Methods for individual participant data meta-analysis in prediction research

Valentijn M.T. de Jong

Methods for individual participant data meta-analysis in prediction research

PhD thesis, Utrecht University, the Netherlands

Author: Valentijn M.T. de Jong
Design: Valentijn M.T. de Jong
Cover: Valentijn M.T. de Jong & Gildeprint
Printed by: Gildeprint
ISBN: 978-90-393-7261-6

The research described in this thesis was financially supported by the Netherlands Organization for Health Research and Development, grant number 91810615, and the European Unions Horizon 2020 Research and Innovation Programme under ReCoDID Grant Agreement 825746. Financial support by the Julius Center for Health Sciences and Primary Care for the printing of this thesis is gratefully acknowledged.

©Valentijn M.T. de Jong. All rights reserved. No part of this thesis may be reproduced, stored or transmitted in any form or by any means without the permission of the author.

Methods for individual participant data meta-analysis in prediction research

Methoden voor meta-analyse van data van individuele proefpersonen in voorspellingsonderzoek
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht
op gezag van de rector magnificus, prof. dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties in het openbaar
te verdedigen op

dinsdag 8 december 2020
des middags te 4.15 uur

door

Valentijn Marnix Theodoor de Jong

geboren op 19 juni 1990
te Montfoort

Promotoren: Prof. dr. K.G.M. Moons
Prof. dr. ir. M.J.C. Eijkemans
Copromotor: Dr. T.P.A. Debray

Contents

Chapter 1

General introduction

Prediction research

Prediction models are an important asset in modern medicine. [1, 2] They are commonly developed, validated and used for the prediction of a patient's current (diagnostic prediction models) or future (prognostic prediction models) health status, and may thereby aid in medical decision making and to inform patients on their health. [3, 4, 1, 5] Risk predictions can be used to make decisions regarding the need for additional diagnostic tests, initiating life-style changes or other preventive strategies, identifying the most effective treatment for an individual and for benchmarking the quality of medical centers. Well known examples are QRISK 3, which was developed to facilitate prevention of future heart disease and stroke [6], EuroSCORE II for predicting mortality after cardiac surgery in order to facilitate better decision-making and which may be used as a benchmark in the assessment of the quality of cardiac surgery, [7] and the APRI for predicting the risk of fibrosis and cirrhosis in chronic hepatitis C patients. [8]

As prediction models are developed using data from real persons, they allow for an objective assessment of current or future disease status and quantification of the uncertainty regarding that assessment. This requires a high quality of measuring the predictors in a prediction model. These predictors may include individual participants' characteristics, signs, symptoms, biomarker or imaging test values, genetic test results, biopsy results, etc. A prediction model uses a weighted combination of these predictors to assign a probability to a patient that a health status is present or will be present within a certain time frame.

Prediction models are commonly developed on the data from observational studies or health care records, though data from randomized controlled clinical trials are also sometimes used. Once developed, a prediction model's performance needs to be estimated in other individuals than from which the prediction model was developed. [3, 9] More specifically, it needs to be shown that the model has adequate discrimination and calibration. [10, 9] Discrimination refers to the model's ability to separate individuals with the outcome (health status) from those without that outcome. Calibration refers to the agreement between the outcome probabilities predicted by the model and the true probabilities of the outcome. That is the absolute values of the predicted probabilities need to agree with the observed frequencies of the outcome.

Although these prediction model performance measures can be estimated directly in the development data, this approach yields performance measures that tend to be over-optimistic. The general trend is that model performance decreases when a model is applied to new participants and this decrease can be substantial. [11] A primary cause of this is overfitting, which means that the model's predictor weights (called predictor coefficients) are adapted to idiosyncrasies in the development data at hand rather than the true underlying patterns. An overfitted model yields predicted outcome risks or values for new individuals that are too extreme and are expressed with too much certainty. Besides invalid predictor coefficients, model performance may also be affected by differences in the measurement method of predictor or outcome variables and by differences in patient characteristics (case-mix). [12, 13, 14, 15, 16]

In order to ascertain whether a developed prediction model is sufficiently robust

against aforementioned issues, it has been recommended to assess its performance in data from new individuals not used to develop the prediction model in so-called external validation data sets. [3, 9] These prediction model validation studies are preferably conducted on data from a different population, setting or time period, to assess the model's geographic and temporal accuracy. [17] This may also highlight the need for tailoring the model to these populations and settings, to update the model or re-estimate the prediction model entirely. [18, 19, 20, 21] The tailored or updated model then needs further validation to ensure that predictions are sufficiently accurate in new individuals.

It has become increasingly common that developed prediction models are externally validated before publication. [22, 9] Obviously there is merit to this practice, yet it also has drawbacks. First, separation of development and validation data implies that a smaller than necessary sample is used for both the development and the validation of the prediction model, which reduces the precision of the estimates. Second, when data sets from multiple studies are available for model development, as in an individual participant data meta-analysis (IPD-MA) setting, the choice for which data sets are to be used for development and which for validation can be arbitrary. Third, adequate performance in a single validation study does not necessarily imply that performance will be adequate in practice; for this the validation study needs to reflect the target population and setting. Accordingly, poor performance in a validation study may be a consequence of a sample from a non-target population or setting being used for validation.

Individual participant data meta-analysis

Prediction models should ideally be developed and validated in large samples from multiple populations and settings. [17, 23, 20, 24, 25] This requires research groups to join efforts by sharing their individual participant data and subsequently applying adequate statistical methods to synthesize the data across studies or research centers. To account for heterogeneity between settings and populations (random-effects) meta-analysis can be used, which appropriately weights the evidence from each study. When applied to the combined data of individuals from multiple studies, this is referred to as individual participant data meta-analysis (IPD-MA). In this thesis we address several of the aforementioned issues in prediction research and how these can be resolved in IPD-MA or other large clustered data sets where IPD are available from multiple settings or populations, such as electronic healthcare records. Two prime examples of this are electronic health care records and IPD meta-analysis, where the IPD from multiple centers or studies are combined into a single data set. Notably, the use of these large clustered data sets has several advantages. [26]

1. It improves model development, by enabling the evaluation of the prediction model's heterogeneity across populations and settings during the model development. Commonly, the predictive performance of developed models varies across populations and settings. Having this data already available during model development allows one to adapt the model during model development

so that it performs adequately in each setting. As the variety of participant characteristics (case-mix) is larger in clustered data sets, this also allows one to better account for non-linear effects of patient characteristics.

