SPi	Journal Code:	Article ID	Dispatch: 04.01.13	CE: Dumalag, Kristian Arniebel
	S I M	5 7 3 2	No. of Pages: 23	ME:

Statistics in Medicine

Research Article

Received 20 June 2012, Accepted 18 December 2012 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5732

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

Thomas P. A. Debray,[a*†] **Karel G. M. Moons,**[a] **Ikhlaaq Ahmed,**[b] **Hendrik Koffijberg**[a] **and Richard David Riley**[b]

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. Copyright © 2013 John Wiley & Sons, Ltd.

Keywords: prediction research; risk prediction models; meta-analysis; logistic regression; multivariable; individual participant data (IPD); internal–external validation

1. Introduction

Clinical prediction models are an increasingly important tool in evidence-based medical decision making [1–3]. 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 [1, 4–7]. An example is the diagnostic model developed by Oudega *et al.* [6], which aims to predict the presence of deep vein thrombosis (DVT) in patients suspected of DVT 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 [3, 8, 9]. 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

[a] *Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands*
[b] *MRC Midlands Hub for Trials Methodology Research, School of Health and Population Sciences, University of Birmingham, Birmingham, U.K.*
**Correspondence to: Thomas P. A. Debray, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Stratenum 6.131, PO Box 85500, 3508GA Utrecht, The Netherlands.*
[†] *E-mail: T.Debray@umcutrecht.nl*

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 [10, 11]. Examples of IPD-MA that have led (or will lead) to the development and validation of risk prediction models are abound in the literature [7, 12–14].

Prediction models resulting from IPD-MA are appealing as they may be seen to be more generalizable as compared with 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, that is, 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, that is, outcome prevalences (for diagnostic models) or incidences (for prognostic models), or in the predictor–outcome associations [15–17]. Potential causes of such heterogeneity in otherwise related study populations are differences in study design, inclusion and exclusion criteria, disease severity, and interventions undergone [18, 19].

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 [17, 20]. Consequently, the implementation of random effects modeling that effectively account for heterogeneity across the included studies seems highly recommended [11, 21, 22]. 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 [23, 24]. 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 [22]. 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 [12]. 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 (PH) modeling techniques. Although this framework appears to be a useful strategy for accounting and adjusting for heterogeneity in an IPD-MA 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 [20, 25, 26]. Furthermore, we extend the IECV approach proposed by Royston *et al.* [12] 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 DVT.

2. 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. We summarize this framework in Figure 1 and 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 while 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-MA, 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, we consider the value of the framework in the presence of additional heterogeneity in the predictor–outcome associations in Section 4.2.

2.1. *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 the sake of simplicity, we assume that a pre-selection (based on, e.g., prior knowledge or clinical expertise) of the candidate predictors has been carried out and that their specification (e.g., linear or nonlinear 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 [8, 27] 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 are described by K independent predictors, a dichotomous outcome $\boldsymbol{y}$, and contains N_j subjects. Let $\boldsymbol{X}_{ij}$

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 equation 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.

Figure 1. Recommended steps for developing, implementing, and evaluating a risk prediction model when individual participant data from multiple studies are available.

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.

2.1.1. 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:

$$\begin{aligned} y_i &\sim \text{Berno}u\cancel{i}\text{lli}\left(\pi_i\right) \\ \text{logit}\left(\pi_i\right) &= \alpha + \boldsymbol{\beta}'\boldsymbol{X}_i \end{aligned} \tag{1}$$

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 [28].

2.1.2. 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-MA [29–31]. 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 that 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 corresponding logistic regression model consists of $K + 2$ parameters and is specified as follows:

$$\begin{aligned} y_{ij} &\sim \text{Berno}u\cancel{i}\text{lli}\left(\pi_{ij}\right) \\ \text{logit}\left(\pi_{ij}\right) &= a_j + \boldsymbol{\beta}'\boldsymbol{X}_{ij} \;\text{ with }\; a_j \sim \mathcal{N}\left(\alpha, \tau_\alpha^2\right) \end{aligned} \tag{2}$$

By assuming random effects, it becomes possible to model heterogeneity in baseline risk with relatively few parameters. Unfortunately, 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 specialized software packages, which may not always be available [32].

2.1.3. 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.

$$\begin{aligned} y_{ij} &\sim \text{Berno}u\cancel{i}\text{lli}\left(\pi_{ij}\right) \\ \text{logit}\left(\pi_{ij}\right) &= \sum_{m=1}^{M}\left(\alpha_m I_{m=j}\right) + \boldsymbol{\beta}'\boldsymbol{X}_{ij} \end{aligned} \tag{3}$$

where I represents an indicator variable that equals 1 when $m = j$ and 0 otherwise. It is by using an indicator variable to estimate a separate intercept for each study that the normality assumption from expression (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. We further address on how to deal with this in Section 2.2. 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.

2.2. 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) 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 (Section 2.2.1), 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 (Sections 2.2.2 and 2.2.3), whereas the third method ensures intercept optimality by re-estimating the intercept using IPD (Section 2.2.4). All methods can be implemented without the original participant data, as long as some basic information about these data is reported. In summary, 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.

2.2.1. 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 earlier. Royston *et al.* proposed this approach, in which they pool the baseline hazard distribution functions of the studies in an IPD-MA for deriving a prognostic model [12]. 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.

2.2.2. 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 [13]. 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 (Appendix A). 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 [12], given that these estimates are reported.

2.2.3. 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:

$$\hat{\alpha} = \ln\left(\frac{\text{prev}_{\text{new}}}{1 - \text{prev}_{\text{new}}}\right) \tag{4}$$

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 the following:

$$\pi_i = \text{logit}^{-1}\left(\hat{\alpha} + \hat{\boldsymbol{\beta}}'\left(\boldsymbol{X}_i - \overline{\boldsymbol{X}}\right)\right) \tag{5}$$

where the beta estimates are taken from the developed prediction model in step 1 (e.g., the aforementioned stratified or random effects) and the alpha term is from Equation (4).

2.2.4. 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 [8, 17, 20, 33]. 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.

2.3. 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 much 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 because of 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 [2, 17, 34–36]. 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 IECV technique originally proposed by Royston *et al.* [12]. 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 Section 2.1.
3. Choose a model intercept that is appropriate for the validation sample, using one of the approaches described in Section 2.2. 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 [36, 37]. 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 [8, 38, 39]. 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 [8, 40]. 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 [41]. 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 [42].

Results from the IECV technique can be interpreted as follows. In general, if the derived models (i.e., 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 (Section 2.2) 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-MA 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 [43–45]. Although the inclusion of additional covariates, nonlinear 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 [37]. 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-MA 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 [46, 47]. Similarly, if few studies are available in the IPD-MA, 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 four or five 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?) [1, 36, 48] and evaluated alongside the consequences that would result from implementing the model in routine care [49, 50]. 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.

3. 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 modeled using a Poisson model, where expression (1) becomes the following:

$$\begin{gathered} y_i \sim \text{Poisson}\,(\lambda_i) \\ \ln\,(\lambda_i) = \alpha + \boldsymbol{\beta}' \boldsymbol{X}_i \end{gathered} \tag{6}$$

In this expression, α represents the log of the baseline rate and can be modeled using random effects or stratified estimation similar to expressions (2) and (3). This model can further be extended to estimate PH models when time-to-event data are available such that each patient can have a different length of follow-up [51–54]:

$$\begin{gathered} y_i \sim \text{Poisson}\,(\lambda_i) \\ \ln\,(\lambda_i) = \ln(t_i) + \alpha + \boldsymbol{\beta}' \boldsymbol{X}_i \end{gathered} \tag{7}$$

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 period. Although it is possible to relax this assumption by adopting a Cox PH model (which still assumes proportional hazards at all times) [12], there are several limitations to this approach. Most importantly, Cox PH models have an unspecified baseline hazard that hampers prediction of survival times [25, 55, 56]. 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 [57]:

$$h(t|\boldsymbol{X}_i) = g(\boldsymbol{a}, t)e^{\boldsymbol{\beta}'\boldsymbol{X}_i} \tag{8}$$

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, that is, $g(\cdot) = \lambda$. Here, a random baseline hazard effect can be modelled as follows:

$$h(t|\boldsymbol{X}_{ij}) = \zeta_j \lambda e^{\boldsymbol{\beta}'\boldsymbol{X}_{ij}} \text{ with } \zeta_j \sim \Gamma(1, \theta_0) \tag{9}$$

This expression is similar to the Gamma frailty model [58], 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, 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. In addition, 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 modeled using restricted cubic splines within a flexible parametric framework [12, 59]. 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.

