijden en aandacht thuis. Lieve peter, ook jij hebt me erg gesteund bij m'n studies. Mede dankzij jouw inzichten ben ik in Maastricht een epidemiologische richting ingeslagen, en heeft mijn master thesis de grondslag gelegd voor dit proefschrift. Lieve oma, bij jou kon ik steeds terecht voor een kop koffie (sinds ik 5 was) en een mattentaart (toen ik wat ouder was). Je was iemand die graag verhalen vertelde, en ik was iemand die graag luisterde. Het is jammer dat je geestelijke gezondheid de laatste jaren achteruit gegaan is, en dat ik niet meer in Hoeilaart bij je op bezoek kan. Ik zal onze zaterdagochtends echter nooit vergeten. Beste familie, ik besef zeer goed dat het afronden van dit proefschrift nooit mogelijk was geweest zonder jullie steun!

Liefste Welling, het is inmiddels een jaar geleden dat we toenadering zochten tot elkaar en onze vriendschap naar een relatie vertaalden. Hoewel de afgelopen maanden erg turbulent voor je waren, heb ik steeds 100% op je kunnen rekenen en hebben we samen vele mooie momenten beleefd. Je liefde doet me leven, en je wilskracht verbaast me keer op keer. Bedankt voor je betrokkenheid!!

Curriculum Vitae

Thomas P.A. Debray (1985, Leuven, Belgium) studied at the University College of Ghent (2003–2007), the University of Maastricht (2007–2009) and the University of Utrecht (2010–2013), where he respectively attained a bachelor's degree in Applied Informatics, a master's degree in Artificial Intelligence and a master's degree in Clinical Epidemiology. His thesis in Maastricht was performed in collaboration with the Belgian Centre for Evidence-Based Medicine, providing the acquired theoretical concepts with a realistic context and leading to a novel approach for diagnosing diseases with a low prevalence. From 2009–2013 he worked on the present thesis as a PhD student at the Julius Center for health sciences and primary care (University Medical Center Utrecht, the Netherlands) under supervision of Prof. dr. K.G.M. Moons and dr. H. Koffijberg. This thesis was realized in collaboration with the Erasmus Medical Center in Rotterdam under supervision of Prof. dr. E.W. Steyerberg and dr. Y. Vergouwe. Between November 2011 and March 2012, he was a visiting academic at the department of public health, epidemiology and biostatistics at the University of Birmingham (the United Kingdom) under supervision of dr. R.D. Riley.

At present, Thomas P.A. Debray is working as a post-doctoral researcher on methodological aspects related to meta-analysis (Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, the Netherlands).