2. It allows for more informative prediction model validation. As estimates of discrimination and calibration are readily available in each of the samples in the combined data set, the variation in this performance can be explored, its causes can be investigated, the model's generalizability to other populations and settings can be quantified and the need updating the model can immediately be investigated.

3. The observed data can be used more efficiently, as it does not require the splitting of samples into development and validation sets and thereby increases the sample sizes available for both these tasks. This reduces the danger of overfitting in prediction model development, which implies that the resulting models will be more robust. In turn, increased sample sizes for validation imply that the estimates of discrimination and calibration are more precise. It also allows for borrowing information across studies. Apart from borrowing across studies to reduce the variance of estimates, it also allows for the borrowing of information on the quality of measurements. This may enable the estimating of a predictor-outcome association in multiple populations and settings even when a high quality measurement of a certain predictor may be entirely unavailable in some studies.

Outline of this thesis

In **chapter 2** we provide a review of methods for performing an IPD-MA of therapeutic intervention studies where the time to an event is the outcome of primary interest. This is the case when the event is certain to occur but the time until the event is unknown, such as death. Time-to-event analysis, commonly named survival analysis, allows one to estimate the effectiveness of therapeutic interventions and to predict (average) survival times. We provide guidance on the analysis of individual participant data with time-to-event outcomes from multiple therapeutic intervention studies. We illustrate the methods in a real IPD-MA of randomized clinical trials on the effectiveness of Carbamazepine and Valproate to increase the time to epileptic seizures in epilepsy patients.

In **chapter 3** we discuss Stepwise Internal-External Cross-Validation (SIECV) for the development of more generalizable prediction models when multiple individual participant data sets are available. We show how this method can be used to assess and improve the generalizability of prediction models during prediction model development. We illustrate our methodology on two motivating examples: the diagnosis of deep vein thrombosis and the prediction of atrial fibrillation.

In **chapter 4** we develop methods for the standardization of different data sets in an IPD-MA of prediction model validation studies. We show how propensity score methods can be applied to use data from a non-target population or setting in prediction model validation. This effectively increases the sample size available for model validation and thereby improve the reproducibility of performance estimates.

It facilitates the interpretation of (heterogeneity in) prediction model performance in these data sets in terms of the intended population and setting.

It is common in research that a variable of interest is measured with error, that is the preferred measurement method is not available for some participants. For categorical predictors this implies that misclassification may occur, which will result in a biased predictor-outcome relation (or exposure-outcome relation) [27, 28, 29, 30, 31, 32, 33, 34, 35] So far, methods for handling predictor misclassification have been restricted to single studies and aggregate data meta-analysis (that is, based on estimates reported in the literature). In **chapter 5** we discuss methods for restoring the predictor-outcome association in individual participant data meta-analyses where the ideal measurement method of a predictor is unavailable for some of the participants in the IPD-MA or even for all participants in some studies included in the IPD-MA.

In **chapter 6** we provide recommendations for the minimum sample size for the development of prediction models for multinomial outcomes using penalized and unpenalized estimation methods. We base these recommendations on a full factorial simulation study and a motivating example on predicting the correct diagnosis in patients suspected of ovarian cancer.

Finally, in **chapter 7** we provide an overview of meta-analysis methods for prognosis research, when (possibly a combination of) individual participant data as well as aggregate prediction model study data are available from the literature. We finish with providing general recommendations for performing an IPD-MA in prediction modeling research.

Chapter 2

Individual participant data meta-analysis of intervention studies with time-to-event outcomes: A review of the methodology and an applied example

Valentijn M.T. de Jong, Karel G.M. Moons, Richard D. Riley, Catrin Tudur Smith, Anthony G. Marson, Marinus J.C. Eijkemans, Thomas P.A. Debray. Individual participant data meta-analysis of intervention studies with time-to-event outcomes: A review of the methodology and an applied example. *Research Synthesis Methods*, 2020; 1–21. DOI: 10.1002/jrsm.1384

Abstract
Many randomized trials evaluate an intervention effect on time-to-event outcomes. Individual participant data (IPD) from such trials can be obtained and combined in a so-called IPD meta-analysis (IPD-MA), to summarize the overall intervention effect.

We performed a narrative literature review to provide an overview of methods for conducting an IPD-MA of randomized intervention studies with a time-to-event outcome. We focused on identifying good methodological practice for modeling frailty of trial participants across trials, modeling heterogeneity of intervention effects, choosing appropriate association measures, dealing with (trial differences in) censoring and follow-up times, and addressing time-varying intervention effects and effect modification (interactions).

We discuss how to achieve this using parametric and semi-parametric methods, and describe how to implement these in a one-stage or two-stage IPD-MA framework. We recommend exploring heterogeneity of the effect(s) through interaction and non-linear effects. Random effects should be applied to account for residual heterogeneity of the intervention effect. We provide further recommendations, many of which specific to IPD-MA of time-to-event data from randomized trials examining an intervention effect.

We illustrate several key methods in a real IPD-MA, where IPD of 1225 participants from 5 randomized clinical trials were combined to compare the effects of Carbamazepine and Valproate on the incidence of epileptic seizures.

2.1 Introduction

Relative intervention effects (e.g. hazard ratios) are most reliably evaluated in randomized clinical trials (RCT). However, multiple RCTs of the same intervention may provide inconclusive or conflicting evidence on efficacy or safety. Discrepancies between evidence from different RCTs may arise due to chance, or in particular due to heterogeneity in the true intervention effect. This heterogeneity is commonly caused by across-trial differences in, for example, study design (e.g. recruitment strategy, length of follow-up, or analysis methods), case-mix of participants, definition of the studied outcome(s), the implementation (e.g. dosage or intensity) of the intervention. This motivates the need to systematically integrate and summarize evidence across trials, to facilitate evidence-based-medicine.