4. Case studies

To demonstrate the potential value of the aforementioned approaches for model development, intercept choice, and IECV, we now consider three scenarios that use the IPD of 12 studies conducted for diagnosing 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-MA. We **T1** summarize the studies in Table I, and the studies contained a total of 10,014 patients of which 1897 (18.9%) truly have DVT. The corresponding IPD were collected between 1994 and 2007 in the USA, Sweden, Canada, and the Netherlands.

In each scenario, we apply the three steps of Section 2. 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 (Section 2.2.1). For the stratified approach, three different strategies were evaluated: intercept *selection* based on the outcome proportion in the new study population (Section 2.2.2), intercept *selection* based on similarities of baseline descriptives (Section 2.2.2), and intercept *estimation* based on the outcome proportion observed in the IPD of the new study population (Section 2.2.3). 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 Section 2.3 (proportion of predicted and observed outcomes, average percentage bias of the predictor–outcome associations, and the AUC) and visually inspecting the calibration plots.

We performed all analyses 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 is available on request.

Table I. Baseline descriptives of the IPD in the DVT case study: *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).

ID	*N*	Period	SID_1	SID_2	$sex = 1$ (%)	$surg = 1$ (%)	$malign = 1$ (%)	$calfdif3 = 1$ (%)	$ddimdich = 1$ (%)	$dvt = 1$ (%)	Ref.
1	1028	2005–2007	2	12	37	15	5	30	46	13	[a]
2	814		8	8	38	9	11	43	73	39	[b]
3	153	1994–1996	6	5	48	16	5	39	24	17	[c]
4	1756	1995–1999	11	8	37	16	13	24	52	23	[d]
5	532	2003–2005	9	3	40	19	4	41	75	17	[e]
6	1075	1997–1999	3	5	44	4	5	28	39	18	[f]
7	1768	1994–2001	8	1	38	6	8	20	NA	8	[g]
8	357	2003–2005	2	4	39	5	4	19	69	24	[h]
9	1295	2002–2003	5	11	36	20	6	43	69	22	[i]
10	436	2000–2001	6	12	33	13	6	15	NA	14	[j]
11	541		4	9	44	7	18	30	49	22	[k]
12	259	2002–2006	5	10	33	21	7	41	68	14	[l]

Predictor variables that were not measured in a particular study are indicated by NA. For each study IPD, the ID of the most similar IPD is indicated by SID_1 (similarity based on 12 predictor variables that were measured in all studies and the outcome) and SID_2 (similarity based on outcome prevalence). References are provided in Appendix B.

4.1. *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-MA. In this manner, we ensure validity of the fixed effects assumption for the predictor effects of the proposed methods described in Section 2.1. 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 [60]. The corresponding model is specified as follows and was estimated using all 12 studies:

$$
\begin{aligned}
y_{ij} &\sim \text{Bernouilli}\,(\pi_{ij}) \\
\text{logit}\,(\pi_{ij}) &= [a]_j + [b_{\text{sex}}]_j\,[\mathbf{X}_{\text{sex}}]_{ij} + [b_{\text{surg}}]_j\,[\mathbf{X}_{\text{surg}}]_{ij} \\
\begin{bmatrix} a \\ b_{\text{sex}} \\ b_{\text{surg}} \end{bmatrix}_j &\sim \text{MVN}\left(\begin{bmatrix} \alpha \\ \beta_{\text{sex}} \\ \beta_{\text{surg}} \end{bmatrix}, \begin{bmatrix} \tau_\alpha^2 & \tau_{\alpha\beta_{\text{sex}}} & \tau_{\alpha\beta_{\text{surg}}} \\ \tau_{\alpha\beta_{\text{sex}}} & \tau_{\beta_{\text{sex}}}^2 & \tau_{\beta_{\text{sex}}\beta_{\text{surg}}} \\ \tau_{\alpha\beta_{\text{surg}}} & \tau_{\beta_{\text{sex}}\beta_{\text{surg}}} & \tau_{\beta_{\text{surg}}}^2 \end{bmatrix} \right)
\end{aligned}
\tag{10}
$$

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 2.1. We present results from the IECV in Table II, for each of the stacking, random effects on intercept, and stratified intercept approaches.

4.1.1. 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 II), which indicates that the modeled 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 standard error $= 0.46$) and in study 8 ($\hat{\beta}_{\text{sex,val}} = 0.16$ with standard error $= 0.26$). 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.

4.1.2. 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-MA (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 ($E/O \geqslant 1.2$ in 4 of the 12 validation samples) or underestimate ($E/O \leqslant 0.8$ 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 II). 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) 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

Table II. 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.

		Model performance							
Model development	Model implementation	ID	E/O	b_{overall}	AUC	ID	E/O	b_{overall}	AUC
Stacking	Intercept is overall estimate	1	1.50	0.89 (0.32)	0.58 (0.02)	7	2.60	0.94 (0.30)	0.57 (0.02)
Random effects	Intercept is a weighted average		1.42	0.89 (0.31)	0.58 (0.02)		2.38	0.89 (0.29)	0.57 (0.02)
Stratified	Intercept is selected based on outcome proportion		1.02	0.89 (0.32)	0.58 (0.02)		1.58	0.94 (0.30)	0.57 (0.02)
Stratified	Intercept is selected based on baseline descriptives		3.02	0.89 (0.32)	0.58 (0.02)		2.95	0.94 (0.30)	0.57 (0.02)
Stratified	Intercept is estimated from the outcome proportion		1.03	0.89 (0.32)	0.58 (0.02)		1.03	0.94 (0.30)	0.57 (0.02)
Stacking	Intercept is overall estimate	2	0.43	0.81 (0.23)	0.57 (0.02)	8	0.77	0.82 (0.46)	0.55 (0.03)
Random effects	Intercept is a weighted average		0.42	0.86 (0.24)	0.57 (0.02)		0.72	0.77 (0.45)	0.55 (0.03)
Stratified	Intercept is selected based on outcome proportion		0.62	0.81 (0.23)	0.57 (0.02)		0.96	0.82 (0.46)	0.55 (0.03)
Stratified	Intercept is selected based on baseline descriptives		0.62	0.81 (0.23)	0.57 (0.02)		1.63	0.82 (0.46)	0.55 (0.03)
Stratified	Intercept is estimated from the outcome proportion		1.00	0.81 (0.23)	0.57 (0.02)		1.01	0.82 (0.46)	0.55 (0.03)
Stacking	Intercept is overall estimate	3	1.18	1.32 (0.69)	0.65 (0.06)	9	0.84	0.98 (0.20)	0.58 (0.02)
Random effects	Intercept is a weighted average		1.14	1.22 (0.69)	0.65 (0.06)		0.80	0.98 (0.20)	0.58 (0.02)
Stratified	Intercept is selected based on outcome proportion		1.04	1.32 (0.69)	0.65 (0.06)		0.95	0.98 (0.20)	0.58 (0.02)
Stratified	Intercept is selected based on baseline descriptives		1.04	1.32 (0.69)	0.65 (0.06)		0.76	0.98 (0.20)	0.58 (0.02)
Stratified	Intercept is estimated from the outcome proportion		1.03	1.32 (0.69)	0.65 (0.06)		1.02	0.98 (0.20)	0.58 (0.02)
Stacking	Intercept is overall estimate	4	0.77	1.24 (0.18)	0.60 (0.02)	10	1.36	1.20 (0.39)	0.61 (0.04)
Random effects	Intercept is a weighted average		0.75	1.24 (0.18)	0.60 (0.02)		1.32	1.21 (0.39)	0.61 (0.04)
Stratified	Intercept is selected based on outcome proportion		1.03	1.24 (0.18)	0.60 (0.02)		0.96	1.20 (0.39)	0.61 (0.04)
Stratified	Intercept is selected based on baseline descriptives		0.90	1.24 (0.18)	0.60 (0.02)		1.19	1.20 (0.39)	0.61 (0.04)
Stratified	Intercept is estimated from the outcome proportion		1.02	1.24 (0.18)	0.60 (0.02)		1.03	1.20 (0.39)	0.61 (0.04)
Stacking	Intercept is overall estimate	5	1.14	1.00 (0.35)	0.58 (0.03)	11	0.89	1.08 (0.29)	0.61 (0.03)
Random effects	Intercept is a weighted average		1.09	1.00 (0.35)	0.58 (0.03)		0.85	1.06 (0.29)	0.61 (0.03)
Stratified	Intercept is selected based on outcome proportion		0.95	1.00 (0.35)	0.58 (0.03)		1.04	1.08 (0.29)	0.61 (0.03)
Stratified	Intercept is selected based on baseline descriptives		1.31	1.00 (0.35)	0.58 (0.03)		1.10	1.08 (0.29)	0.61 (0.03)
Stratified	Intercept is estimated from the outcome proportion		1.03	1.00 (0.35)	0.58 (0.03)		1.02	1.08 (0.29)	0.61 (0.03)
Stacking	Intercept is overall estimate	6	1.14	0.84 (0.22)	0.59 (0.02)	12	1.41	0.76 (0.53)	0.55 (0.05)
Random effects	Intercept is a weighted average		1.09	0.84 (0.22)	0.59 (0.02)		1.37	0.76 (0.53)	0.55 (0.05)
Stratified	Intercept is selected based on outcome proportion		1.00	0.84 (0.22)	0.59 (0.02)		1.03	0.76 (0.53)	0.55 (0.05)
Stratified	Intercept is selected based on baseline descriptives		0.95	0.84 (0.22)	0.59 (0.02)		1.23	0.76 (0.53)	0.55 (0.05)
Stratified	Intercept is estimated from the outcome proportion		1.03	0.84 (0.22)	0.59 (0.02)		1.03	0.76 (0.53)	0.55 (0.05)