This can be achieved using a systematic review with meta-analysis (MA). Whereas most meta-analyses are based on aggregated data (AD) from available literature, individual participant (or patient) data meta-analyses (IPD-MA) of multiple intervention studies are considered the gold standard. [36, 37, 38] IPD-MA offers several advantages, as the meta-analyst has full control of the data analysis and uses the data at the individual participant level. [39] Key advantages are the standardisation of outcome and follow-up definitions, checking of data and quality, proper modelling of time-to-event outcomes, and the exploration of intervention-covariate interactions at the participant level. [39, 40] It may thus come to no surprise that IPD-MA are increasingly common. [41, 42]

Extensive guidance has previously been provided for conducting an IPD-MA of intervention effects, for various types of outcome data, such as binary, [43, 42, 44] continuous, [41, 42, 45, 46] ordinal [42] and count outcomes. [42] Yet, IPD-MA are especially useful when analyzing time-to-event outcomes in intervention studies, as censored outcomes can be reassessed for the meta-analysis, survival measures (e.g. hazard ratios, median survival) can be calculated directly and independent to trial reporting, follow-up length can often be increased, time-varying hazard ratios can be examined, and effect modifiers (intervention-covariate interactions) can be assessed. [47, 48]

Whereas a wealth of methods have been developed for analyzing and predicting time-to-event outcomes in single studies, [49, 50, 51, 52] limited guidance exists on their application in IPD-MA settings. In this article, we aim to provide readers with this guidance, by means of our systematic search of databases, narrative review and explanation, and an applied example. Although we focus IPD-MA of trials, the methods we describe are also applicable to multi-center trials.

In the next section, we provide the principles as well as several major issues of time-to-event analyses, that are common in not only IPD-MA but also in single studies. In section 2.3 we provide details of our systematic literature search of methodology for IPD-MA of time-to-event outcomes, and then a narrative review thereof follows in section 2.4 where we discuss the one- and two-stage approaches to meta-analysis, and in section 2.5 where we discuss issues in more detail. Then, in section 2.6 we apply several key methods of the review to a real IPD meta-analysis of clinical trials. Finally, we give provide a discussion in section 2.7 and concluding remarks in section 2.8.

2.2 Principles of time-to-event analysis

The analysis of trials with a survival outcome (e.g. death) typically involves statistical models that account for the time $T_{\text{surv},i}$ elapsed until subject $i, i = 1, .., n$ developed the event of interest. We here denote the probability for subject i to remain event-free for at least t time by the survival function $S(t) = Pr(T_{\text{surv},i} > t)$. A key challenge in time-to-event (TTE) data is that for many participants $T_{\text{surv},i}$ is censored to $T_{\text{cens},i}$, for instance due to dropout or the end of the study. This implies that for those participants $T_{\text{surv},i} > T_{\text{cens},i}$. Hence, the outcome for subject i is typically summarized by the observed event-free or survival time $T_i = \min(T_{\text{surv},i}, T_{\text{cens},i})$ and the event status D_i (where $D = 0$ when censored, and $D = 1$ when the event of interest was observed to have occured). We can compare the survival times of intervention groups and control, while accounting for censoring, with a variety of regression methods.

A commonly used method for analyzing right-censored TTE data is the Cox proportional hazards (PH) model. [53] In this semi-parametric model the effect of the covariates is modeled parametrically, whereas the baseline is left unspecified. It is typically assumed that the ratio of the hazards for any two individuals is constant, irrespective of t. The hazard $h(t|\boldsymbol{X})$ for an individual with covariate vector $\boldsymbol{X'} = (X_1, \ldots, X_k)$ is given by equation 2.1.1 (Table 2.1),
where $\boldsymbol{\beta^T} = (\beta_1, ..., \beta_k)$ is a vector of regression parameters. The function $h_0(t)$ represents the baseline hazard, which is left unspecified. [49, 50] The hazard ratio for two individuals $i = 1, 2$ is then given by $\exp\{\boldsymbol{\beta}'(\boldsymbol{X}_1 - \boldsymbol{X}_2)\}$. For the analysis of randomized trials, $\boldsymbol{X}$ typically just contains a single covariate representing the intervention indicator (e.g. $X_i = 0$ for subjects in the control arm and $X_i = 1$ for subjects in the intervention arm) such that $\exp(\beta)$ can directly be interpreted as the relative intervention effect.

An important consideration is whether to include other (prognostic) covariates in the Cox PH model alongside treatment. In many time-to-event models, including the Cox PH model, the observed unadjusted intervention effect of a protective intervention may change over time due to covariates (i.e. frailty), even if these covariates are perfectly balanced between the intervention groups. [54, 55] Frail participants will have a higher incidence rate than less frail participants. If the intervention is protective, frail participants in the intervention group will have a lower incidence rate than frail participants in a control (or an ineffective intervention) group and participants that are not frail. Over time, the proportion in the control group that is still at risk will increasingly consist of participants that are not frail, whereas this will take longer for the intervention group, thereby resulting in an imbalance in frailty. For trials with a high event rate and most frailty distributions, the unadjusted intervention effect will attenuate towards the null (hazard ratio of 1) as time progresses, which violates the proportional hazards assumption. [56] The unadjusted intervention effect is then the marginal intervention effect, [57] i.e. the average intervention effect for the population as a whole, averaged across all time-points. Hence, it is dependent on the length of the follow-up.

Table 2.1: Models for two-stage time-to-event meta-analysis

Type	Model	Hazard function	Survival function	Ref.	No.
Proportional Hazards	General model[1]	$h_0(t)\exp(\boldsymbol{\beta'X})$	$S(t\|\boldsymbol{X}) = S_0(t)^{\exp(\boldsymbol{\beta'X})}$	[49, 47, 58]	2.1.1
	Exponential	$\lambda\exp(\boldsymbol{\beta'X})$	$S(t\|\boldsymbol{X}) = \exp(-\lambda t\exp(\boldsymbol{\beta'X}))$	[49, 51, 58]	2.1.2
	Weibull[2]	$\lambda\nu t^{\nu-1}\exp(\boldsymbol{\beta'X})$	$S(t\|\boldsymbol{X}) = \exp(-\lambda t^{\nu}\exp(\boldsymbol{\beta'X}))$	[49, 59, 51, 58]	2.1.3
	Gompertz[3]	$\lambda\exp(\psi t)\exp(\boldsymbol{\beta'X})$	$S(t\|\boldsymbol{X}) = \exp(-\frac{\lambda}{\psi}(\exp(\psi t)-1)\ \exp(\boldsymbol{\beta'X}))$	[49, 60, 51]	2.1.4
Accelerated Failure Time	General model	$h_0(t\exp\{\boldsymbol{\beta'X}\})\exp(\boldsymbol{\beta'X})$	$S(t\|\boldsymbol{X}) = S_0(t\exp(\boldsymbol{\beta'X}))$	[49, 51, 61]	2.1.5
	Weibull	$\lambda\nu t^{\nu-1}(\exp(\boldsymbol{\beta'X}))^{\nu}$	$S(t\|\boldsymbol{X}) = \exp(-\lambda t^{\nu}\exp(\nu\boldsymbol{\beta'X}))$	[51, 49]	2.1.6
	Log-logistic[4]	$\frac{\varphi}{t\{1+t^{-\varphi}exp(-\boldsymbol{\beta'X})\}}$	$\log\frac{1-S(t\|\boldsymbol{X})}{S(t\|\boldsymbol{X})} = \varphi log(t) + \boldsymbol{\beta'X}$	[61, 62, 63]	2.1.7

[1] In the Cox Proportional Hazards model, the baseline hazard $h_0(t)$ is left unspecified.
[2] ν is a shape parameter, λ is a scale parameter.
[3] The Gompertz distribution can be generalized to the Gompertz-Makeham distribution by adding a constant to the hazard function. [64]
[4] The log-logistic model is a proportional odds model, where the β parameters can be interpreted as log-odds ratios.

If the intention is to measure a conditional intervention effect, i.e. the intervention effect for a participant with given covariate values, the observed unadjusted intervention effect is often not valid. Instead, covariates should be included in the model, to obtain a conditional intervention effect. [65, 66] Further, the adjustment for a prognostic covariate often increases the power for finding an intervention effect. [67] Alternatively, an AFT model could be used (sections 2.5.1 and 2.5.2), for which the effect of missing covariates is absorbed into the baseline parameters, leaving the unadjusted intervention effect unaffected. [56]

The Cox PH model has numerous appealing properties, in particular allowing the estimation of hazard ratios for included covariates without requiring the shape of the baseline hazard to be specified. However, its implementation is not always justified. For instance, difficulties may arise when hazards are non-proportional. Although effects to model non-PH can be included (e.g. with splines, interactions or time-varying effects) in a Cox PH model, this usually complicates the interpretation of the estimated intervention effect.

For these reasons it is often recommended to adopt a model where proportionality occurs on another scale when proportionality of hazards is violated, which is discussed in section 2.5.2. When absolute survival probabilities for individual participants are of primary interest, it can be useful to define a parametric function for $h_0(t)$, and thus to abandon Cox PH models altogether, [68, 69] which is discussed in section 2.5.1. Indeed, even when the focus is mainly on an intervention effect, translation of its hazard ratio to the absolute risk scale is important, which requires the baseline survival to be modelled, either parametrically or non-parametrically. For a full overview of R packages on time-to-event analysis, see `cran.r-project.org/web/views/Survival.html`.

2.3 IPD meta-analysis methods: review

Increasingly often, IPD from multiple studies are available for analysis. This introduces new challenges and allows for different approaches for analysis, which we set out to identify. We conducted a literature review to identify scientific articles concerning statistical methods for IPD-MA of time-to-event data.

2.3.1 Methods

We systematically searched through Pubmed and Web of Science using the search filters supplied in Supporting Information 1 (https://doi.org/10.1002/jrsm.1384), from conception until December 31st, 2018. In addition, we added suggestions and performed cross-reference checks of the obtained articles. Articles were considered eligible for inclusion if they described statistical methods for analyzing multiple or clustered individual participant data sets with a time-to-event outcome. Publications that met at least one of the following criteria were excluded from our review:

- Full text of the manuscript not available,
- Not published in English,
- Not a peer reviewed article,

- Application of methods without methodological focus,
- No focus on at least one of the following topics:
  - time-to-event outcomes,
  - IPD,
  - estimation of intervention effects,
  - meta-analysis or analysis of clustered data.

2.3.2 Results

A total of 1887 unique records were identified through our search strategy, and were deemed eligible for title and abstract screening (Figure 2.1). Of these, 1713 were removed during screening because the titles did not have a methodological focus. The remaining 174 records were assessed on the full-text, of which 58 met the inclusion criteria and 116 did not. Further, a total of 159 unique records were assessed after being suggested or found through cross-referencing. Of these, 16 suggestions and 54 cross-references met the inclusion criteria and were included in the review. A total of 128 articles were included in the review, of which a complete list can be found in Supporting Information 3 (https://doi.org/10.1002/jrsm.1384).

The core methods for analyzing TTE outcomes in IPD-MA are described in section 2.4. The structure of this section was defined independent of the review, yet the description of methods therein has resulted from the review. Further, extensions to these methods, such as relaxing the proportionality of hazards assumption, modeling multiple interventions or outcomes, and methods for missing data are described in section 2.5.1, which was grouped according to the topics identified in the review. The review has resulted in ten key recommendations backed by references, which are summarized in Table 2.2.

Figure 2.1: Flowchart of inclusion and exclusion of papers for review.

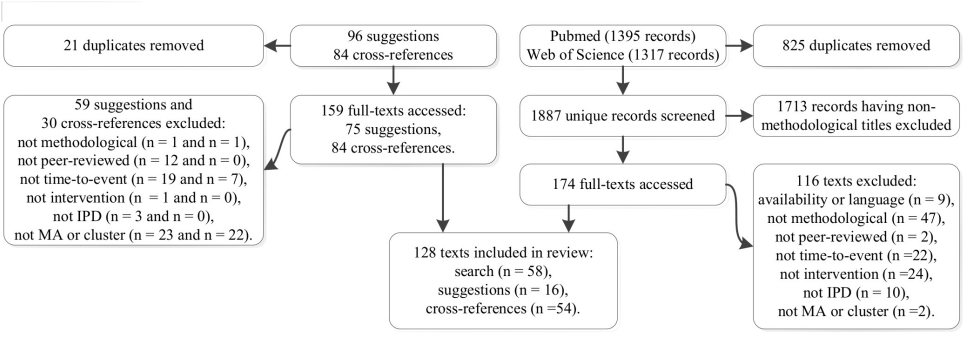


Table 2.2: Ten Recommendations for the IPD-MA of TTE data from Randomized Trials Examining an Intervention Effect