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 each measure is indicated between brackets.

the included predictor variables (cf. Section 2.2.3). Here, the E/O statistic is always close to 1, ranging between 1.00 and 1.03.

4.1.3. Quality of model predictions. Visual inspection of the calibration plots (Figure A.1 in the 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 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 studies 3 and 8, where the least data were available. Calibration curves for other approaches were similar (results not included), with curves shifted upward and downward according to underestimation ($E/O < 1$) and overestimation ($E/O > 1$), respectively, of the outcome presence. Evaluation of the AUC 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% to 22% and 13% to 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.

4.1.4. 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 with 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.

4.2. 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 [24]. 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 Section 4.1, we found $\hat{\alpha} = -3.98\ (\hat{\tau}_{\alpha} = 0.31)$, $\hat{\beta}_{\text{malign}} = 0.38\ \left(\hat{\tau}_{\beta_{\text{malign}}} = 0.35\right)$, $\hat{\beta}_{\text{calfdif3}} = 1.05\ \left(\hat{\tau}_{\beta_{\text{calfdif3}}} = 0.16\right)$, $\hat{\beta}_{\text{surg}} = 0.25\ \left(\hat{\tau}_{\beta_{\text{surg}}} = 0.09\right)$, and $\hat{\beta}_{\text{ddimdich}} = 2.76\ \left(\hat{\tau}_{\beta_{\text{ddimdich}}} = 0.41\right)$. Clearly, the heterogeneity estimates (τ values) are quite large for most variables. We present results from the IECV in Table III. **T3**

Results in Table III 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 overestimate 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 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.

Table III. 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.

		Model performance							
Model development	Model implementation	ID	E/O	b_{overall}	AUC	ID	E/O	b_{overall}	AUC
Stacking	Intercept is overall estimate	1	1.42	0.88 (0.09)	0.80 (0.02)	6	0.87	0.95 (0.07)	0.84 (0.02)
Random effects	Intercept is a weighted average		1.35	0.84 (0.09)	0.80 (0.02)		0.83	0.92 (0.07)	0.84 (0.02)
Stratified	Intercept is selected based on outcome proportion		0.61	0.88 (0.09)	0.80 (0.02)		0.50	0.95 (0.07)	0.84 (0.02)
Stratified	Intercept is selected based on baseline descriptives		1.43	0.88 (0.09)	0.80 (0.02)		1.14	0.95 (0.07)	0.84 (0.02)
Stratified	Intercept is estimated from the outcome proportion		1.68	0.88 (0.09)	0.80 (0.02)		1.49	0.95 (0.07)	0.84 (0.02)
Stacking	Intercept is overall estimate	2	0.68	1.02 (0.09)	0.76 (0.02)	8	1.00	1.15 (0.19)	0.77 (0.02)
Random effects	Intercept is a weighted average		0.67	0.99 (0.09)	0.76 (0.02)		0.96	1.09 (0.18)	0.77 (0.02)
Stratified	Intercept is selected based on outcome proportion		0.67	1.02 (0.09)	0.76 (0.02)		1.14	1.15 (0.19)	0.77 (0.02)
Stratified	Intercept is selected based on baseline descriptives		0.80	1.02 (0.09)	0.76 (0.02)		1.34	1.15 (0.19)	0.77 (0.02)
Stratified	Intercept is estimated from the outcome proportion		1.13	1.02 (0.09)	0.76 (0.02)		1.33	1.15 (0.19)	0.77 (0.02)
Stacking	Intercept is overall estimate	3	0.76	1.35 (0.22)	0.92 (0.03)	9	1.25	0.95 (0.09)	0.76 (0.01)
Random effects	Intercept is a weighted average		0.71	1.32 (0.22)	0.92 (0.03)		1.17	0.93 (0.08)	0.76 (0.01)
Stratified	Intercept is selected based on outcome proportion		0.44	1.35 (0.22)	0.92 (0.03)		1.27	0.95 (0.09)	0.76 (0.01)
Stratified	Intercept is selected based on baseline descriptives		0.81	1.35 (0.22)	0.92 (0.03)		0.70	0.95 (0.09)	0.76 (0.01)
Stratified	Intercept is estimated from the outcome proportion		1.39	1.35 (0.22)	0.92 (0.03)		1.37	0.95 (0.09)	0.76 (0.01)
Stacking	Intercept is overall estimate	4	0.80	1.16 (0.07)	0.85 (0.01)	11	0.90	0.96 (0.10)	0.83 (0.02)
Random effects	Intercept is a weighted average		0.78	1.15 (0.07)	0.85 (0.01)		0.84	0.92 (0.10)	0.83 (0.02)
Stratified	Intercept is selected based on outcome proportion		0.80	1.16 (0.07)	0.85 (0.01)		0.70	0.96 (0.10)	0.83 (0.02)
Stratified	Intercept is selected based on baseline descriptives		0.88	1.16 (0.07)	0.85 (0.01)		1.03	0.96 (0.10)	0.83 (0.02)
Stratified	Intercept is estimated from the outcome proportion		1.35	1.16 (0.07)	0.85 (0.01)		1.41	0.96 (0.10)	0.83 (0.02)
Stacking	Intercept is overall estimate	5	1.66	1.01 (0.17)	0.73 (0.02)	12	1.95	0.97 (0.24)	0.76 (0.04)
Random effects	Intercept is a weighted average		1.61	0.99 (0.16)	0.73 (0.02)		1.95	0.96 (0.24)	0.76 (0.04)
Stratified	Intercept is selected based on outcome proportion		2.03	1.01 (0.17)	0.73 (0.02)		1.34	0.97 (0.24)	0.76 (0.04)
Stratified	Intercept is selected based on baseline descriptives		1.27	1.01 (0.17)	0.73 (0.02)		1.51	0.97 (0.24)	0.76 (0.04)
Stratified	Intercept is estimated from the outcome proportion		1.40	1.01 (0.17)	0.73 (0.02)		1.57	0.97 (0.24)	0.76 (0.04)

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 each measure is indicated between brackets.

Table IV. 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.

		Model performance							
Model development	Model implementation	ID	E/O	b_{overall}	AUC	ID	E/O	b_{overall}	AUC
Stacking	Intercept is overall estimate	1	1.51	0.89 (0.13)	0.68 (0.02)	7	2.27	1.04 (0.13)	0.69 (0.02)
Random effects	Intercept is a weighted average		1.41	0.90 (0.14)	0.68 (0.02)		2.07	1.01 (0.12)	0.69 (0.02)
Stratified	Intercept is selected based on outcome proportion		0.91	0.89 (0.13)	0.68 (0.02)		1.38	1.04 (0.13)	0.69 (0.02)
Stratified	Intercept is selected based on baseline descriptives		2.73	0.89 (0.13)	0.68 (0.02)		2.65	1.04 (0.13)	0.69 (0.02)
Stratified	Intercept is estimated from the outcome proportion		1.15	0.89 (0.13)	0.68 (0.02)		1.15	1.04 (0.13)	0.69 (0.02)
Stacking	Intercept is overall estimate	2	0.51	0.74 (0.10)	0.65 (0.02)	8	0.76	0.92 (0.18)	0.66 (0.03)
Random effects	Intercept is a weighted average		0.49	0.74 (0.10)	0.65 (0.02)		0.71	0.93 (0.18)	0.66 (0.03)
Stratified	Intercept is selected based on outcome proportion		0.70	0.74 (0.10)	0.65 (0.02)		1.02	0.92 (0.18)	0.66 (0.03)
Stratified	Intercept is selected based on baseline descriptives		0.70	0.74 (0.10)	0.65 (0.02)		1.46	0.92 (0.18)	0.66 (0.03)
Stratified	Intercept is estimated from the outcome proportion		1.03	0.74 (0.10)	0.65 (0.02)		1.08	0.92 (0.18)	0.66 (0.03)
Stacking	Intercept is overall estimate	3	1.28	1.36 (0.35)	0.76 (0.06)	9	0.98	0.97 (0.10)	0.69 (0.02)
Random effects	Intercept is a weighted average		1.23	1.37 (0.35)	0.76 (0.06)		0.92	0.99 (0.10)	0.69 (0.02)
Stratified	Intercept is selected based on outcome proportion		1.02	1.36 (0.35)	0.76 (0.06)		1.10	0.97 (0.10)	0.69 (0.02)
Stratified	Intercept is selected based on baseline descriptives		1.18	1.36 (0.35)	0.76 (0.06)		0.78	0.97 (0.10)	0.69 (0.02)
Stratified	Intercept is estimated from the outcome proportion		1.14	1.36 (0.35)	0.76 (0.06)		1.10	0.97 (0.10)	0.69 (0.02)
Stacking	Intercept is overall estimate	4	0.70	1.35 (0.09)	0.74 (0.01)	10	1.12	0.68 (0.21)	0.64 (0.04)
Random effects	Intercept is a weighted average		0.69	1.43 (0.10)	0.74 (0.01)		1.06	0.69 (0.21)	0.64 (0.04)
Stratified	Intercept is selected based on outcome proportion		0.99	1.35 (0.09)	0.74 (0.01)		0.68	0.68 (0.21)	0.64 (0.04)
Stratified	Intercept is selected based on baseline descriptives		0.87	1.35 (0.09)	0.74 (0.01)		1.01	0.68 (0.21)	0.64 (0.04)
Stratified	Intercept is estimated from the outcome proportion		1.07	1.35 (0.09)	0.74 (0.01)		1.10	0.68 (0.21)	0.64 (0.04)
Stacking	Intercept is overall estimate	5	1.28	0.72 (0.16)	0.65 (0.03)	11	0.88	0.91 (0.14)	0.70 (0.02)
Random effects	Intercept is a weighted average		1.22	0.74 (0.16)	0.65 (0.03)		0.83	0.92 (0.14)	0.70 (0.02)
Stratified	Intercept is selected based on outcome proportion		1.00	0.72 (0.16)	0.65 (0.03)		0.91	0.91 (0.14)	0.70 (0.02)
Stratified	Intercept is selected based on baseline descriptives		1.30	0.72 (0.16)	0.65 (0.03)		1.17	0.91 (0.14)	0.70 (0.02)
Stratified	Intercept is estimated from the outcome proportion		1.14	0.72 (0.16)	0.65 (0.03)		1.10	0.91 (0.14)	0.70 (0.02)
Stacking	Intercept is overall estimate	6	1.10	1.02 (0.11)	0.70 (0.02)	12	1.59	0.91 (0.25)	0.67 (0.05)
Random effects	Intercept is a weighted average		1.04	1.03 (0.11)	0.70 (0.02)		1.53	0.92 (0.26)	0.67 (0.05)
Stratified	Intercept is selected based on outcome proportion		0.87	1.02 (0.11)	0.70 (0.02)		1.43	0.91 (0.25)	0.67 (0.05)
Stratified	Intercept is selected based on baseline descriptives		0.85	1.02 (0.11)	0.70 (0.02)		1.25	0.91 (0.25)	0.67 (0.05)
Stratified	Intercept is estimated from the outcome proportion		1.12	1.02 (0.11)	0.70 (0.02)		1.17	0.91 (0.25)	0.67 (0.05)

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 each measure is indicated between brackets.

In conclusion, when predictor–outcome associations in the IPD-MA 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.

4.3. *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

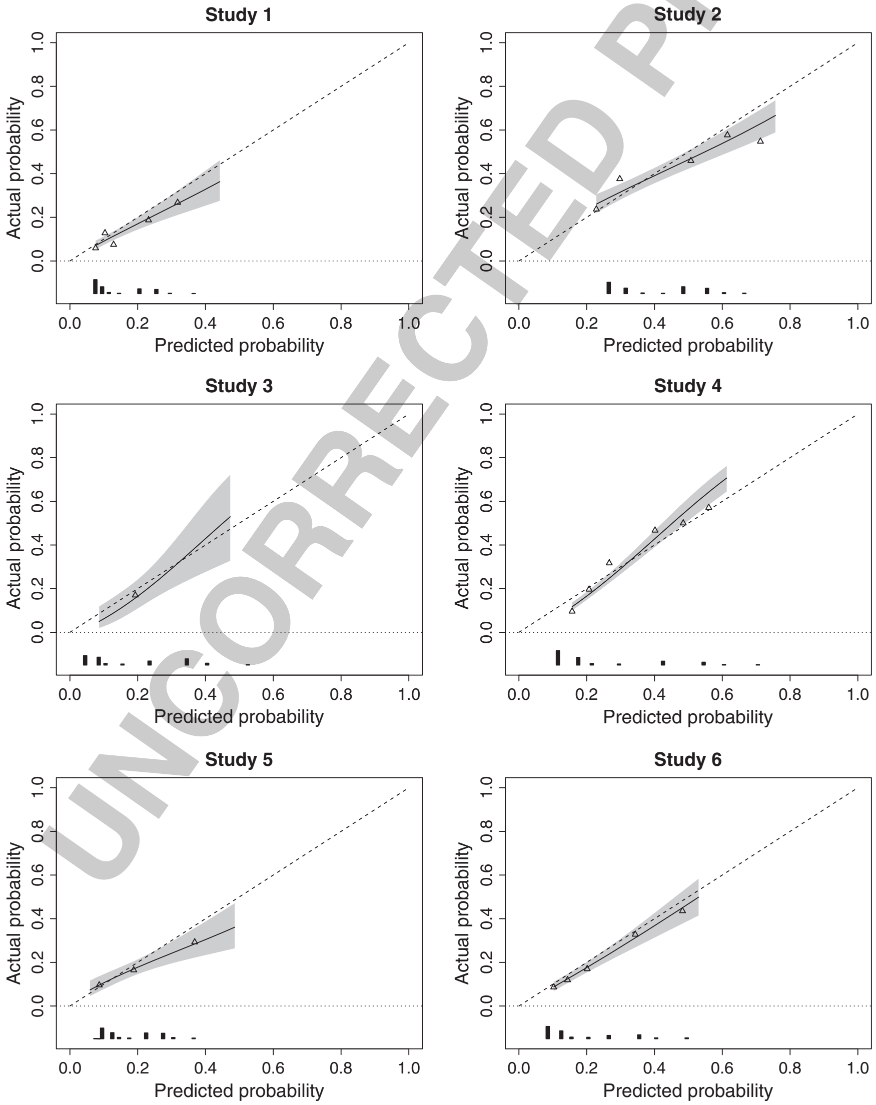


Figure 2. Calibration plots of models derived by stratified estimation of the intercept (where the final intercept is estimated from the outcome proportion in the validation study) in the validation studies of case study 3. The triangles indicate groups of observations with similar predicted probabilities and their corresponding outcome proportion. Note that a maximum of eight groups can be generated because the included predictor variables *sex*, *surg*, and *calfdif3* are dichotomous.