Recommendation	Reference
The Cox model may be the default model of choice, but proportionality of hazards should be tested, e.g. with interaction or time-varying effects for the intervention.	[70, 71]
Consider non-PH models.	[72, 73, 74, 61, 75]
Account for clustering in one-stage models, preferably by stratification of the baseline.	[54, 76, 77, 78, 79, 80]
Adjust for covariates measured before randomization.	[59, 81, 67]
Apply one-stage models if trials are very small or the outcome very rare.	[51, 82]
In one-stage models, center covariates within trials.	[83]
Model participant-level interactions on the participant-level.	[84]
For the intervention effect (& its interaction effects), apply random effects & investigate heterogeneity.	[47, 40, 85, 86]
If competing risks are present & absolute risks are of interest, apply competing risks models.	[87, 88, 89, 90, 91]
Multiple imputation of missing covariates must account for clustering & time-to-event, using the event indicator and the Nelson-Aalen cumulative hazard.	[92, 93, 94, 95, 96]

2.4 Description of methods

2.4.1 Time-to-event analysis in individual participant data meta-analysis

When IPD from multiple trials are available, summary estimates for relative intervention effects can be obtained using the so-called one-stage or two-stage approaches. [82] In the conceptually simpler two-stage approach (section 2.4.2), the IPD from each trial is analyzed separately to produce trial-specific estimates of relative intervention effect (e.g. hazard ratios), using the same methodology in each trial (e.g. Cox regression). In the second stage, estimates of intervention effect are combined into a weighted average using traditional meta-analysis methods that ideally account for possible between-trial heterogeneity. In the one-stage approach (section 2.4.3), data from all studies are analyzed in one analysis, and a variety of methods can be used to account for clustering of participants within studies. [51, 80, 49, 97, 42] In both the one- and two-stage approaches, methods to account for heterogeneity in intervention effects across studies are available (Table 2.3). [36, 97, 42] In the one-stage approach, one must also decide how to model or account for heterogeneity in other parameters (such as adjustment factors or terms defining the baseline hazard). For a discussion on the choice between the one-stage and two-stage approaches see section 2.8.

Table 2.3: Methods for Modeling Heterogeneity

Baseline	Coefficients	Modeled difference between trials
Common	Common	No difference, same for every trial
Frailty	Random Effects	Proportional differences, difference between trials follows distribution
Fixed [a]	Fixed [b]	Proportional differences, estimated per trial. Same shape between trials.
Stratified		Non-proportional differences. Estimated per trial, with different shapes.

These methods are possible in one-stage meta-analysis. In a two-stage meta-analysis the baseline is stratified and the given options for the coefficients can be used.
[a] By adding trial indicators to the model.
[b] By adding trial indicators $*$ variable interaction to the model.

2.4.2 Two-stage approach

The two-stage approach is often considered the most convenient approach for IPD meta-analysis, as it does not necessarily require IPD to be exchanged. For instance, each trial can be analyzed separately, and only their summary statistics are combined. The approach is particularly appealing when not all trials provide IPD, as it allows reported intervention effects and their respective standard errors from non-

IPD trials to be analyzed in the second stage, together with the estimates from the IPD trials.

In the first stage, common methods for TTE analysis can be used to obtain estimates of relative intervention effect for each trial (so-called aggregate data). For instance, when applying Cox regression (equation 2.1.1), this yields the log hazard ratio estimates $\hat{\beta}_j$ and their corresponding error variance $V(\hat{\beta}_j)$, for trial $j = 1, ..., J$. Afterwards, the estimated intervention effects can be summarized by calculating a weighted average. For instance, in a so-called common (or fixed) effect meta-analysis it is assumed that all trials share a common intervention effect β_{IV}, which can be derived as follows:

$$\begin{aligned} \beta_{\text{IV}} &= \frac{\sum_{j=1}^{J} \frac{\hat{\beta}_j}{V(\hat{\beta}_j)}}{\sum_{j=1}^{J} \frac{1}{V\hat{\beta}_j)}} \\ V(\beta_{\text{IV}}) &= \frac{1}{\sum_{j=1}^{J} \frac{1}{V(\hat{\beta}_j)}} \end{aligned} \tag{2.1}$$

where V is the variance. Hereby, it is assumed that the within-trial variances $V(\hat{\beta}_j)$ are known (i.e. estimated without uncertainty). The common effect meta-analysis model can also be formulated as follows:

$$\hat{\beta}_j \sim \mathcal{N}\left(\beta_{\text{IV}}, V(\hat{\beta}_j)\right) \tag{2.2}$$

If certain trials provide no IPD, but the intervention effect and its variance are available in the literature, these can be included in the second stage of the two-stage framework, [98] provided that the models in the first stage are specified the same. If a trial has a small sample size, the Maximum Likelihood estimator of the intervention effect can be affected by small sample bias. [99] Worse still, if considerable censoring is present, the likelihood may be monotone and the Maximum Likelihood may be inestimable, depending on the intervention and covariate distributions. [100] This can be resolved by applying Firth's correction to the likelihood in the first stage, [99, 100] or by opting for a one-stage model instead.

The assumption that an intervention effect is common across trials is often unrealistic, as trials are often affected by between-trial heterogeneity. [101, 102] This heterogeneity may, for instance, appear when participant-level covariates interact with the intervention effect (i.e. effect modification), when small sample bias is present in some estimates of the intervention effect, or when aggregate data are based on invalid modeling assumptions (e.g. in the presence of non-proportional hazards, non-PH). For time-to-event analysis, between-trial heterogeneity may also arise due to selection effects. In particular, participants who are more frail and therefore more susceptible to the outcome, are no longer at risk after having an event. Therefore, over time, the most frail participants are removed from the risk set, whereas the less frail participants remain at risk (see section 2.2). [54, 65, 103, 49] This, in turn, may lead to different intervention effects across trials if the follow-up length differs across trials. For these reasons, in the two-stage approach

it is generally recommended to adopt a random effects meta-analysis model, which is typically specified as:

$$\begin{aligned} \hat{\beta}_j &\sim \mathcal{N}\left(\beta_j, V(\hat{\beta}_j)\right) \\ \beta_j &\sim \mathcal{N}\left(\beta_{\mathrm{RE}}, \tau^2\right) \end{aligned} \tag{2.3}$$