Case Study 1 and one strong predictor *calfdif3* from case study 2. By performing a full multivariate random effects meta-analysis similar to Section 4.1, 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. T4 We present results from the IECV in Table IV, and the results 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 F2 F3 of the calibration plots (Figures 2 and 3) 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

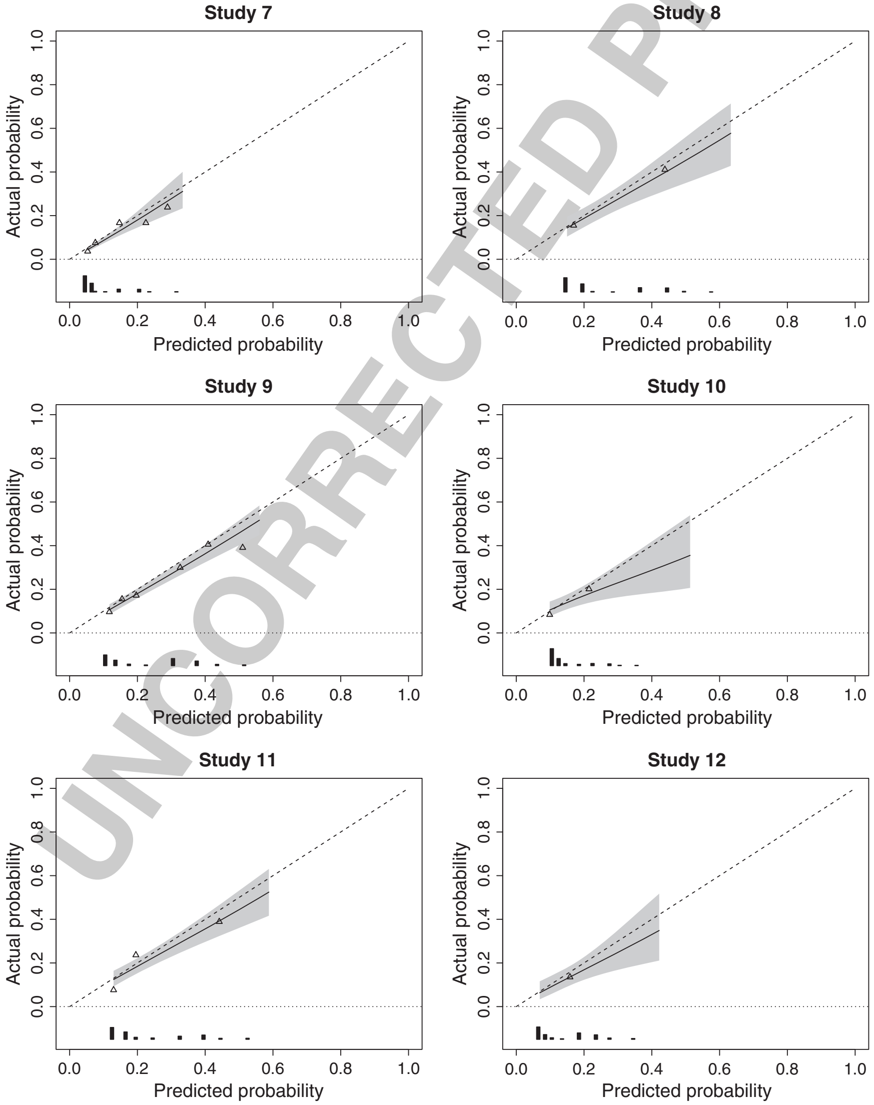


Figure 3. Calibration plots of models derived by stratified estimation of the intercept (where the final intercept is estimated from the outcome proportion in the validation study) in the validation studies of case study 3. The triangles indicate groups of observations with similar predicted probabilities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

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.

5. 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 II), 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 (Section 2.2.3). 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 (Section 2.2.2). 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 (Sections 2.1.1 and 2.1.2). In such situations, the IECV technique may be particularly helpful to identify the generalizability of derived prediction models across other study populations [12]. 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.

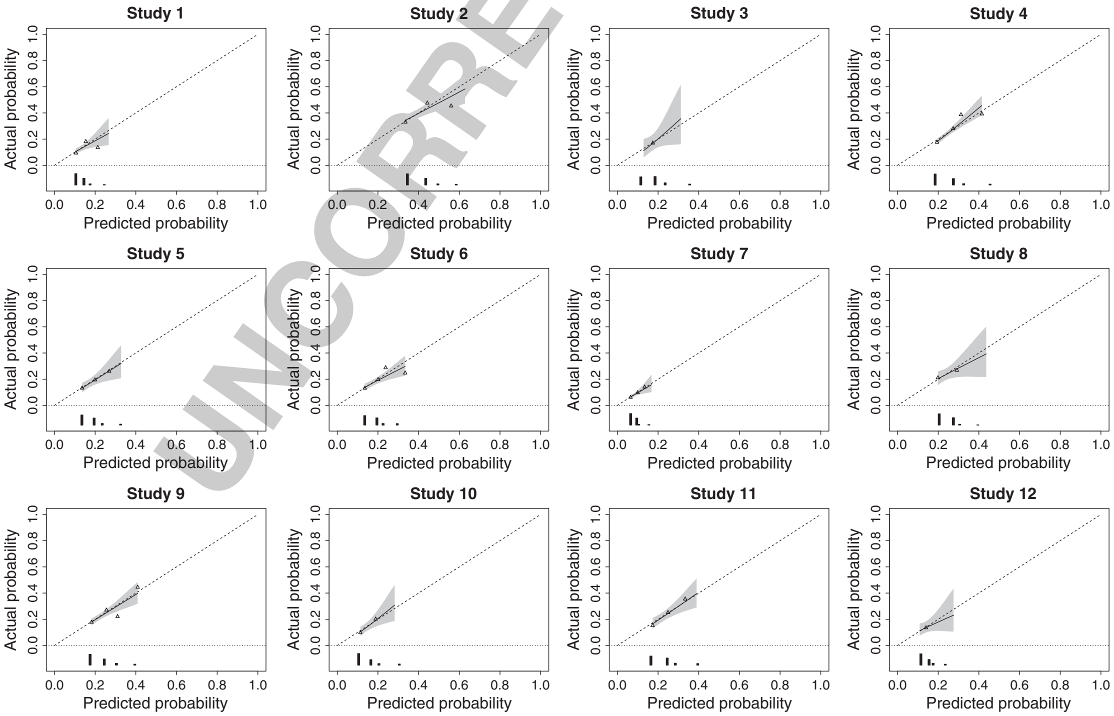
Some important limitations need to be considered to fully appraise the findings of this study. Firstly, the inclusion of homogeneous predictors may not always yield highly discriminative prediction models. Weakly heterogeneous but strong predictors may therefore be included to improve discrimination at the cost of model calibration. Heterogeneity may further be reduced by including additional covariates, nonlinear associations, or interaction terms, or by applying bootstrap and shrinkage techniques [8, 61–64]. Secondly, when many but relatively small studies are available, stratified estimation may no longer be feasible because of its inherent model complexity. In such scenarios, random intercept effects modeling may considerably reduce the amount of unknown parameters while 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 are 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 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.

Appendix A

Comparison of baseline descriptives to select a model intercept.

1. For each predictor and/or outcome,
   (a) calculate difference in mean (continuous variables) or proportion (discrete variables) of observed individuals for each study.
   (b) assign a rank for each study according to similarity, where increasing ranks indicate a decreasing distance (i.e., more similarity).
2. For each study, calculate the median rank over all variables.
3. Select the model intercept from the study with the largest median rank. Alternatively, it is possible to weight estimated intercepts according to the achieved ranks.



Q1 **Figure A.1.** 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 1. The triangles indicate groups of observations with similar predicted probabilities and their corresponding outcome proportion. Note that a maximum of four distinct groups can be generated because the included predictor variables *sex* and *surg* are dichotomous.

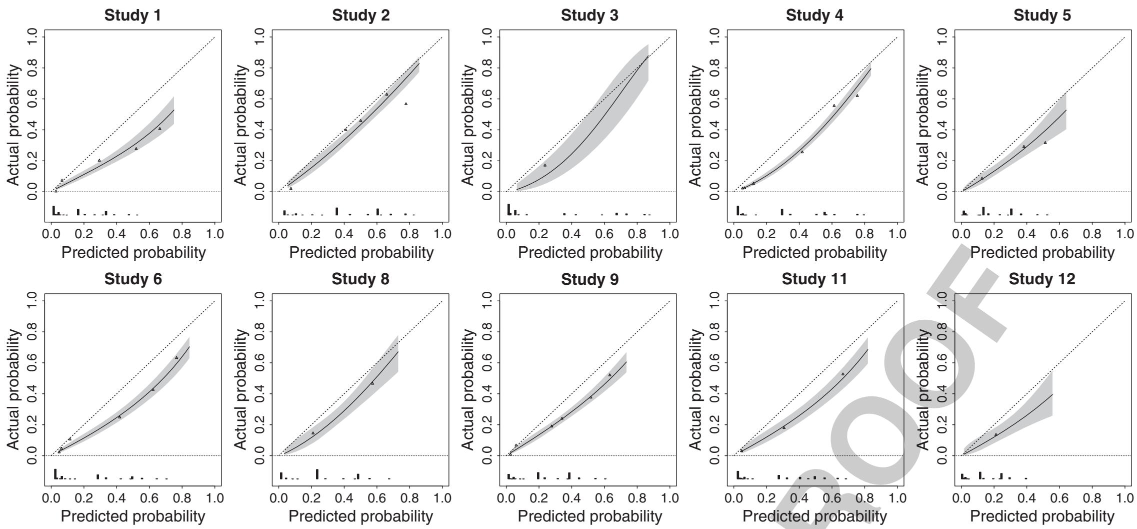


Figure A.2. Calibration plots of models derived by stratified estimation of the intercept (where the final intercept is estimated from the outcome proportion in the validation study) in the validation studies of case study 2. The triangles indicate groups of observations with similar predicted probabilities and their corresponding outcome proportion. Note that a maximum of 16 groups can be generated because the included predictor variables *malign*, *calfdif3*, *surg*, and *ddimdich* are dichotomous.

Appendix B

Reference list of studies used in the case studies.

a. Büller HR, Ten Cate-Hoek AJ, Hoes AW, Joore MA, Moons KGM, Oudega R, Prins MH, Stoffers HEJH, Toll DB, van der Velde EF, van Weert HCPM. Safely ruling out deep venous thrombosis in primary care. *Annals of Internal Medicine* 2009; **150**(4):229–235.

b. Schutgens REG, Ackermark P, Haas FJLM, Nieuwenhuis HK, Peltenburg HG, Pijlman AH, Pruijm M, Oltmans R, Kelder JC, Biesma DH. 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* Feb 2003; **107**(4):593–607, doi:10.1161/01.CIR.0000045670.12988.1E.

c. Anderson DR, Wells PS, Stiell I, MacLeod B, Simms M, Gray L, Robinson KS, Bormanis J, Mitchell M, Lewandowski B, Flowerdew G. 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* Oct 2000; **19**(3):225–230.

d. Kraaijenhagen RA, Piovella F, Bernardi E, Verlato F, Beckers EAM, Koopman MMW, Barone M, Camporese G, Potter Van Loon BJ, Prins MH, Prandoni P, Büller HR. Simplification of the diagnostic management of suspected deep vein thrombosis. *Archives of Internal Medicine* Apr 2002; **162**(8):907–911.

e. Toll DB, Oudega R, Bulten RJ, Hoes AW, Moons KGM. Excluding deep vein thrombosis safely in primary care. *The Journal of Family Practice* Jul 2006; **55**(7):613–618.

f. Anderson DR, Kovacs MJ, Kovacs G, Stiell I, Mitchell M, Khoury V, Dryer J, Ward J, Wells PS. 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* Apr 2003; **1**(4):645–651.

g. Andreou ER, Koru-Sengul T, Linkins L, Bates SM, Ginsberg JS, Kearon C. Differences in clinical presentation of deep vein thrombosis in men and women. *Journal of Thrombosis and Haemostasis* Oct 2008; **6**(10):1713–1719, doi:10.1111/j.1538-7836.2008.03110.x.

h. 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* Feb 2009; **123**(4):612–616, doi:10.1016/j.thromres.2008.04.007.

i. 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* Jul 2005; **94**(1):200–205, doi:10.1160/TH04-12-0829.

j. 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* Jun 2004; **140**(12):985–991.

k. Wells PS, Anderson DR, Rodger M, Forgie M, Kearon C, Dreyer J, Kovacs G, Mitchell M, Lewandowski B, Kovacs MJ. Evaluation of D-dimer in the diagnosis of suspected deep-vein thrombosis. *The New England Journal of Medicine* Sep 2003; **349**(13):1227–1235, doi:10.1056/NEJMoa023153.

l. 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* Feb 2008; **25**(1):3–8, doi:10.1093/fampra/cmm075.

Acknowledgements

While undertaking this work, T. D., I. A., and R. R. were supported by the MRC Midlands Hub for Trials Methodology Research (Medical Research Council Grant ID G0800808), at the University of Birmingham in the UK. We also gratefully acknowledge the financial support by the Netherlands Organization for Scientific Research (9120.8004, 918.10.615, and 916.11.126). Finally, we thank the Editor, Associate Editor, and two anonymous reviewers for their constructive comments that helped improve this article.

References

1. 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.
2. Moons KGM, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, Grobbee DE. 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.
3. Q2 Adams ST, Leveson SH. Clinical prediction rules. *British Medical Journal* 2012; **344**:d8312. DOI: 10.1136/bmj.d8312.
4. 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.
5. Q3 Conroy RM, Pyörälä K, Fitzgerald AP, Sans S, Menotti A, De Backer G, De Bacquer D, Ducimetière P, Jousilahti P, Keil U, Njølstad I, Oganov RG, Thomsen T, Tunstall-Pedoe H, Tverdal A, Wedel H, Whincup P, Wilhelmsen L, Graham IM, SCORE project group. 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.
6. 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.
7. Schuit E, Kwee A, Westerhuis MEMH, Van Dessel HJHM, Graziosi GCM, Van Lith JMM, Nijhuis JG, Oei SG, Oosterbaan HP, Schuitemaker NW, Wouters MGAJ, Visser GHA, Mol BWJ, Moons KGM, Groenwold RHH. 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.
8. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer: New York, 2009.
9. 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.
10. 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.
11. 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.
12. 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.
13. Steyerberg EW, Mushkudiani N, Perel P, Butcher I, Lu J, McHugh GS, Murray GD, Marmarou A, Roberts I, Habbema JDF, Maas AIR. 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.
14. Phillips RS, Sutton AJ, Riley RD, Chisholm JC, Picton SV, Stewart LA, the PICNICC Collaboration. 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.
15. 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.

16. 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.
17. Moons KGM, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, Woodward M. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* 2012; **98**(9):691–698. DOI: 10.1136/heartjnl-2011-301247.
18. 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.
19. Hailpern SM, Visintainer PF. Odds ratios and logistic regression: further examples of their use and interpretation. *The Stata Journal* 2003; **3**(3):213–225.
20. 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.
21. Greenland S. Principles of multilevel modelling. *International Journal of Epidemiology* 2000; **29**(1):158–167. DOI: 10.1093/ije/29.1.158.
22. Riley RD, Higgins JPT, Deeks JJ. Interpretation of random effects meta-analyses. *British Medical Journal* 2011; **342**:d549. DOI: 10.1136/bmj.d549.
23. 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.
24. 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.
25. 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.
26. 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.
27. Harrell FE Jr. *Regression Modeling Strategies with Applications to Linear Models, Logistic Regression and Survival Analysis*, Springer Series in Statistics. Springer, 2001. Q4
28. Abo-Zaid GMA, Guo B, Deeks JJ, Debray TPA, Steyerberg EW, Moons KGM, Riley RD. Individual participant data meta-analyses should not ignore clustering. *Journal of Clinical Epidemiology* 2012; **Submitted**. Q5
29. Böhning D, Sarol J Jr. 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.
30. 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.
31. Steenbergen MR, Jones BS. Modeling multilevel data structures. *American Journal of Political Science* 2002; **46**(1):218–237.
32. Lee KJ, Thompson SG. Flexible parametric models for random-effects distributions. *Statistics in Medicine* 2008; **27**(3):418–434. DOI: 10.1002/sim.2897.
33. 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.
34. Altman DG, Royston P. What do we mean by validating a prognostic model? *Statistics in Medicine* 2000; **19**(4):453–473. DOI: 10.1002/(SICI)1097-0258(20000229)19:4⟨453::AID-SIM350⟩3.0.CO;2-5.
35. 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.
36. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010; **21**(1):128–138. DOI: 10.1097/EDE.0b013e3181c30fb2.
37. 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.
38. Cox DR. Two further applications of a model for binary regression. *Biometrika* 1958; **45**(3):562–565.
39. Miller ME, Langefeld CD, Tierney WM, Hui SL, McDonald CJ. Validation of probabilistic predictions. *Medical Decision Making* 1993; **13**(1):49–58.
40. Spiegelhalter DJ. Probabilistic prediction in patient management and clinical trials. *Statistics in Medicine* 1986; **5**(5):421–433.
41. Harrell FJ, Lee K, Califf R, Pryor D, Rosati R. Regression modelling strategies for improved prognostic prediction. *Statistics in Medicine* 1984; **3**:143–152. DOI: 10.1002/sim.4780030207.
42. 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.
43. 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.
44. 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.
45. 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.
46. 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.

47. 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.

48. Gerds TA, Cai T, Schumacher M. The performance of risk prediction models. *Biometrical Journal* 2008; **50**(4):457–479. DOI: 10.1002/bimj.200810443.

49. 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.

50. 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.