In contrast to common effect models, random effects models allow for differences in $\hat{\beta}_j$ due to sampling error *within* studies (reflected by $V(\hat{\beta}_j)$) and due to heterogeneity in the true intervention effects β_j *across* studies (reflected by τ^2). Estimates for β_{RE} can thus be interpreted as the average intervention effect across studies. A confidence interval for $\hat{\beta}_{\mathrm{RE}}$ is traditionally constructed as $\hat{\beta}_{\mathrm{RE}} \pm z_{1-\alpha/2} \sqrt{V(\hat{\beta}_{\mathrm{RE}})}$, where $z_{1-\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution. [104] To account for the uncertainty in τ^2 and thereby improve the coverage of the interval, the Hartung-Knapp approach to confidence intervals is given by $\hat{\beta}_{\mathrm{RE}} \pm t_{J-1,1-\alpha/2} \sqrt{V_{\mathrm{HK}}(\hat{\beta}_{\mathrm{RE}})}$, where $t_{J-1,1-\alpha/2}$ is the upper $\alpha/2$ quantile of a t-distribution with $J-1$ degrees of freedom, and $V_{\mathrm{HK}}(\hat{\beta}_{\mathrm{RE}})$ is a modified variance estimate. [105, 106, 107, 108, 109]

Heterogeneity of the intervention effect in the two-stage approach

Statistical heterogeneity in the intervention effect can be recognized by $\hat{\tau} > 0$. The influence of heterogeneity on intervention effects may be explored by constructing a prediction interval, which estimates the interval of the likely intervention effect in a (new) individual trial and can be calculated approximately as follows[110, 85]:

$$\hat{\beta}_{\mathrm{RE}} \pm t_{J-k,1-\alpha/2} \sqrt{\hat{\tau}^2 + V(\hat{\beta}_{\mathrm{RE}})}, \tag{2.4}$$

where $\hat{\beta}_{\mathrm{RE}}$ is an estimate of β_{RE} and $V(\hat{\beta}_{\mathrm{RE}})$ its variance. Typically the $t_{J-2,1-\alpha/2}$ quantile is used here, although similar to the confidence interval there is no consensus on the distribution and its degrees of freedom. [110, 85] When random effects models indicate the presence of important statistical heterogeneity (i.e. $\hat{\tau} > 0$, or a wide prediction interval) of the intervention effect, the interpretation of the overall summary estimate, $\hat{\beta}_{\mathrm{RE}}$, may become difficult or meaningless. Therefore, it is often helpful to identify sources of heterogeneity in intervention effect (see Table 2.4). [47] This can, for instance, be achieved by assessing the relation between relevant trial-level covariates (e.g. level of blinding, or dosage) and the trial effect estimates, also known as meta-regression. [84]

Table 2.4: Potential sources of Heterogeneity in Time-to-event Meta-Analysis

Source	Solutions	Reference
Non PH + Differences in follow-up time	Interaction terms	[84, 111, 83]
	Model effect(s) as time-varying, use splines	[112, 113]
	Use a different model (e.g. AFT)	[72, 69, 73, 74, 114, 113]
Difference in case-mix	Include covariates / prognostic factors	[103, 59]
	AFT model	[59, 73, 74]
Selective dropout or competing risk	Model dropout or competing risk	[87, 89, 115, 91]
Small sample bias in some studies	Bias correction	[99]
	One-stage MA	[116, 115, 42, 82]
	Arcsine transform (for two-stage MA)	[115]

PH: Proportional Hazards; AFT: Accelerated Failure Time; MA: Meta-Analysis. Heterogeneity can be diagnosed by applying frailty and/or random effects terms.[80, 86, 48] If heterogeneity remains, e.g. due to differences in study protocols, stratification of baseline hazard/frailty and/or random effects terms must be applied.[103, 66]

When patient-level associations with treatment effect are of interest, it is better to model interactions between participant-level characteristics (e.g. participant age) on the participant level. In the two-stage approach, the statistical interaction between the relevant covariate and intervention are first estimated separately in each trial, and then the resulting coefficients are meta-analyzed using traditional meta-analysis models. [101, 71] When the intervention effect changes over time, differences in follow-up time between trials will lead to heterogeneous estimates of intervention effect across trials, if unaccounted for. This heterogeneity of intervention effects can be quantified with random-effects meta-analysis, but would preferably be modeled directly (section 2.5.2).

Estimation

A commonly used approach to estimate the heterogeneity from the random effects model (equation 2.3), is to use the method of moments by DerSimonian and Laird (DL). [117] This estimator is biased downwards when the true heterogeneity is moderate or high and sample sizes are low, as the variance estimates are assumed to be known and fixed, [118] leading many researchers to suggest alternatives, the most important of which are mentioned here. The two-step Paule-Mandel method is similar to DL, but iteratively estimates the study weights, and has reduced bias for high values of τ. Another alternative is the Maximum Likelihood (ML) estimator. Although the MSE of the ML estimator for τ is small, it is very biased when τ is large and the included studies are small. [119] The Restricted Maximum Likelihood (REML) estimator yields less biased estimates of τ and has relatively low MSE. [120, 121, 122] Therefore, REML and the two-step Paule-Mandel method are the recommended estimators for τ. [118, 122]

As there may be considerable uncertainty in the heterogeneity estimate regardless of which estimator is used, [122] it is recommended to report a confidence interval for the heterogeneity as well. [123] This may be estimated with the Q-profile method [108, 124] or the generalised Cochran between-study variance method. [119] Further, it should be noted that when fewer than 10 trials are included in the meta-analysis, or when trials are small or the outcome rare, no currently available method can reliably estimate the heterogeneity. [122]

Even though estimates for heterogeneity in meta-analysis tend to be biased in many situations, this barely biases the summary effect estimate, unless there are very few events. [122] The confidence intervals of the summary effect can be constructed by applying the Hartung-Knapp-Sidik-Jonkman HKSJ method for confidence intervals, [125, 126] which had good coverage in simulations for a minimum of two studies, unless the number of events was very low. [127, 122] This may be corrected by applying a modification that ensures that the confidence intervals are at least as wide as a fixed-effects meta-analysis confidence interval. [122] Hence, it is currently recommended to apply a random effects model estimated with REML or two-step Paule-Mandel, and to use the HKSJ method for confidence intervals. [122] Alternatively, Bayesian random-effects models may be used. However, in the simulation studies discussed here either aggregate data or non time-to-event IPD were generated, which is a concern considering that it has been suggested that the performance of the estimators may be related to the type of outcome. [119] For a

comprehensive overview of meta-analysis estimators see [128, 119, 129], for a comparison of their performance see [118, 122], for an overview of software see [119] as well as the two recent packages admetan and ipdmetan, [130] and for an up-to-date overview of R packages see `cran.r-project.org/web/views/MetaAnalysis.html`.

2.4.3 One-stage approach

Accounting for clustering

When applying the one-stage approach, within-trial and between-trial relationships are estimated simultaneously, which can give a more complete understanding of the data. [48] As is the case for two-stage meta-analysis, a one-stage meta-analysis must account for clustering (Table 2.3). [80, 97] Participants in different studies may differ on unmeasured covariates, which will lead to a biased estimate of the conditional (i.e. for a participant with given covariate values) intervention effect regardless of balance of these covariates between intervention groups, if not adjusted for (section 2.2). [55] Whereas the two-stage approach naturally deals with this by estimating separate baseline hazards for the different studies, in the one-stage approach we can use stratification (section 2.4.3), frailty models (section 2.4.3) or marginal models (section 2.4.3).

Stratified models

A commonly used approach is to apply a Cox model with stratified baseline hazards but a common intervention effect (equation 2.5.1, Table 2.5). [47, 131, 51, 132] This allows the shapes of the baseline hazards to vary between trials, whereas the hazards of the different intervention groups are assumed to be proportional within trials, and gives a single estimate of overall intervention effect. When the sample sizes per trial are very small and many trials are included, the stratification of baselines is less efficient than the use of frailty terms, [51] though it also requires fewer assumptions as it fully accounts for any differences in baselines between trials. For the meta-analysis of trials that are each powered to detect a clinically significant intervention effect this should not be an issue, thereby making the stratification of the baseline the preferred model specification.

Frailty models

Rather than stratifying the baseline hazard across the trials, it is possible to model their distribution through frailty terms. A frailty term is a random parameter (i.e. random intercept) within the baseline hazard function that is assumed to follow a specified distribution and thereby allows for differences in baseline rate between (groups of) participants that are a result of unmeasured covariates. Shared frailty models (equation 2.5.2, table 2.5) are designed to account for these differences in unmeasured covariates between trials. Therefore, the assumption in a frailty model is that the baseline hazards in each study have the same shape but a different magnitude. The estimated intervention effect is then to be interpreted relative to other participants in the same trial with the same frailty and covariates. If the baseline hazard of this model is left unspecified, this leads to the Cox PH model

Table 2.5: Models for one-stage time-to-event meta-analysis

Type	Model	Hazard function	Survival function	Ref.	No.
Proportional Hazards	Stratified baseline	$h_{0j}(t)\exp(\boldsymbol{\beta'X}_j)$	$S_j(t\|\boldsymbol{X}_j) = S_{0j}(t)^{exp(\boldsymbol{\beta'X}_j)}$	[53, 133, 49, 47, 51]	2.5.1
	Shared frailty	$h_0(t)\eta_j\exp(\boldsymbol{\beta'X}_j)$ where $\eta_j \sim$ Gamma(θ) or $log(\eta_j) \sim$ Normal$(0,\tau^2)$	$S_j(t\|\boldsymbol{X}_j) = S_0(t)^{\eta_j \exp(\boldsymbol{\beta'X}_j)}$	[54, 133, 51, 49, 47]	2.5.2
	Random effects	$h_0(t)\exp(\boldsymbol{\beta'X}_j + \boldsymbol{b}'_j\boldsymbol{Z}_j)$ where $\boldsymbol{b}_j \sim$ MVN$(\boldsymbol{0},\boldsymbol{\Sigma})$	$S_j(t\|\boldsymbol{X}_j) = S_0(t)^{\exp(\boldsymbol{\beta'X}_j+\boldsymbol{b}'_j\boldsymbol{Z}_j)}$	[134, 135, 133, 48, 80, 113, 136]	2.5.3
Accelerated Failure Time	Stratified baseline	$h_{0j}(t\exp\{\boldsymbol{\beta'X}\})\exp(\boldsymbol{\beta'X})$	$S_j(t\|\boldsymbol{X}_j) = S_{0j}(t\exp(\boldsymbol{\beta'X}))$	[51]	2.5.4
	Shared frailty	$h_0(t\ \eta_j\exp\{\boldsymbol{\beta'X}\})\ \eta_j\exp(\boldsymbol{\beta'X})$ where $\eta_j \sim$ Gamma(θ) or $log(\eta_j) \sim$ Normal$(0,\tau^2)$	$S_j(t\|\boldsymbol{X}_j) = S_0(t\ \eta_j\exp(\boldsymbol{\beta'X}))$	[137, 74, 51, 113]	2.5.5
	Random effects	$h_0(t\exp\{\boldsymbol{\beta'X}_j+\boldsymbol{b}'_j\boldsymbol{Z}_j\})\exp(\boldsymbol{\beta'X}_j+\boldsymbol{b}'_j\boldsymbol{Z}_j)$ where $\boldsymbol{b}_j \sim$ MVN$(\boldsymbol{0},\boldsymbol{\Sigma})$	$S_j(t\|\boldsymbol{X}_j,\boldsymbol{Z}_j) = S_0(t\exp(\boldsymbol{\beta'X}_j+\boldsymbol{b}'_j\boldsymbol{Z}_j))$	[137, 74, 51, 113]	2.5.6

In the Cox Proportional Hazards model, the baseline hazard $h_0(t)$ is left unspecified. For the baseline hazard of the parametric models, see Table 2.1.

with random trial intercept. [47, 49] When data from multiple multi-center studies are combined, nested frailty models can be applied. [138]

It is common to assume a gamma distribution for the frailty, for mathematical or computational reasons, [66, 49] or a normal distribution for the log-frailty, as this bears similarity to the generalized linear mixed effects model, [49, 139] though many other distributions including the inverse Gaussian, positive stable, and compound Poisson are possible. [51, 49, 52] Previous studies have demonstrated that the gamma frailty model appears to be fairly robust against misspecification of the frailty distribution, [78, 140] that it describes the frailty of survivors for a large class of hazard models, [66] and that it can have more power than a stratified model. [77, 51, 140] Therefore, frailty models are generally recommended when the number of participants per trial is very low. Yet, when the number of participants per trial is large, as is often the case in meta-analysis when individual trials are designed to have sufficient power to test for an intervention effect, the frailty and stratification approaches will usually yield similar results, given that the assumptions are met.