Q6 51. 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.

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

53. Ma R, Krewski D, Burnett RT. Random effects Cox models: a Poisson modelling approach. *Biometrika* 2003; **90**(1):157–169.

54. 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.

55. Royston P. Estimating a smooth baseline hazard function for the Cox model. *Technical Report*, University College London, 2011.

56. Royston P, Altman D. External validation of a Cox prognostic model: principles and methods. *Statistics in Medicine* 2011; **Submitted**.

57. 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.

58. 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.

59. 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.

60. 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.

61. 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.

62. 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.

63. 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.

64. 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.

Author Query Form

Journal: Statistics in Medicine

Article: sim_5732

Dear Author,

During the copyediting of your paper, the following queries arose. Please respond to these by annotating your proofs with the necessary changes/additions.

- If you intend to annotate your proof electronically, please refer to the E-annotation guidelines.
- If you intend to annotate your proof by means of hard-copy mark-up, please refer to the proof mark-up symbols guidelines. If manually writing corrections on your proof and returning it by fax, do not write too close to the edge of the paper. Please remember that illegible mark-ups may delay publication.

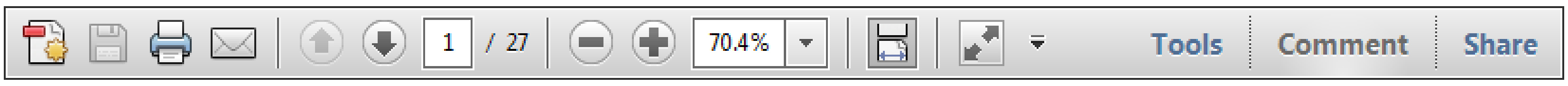
Whether you opt for hard-copy or electronic annotation of your proofs, we recommend that you provide additional clarification of answers to queries by entering your answers on the query sheet, in addition to the text mark-up.

Query No.	Query	Remark
Q1	AUTHOR: Please check that the appendix figures have been labeled correctly.	💬
Q2	AUTHOR: If References 3, 9, 11, 13, 14, 22, 35, 45, and 54 are not one-page articles, please supply the first and last pages for this articles.	
Q3	AUTHOR: Please check list of authors for References 5, 7, and 13 if presented correctly.	
Q4	AUTHOR: Please provide the city location of publisher for Reference 27.	
Q5	AUTHOR: Please provide updates for Reference 28 and 56.	
Q6	AUTHOR: Please provide volume for Reference 51.	

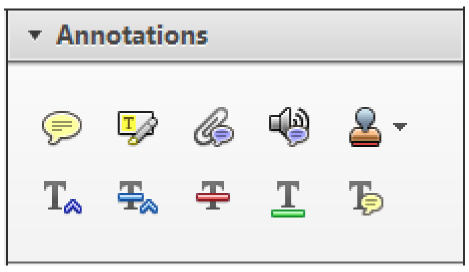
USING e-ANNOTATION TOOLS FOR ELECTRONIC PROOF CORRECTION

Required software to e-Annotate PDFs: Adobe Acrobat Professional or Adobe Reader (version 7.0 or above). (Note that this document uses screenshots from Adobe Reader X)
The latest version of Acrobat Reader can be downloaded for free at: http://get.adobe.com/uk/reader/

Once you have Acrobat Reader open on your computer, click on the Comment tab at the right of the toolbar:



This will open up a panel down the right side of the document. The majority of tools you will use for annotating your proof will be in the Annotations section, pictured opposite. We've picked out some of these tools below:

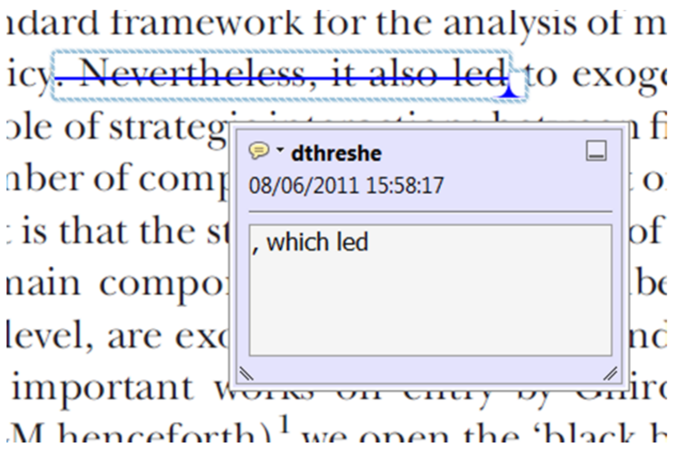


1. Replace (Ins) Tool – for replacing text.

Strikes a line through text and opens up a text box where replacement text can be entered.

How to use it

- Highlight a word or sentence.
- Click on the Replace (Ins) icon in the Annotations section.
- Type the replacement text into the blue box that appears.



ıdard framework for the analysis of m
icy~~. Nevertheless, it also led~~ to exoge
ple of strateg
nber of comp
is that the st
nain compo
level, are exc
important
M henceforth)¹ we open the 'black b

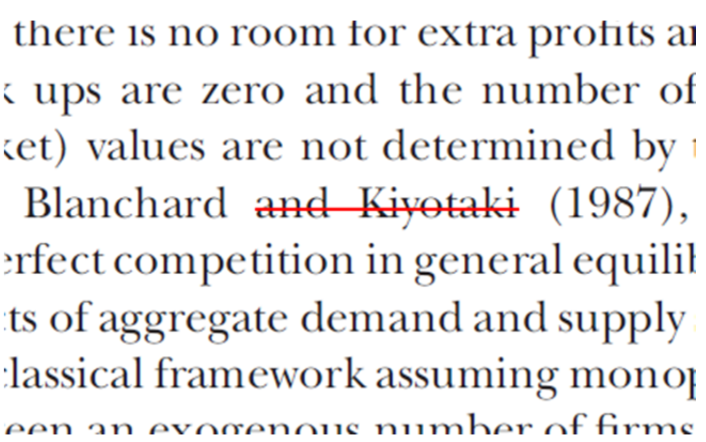
dthreshe
08/06/2011 15:58:17
, which led

2. Strikethrough (Del) Tool – for deleting text.

Strikes a red line through text that is to be deleted.

How to use it

- Highlight a word or sentence.
- Click on the Strikethrough (Del) icon in the Annotations section.



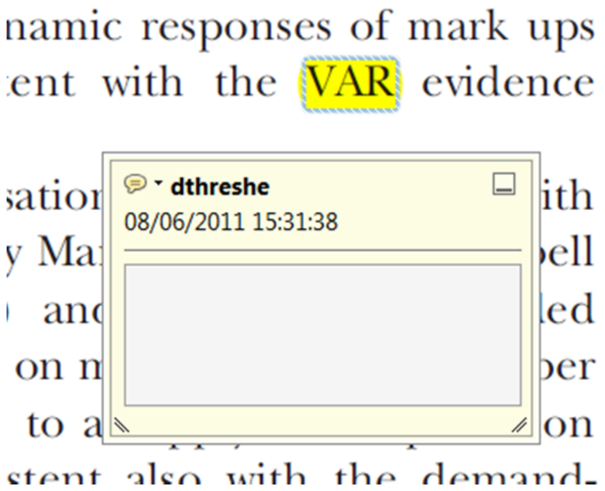
there is no room for extra profits ar
ups are zero and the number of
et) values are not determined by
Blanchard ~~and Kiyotaki~~ (1987),
rfect competition in general equilil
ts of aggregate demand and supply
lassical framework assuming mono
een an exogenous number of firms

3. Add note to text Tool – for highlighting a section to be changed to bold or italic.

Highlights text in yellow and opens up a text box where comments can be entered.

How to use it

- Highlight the relevant section of text.
- Click on the Add note to text icon in the Annotations section.
- Type instruction on what should be changed regarding the text into the yellow box that appears.



namic responses of mark ups
ent with the VAR evidence

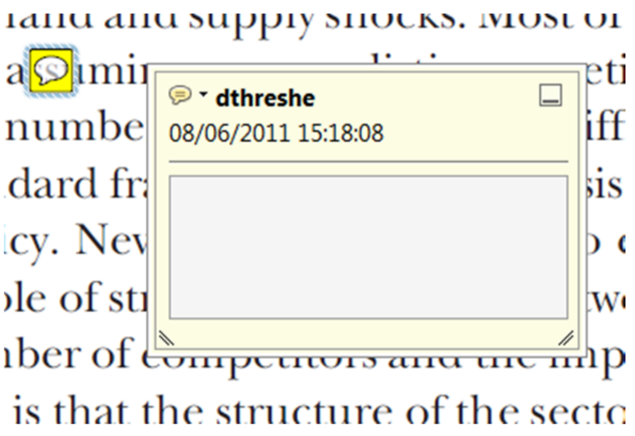
dthreshe
08/06/2011 15:31:38

4. Add sticky note Tool – for making notes at specific points in the text.

Marks a point in the proof where a comment needs to be highlighted.

How to use it

- Click on the Add sticky note icon in the Annotations section.
- Click at the point in the proof where the comment should be inserted.
- Type the comment into the yellow box that appears.



and and supply shocks. Most of
is that the structure of the secto

dthreshe
08/06/2011 15:18:08

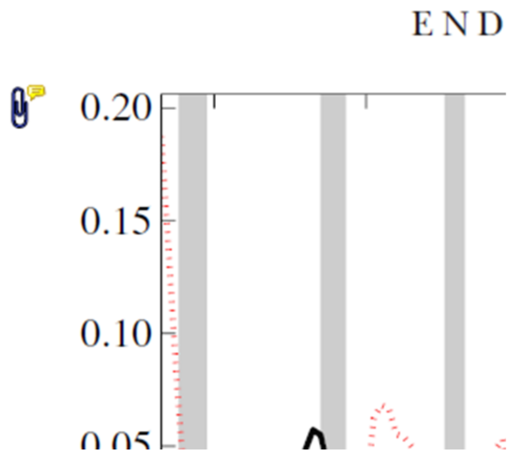
USING e-ANNOTATION TOOLS FOR ELECTRONIC PROOF CORRECTION

5. Attach File Tool – for inserting large amounts of text or replacement figures.

Inserts an icon linking to the attached file in the appropriate pace in the text.

How to use it

- Click on the Attach File icon in the Annotations section.
- Click on the proof to where you'd like the attached file to be linked.
- Select the file to be attached from your computer or network.
- Select the colour and type of icon that will appear in the proof. Click OK.

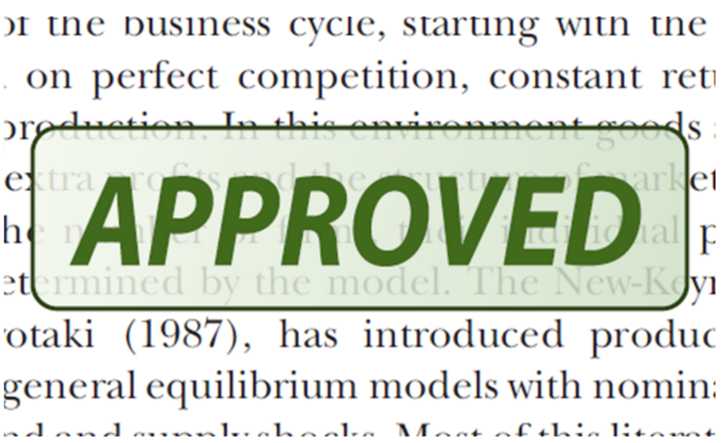


6. Add stamp Tool – for approving a proof if no corrections are required.

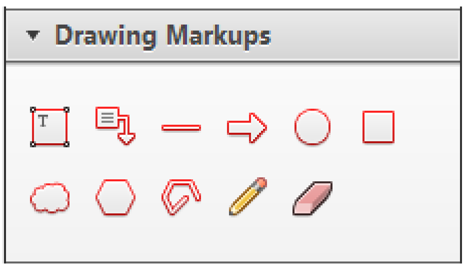
Inserts a selected stamp onto an appropriate place in the proof.

How to use it

- Click on the Add stamp icon in the Annotations section.
- Select the stamp you want to use. (The Approved stamp is usually available directly in the menu that appears).
- Click on the proof where you'd like the stamp to appear. (Where a proof is to be approved as it is, this would normally be on the first page).



or the business cycle, starting with the
on perfect competition, constant ret
production. In this environment goods
extra profits and the structure of market
he number of firms is identical p
etermined by the model. The New-Keyn
otaki (1987), has introduced produc
general equilibrium models with nomin

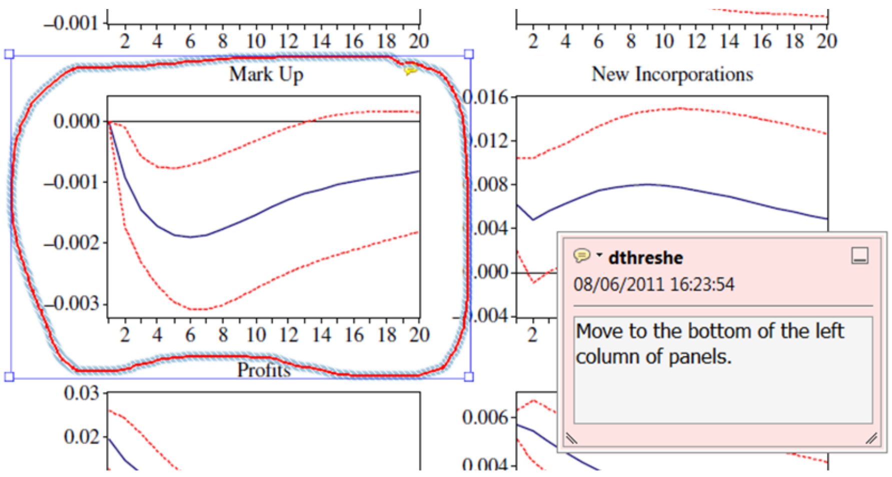


7. Drawing Markups Tools – for drawing shapes, lines and freeform annotations on proofs and commenting on these marks.

Allows shapes, lines and freeform annotations to be drawn on proofs and for comment to be made on these marks..

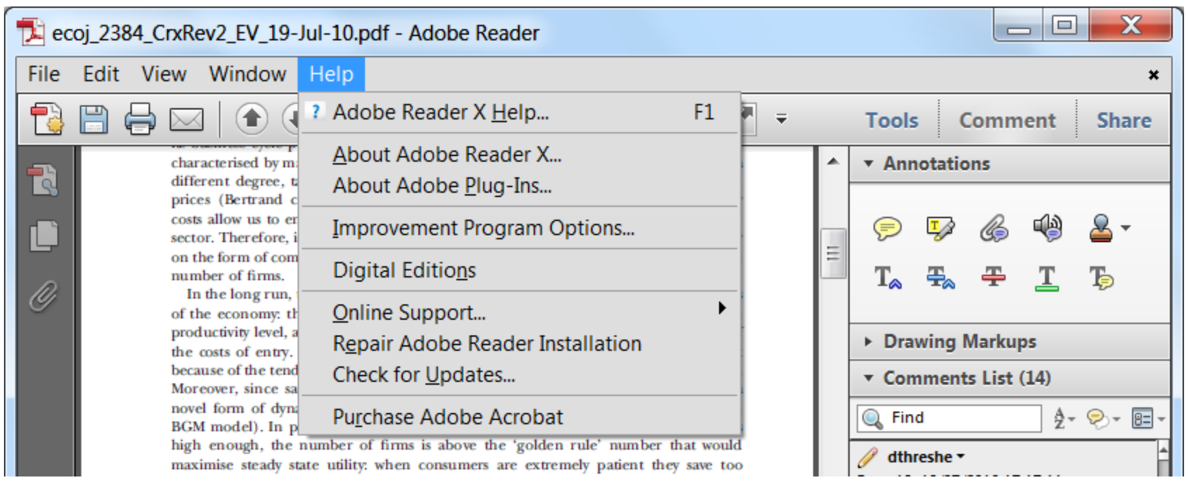
How to use it

- Click on one of the shapes in the Drawing Markups section.
- Click on the proof at the relevant point and draw the selected shape with the cursor.
- To add a comment to the drawn shape, move the cursor over the shape until an arrowhead appears.
- Double click on the shape and type any text in the red box that appears.



dthreshe
08/06/2011 16:23:54
Move to the bottom of the left column of panels.

For further information on how to annotate proofs, click on the Help menu to reveal a list of further options:


ecoj_2384_CrxRev2_EV_19-Jul-10.pdf - Adobe Reader
File Edit View Window Help
Adobe Reader X Help... F1
About Adobe Reader X...
About Adobe Plug-Ins...
Improvement Program Options...
Digital Editions
Online Support...
Repair Adobe Reader Installation
Check for Updates...
Purchase Adobe Acrobat
Tools Comment Share
Annotations
Drawing Markups
Comments List (14)
Find
dthreshe