When a frailty is applied to the baseline hazard, the median hazard ratio (MHR) can be used to evaluate the meaning of this frailty in the context of the different studies. [141, 142, 143] The MHR is the median relative difference in the hazard of the occurrence of the outcome when comparing identical participants from two randomly selected studies ordered by hazard. When a log-normal distribution is assumed for the frailty, the Median Hazard Ratio (MHR) can be computed as $\exp\{\sqrt{2\sigma^2}\Phi^{-1}(0.75)\}$, where Φ^{-1} is the inverse of the standard normal distribution. [142, 143]

Marginal models

In the analysis of clustered data, such as IPD from different studies, where the interest lies in the average intervention effect for the target population as a whole, we may use marginal models. In such models the dependence between participants from the same trial is not modeled explicitly but standard errors are adjusted for it. [144, 145] Intervention effects are interpreted as relative to participants drawn randomly from the entire target population from which the participants are considered to be sampled. [76] When the interest lies in the intervention effect of participants in the individual studies or in the causes of heterogeneity of intervention effects across studies or subgroups, as in an IPD-MA often is the case, conditional models are needed. [79]

Estimation

Maximum Likelihood (ML) estimates of the mixed effects Cox model may be obtained with a Newton-Raphson procedure, [146] with penalization methods by constraining the frailty terms with a penalty, [147, 148, 86] by expectation-maximisation, [135] or by expectation-maximisation and penalization. [136]

Further, residual maximum likelihood (REML) estimates of the mixed effects Cox model can be obtained with a Newton-Raphson procedure, [146, 47, 48] or with penalization methods by constraining the frailty terms with a penalty. [147, 148] As the penalized method does not take uncertainty of τ^2 into account, it has been

suggested that it produces less precise estimates of the intervention effect. [132] However, comparative evidence is currently lacking.

Alternatively, the mixed effects Cox model can be estimated with a poisson model, [137] where the time-scale is split into intervals defined by event times. [149] Mixed effects parametric models can be estimated with Maximum Likelihood by adaptive Gauss-Hermite quadrature. [113] Mixed effects Weibull models can also be estimated with REML. [137]

The Bayesian framework allows for the estimation of a wide range of time-to-event models. For instance, the Cox random effects model can be estimated using Bayesian methods. [150, 134, 51] A random trial effect and an intervention by trial interaction may be evaluated simultaneously in a Bayesian Cox PH model. [151] For a discussion of commensurate priors for incorporating between-trial variability in a Bayesian meta-analysis, see [152]. Finally, an overview of software for the estimation of one-stage time-to-event models is given in Table 2.6.

Heterogeneity of the intervention effect in the one-stage approach

Similar to the two-stage approach, we may expect heterogeneity of the intervention effect in the one-stage approach, which makes the common effects assumption untenable. As such, it is also recommended for one-stage models to assume random effects (equation 2.5.3, Table 2.5), [135] and to investigate the causes of this heterogeneity, if present. [47] One possible cause of heterogeneity of the intervention effect is effect modification (i.e. interaction) at the individual level, which can be investigated by adding an interaction term in the one-stage model. [80] Crucially, when including such an interaction term (e.g. an intervention-covariate interaction) in the one-stage approach, special care must be taken to avoid the amalgamation of within- and across-trial information, as this may lead to ecological bias. This can be achieved by centering the covariates by their mean values within trials, such that the interaction estimate is then only based on within-trial information. [83] To improve the estimation of between-study variance and the coverage of confidence intervals, the intervention variable can be centered within studies as well. To further prevent the borrowing of information across studies that may affect the estimate of the intervention effect in the one-stage approach, a covariate by trial indicator interaction can be included. This stratifies the covariates effects as it allows covariate effects to be estimated separately for each study (see Table 2.3).

When there are differences in follow-up time between trials and the intervention effect changes over time, the estimated intervention effects (as quantified by random effects) will be different per trial. If this is unaccounted for, this will lead to heterogeneity of the intervention effect. This can then be investigated by modeling the effect as time-dependent (section 2.5.2).

In the two-stage approach the influence of trial-level characteristics on the intervention effect can be estimated with meta-regression in the second stage. In the one-stage approach it is possible to simultaneously estimate the heterogeneity of baseline rate of the participants within different studies, the heterogeneity of intervention effects and their correlation. [86]

Table 2.6: Software for One-stage Time-to-event Models

Program	Package/method	Description	Code in	Mentioned in
R, S-Plus	-	Random effects Cox model	[136]	
	survival	Cox and parametric time-to-event models.	[148, 153, 140]	
		Stratified, frailty and marginal specifications	[154]	
	coxme	Mixed effects Cox models		
	frailtypack	Cox and parametric random effects and stratified models. Correlated random effects. Competing events. Joint nested frailty models.		[138, 86, 91]
	SemiCompRisks	Bayesian and frequentist random effects parametric and semi-parametric models for competing events.		[91]
	parfm	Parametric frailty models		
	PenCoxFrail	Regularized Cox frailty models		
	mexhaz	Flexible (excess) hazard regression models, non-proportional effects, and random effects		
	dynfrail	Semiparametric dynamic frailty models		
	frailtyEM	Frailty models with semi-parametric baseline hazard, recurrent events		
	joineR	Joint random effects models of repeated measurements & time-to-event		
	joint.Cox	Joint frailty-copula models with smoothing splines		
	JointModel	Joint model for longitudinal and time-to-event outcomes		
	joineRML	Joint time-to-event and multiple continuous longitudinal outcomes		
	rstanarm	Joint model for hierarchical longitudinal and time-to-event data	[155]	
	surrosurv	Time-to-event surrogate endpoints models	[156]	
SAS	PHREG	Cox models, including stratification or frailty	[153, 140]	[157]
	NLMIXED	Mixed effects parametric survival models		[158]
		Joint model for recurrent events and semi-competing risk	[159]	
	GENMOD	Poisson regression, marginal models		[157]
Stata	stcox	Cox model, stratified and frailty specifications.		
	stmixed	Flexible parametric time-to-event models with mixed effects		[113, 42]
	xtmepoisson	Mixed effects Poisson regression	[149]	
JAGS, OpenBUGS,	-	Bayesian mixed effects models,	[77, 149, 153]	
WinBUGS	-	IPD network meta-analysis	[160, 75]	
MLwiN	-	Mixed effects time-to-event models		[161, 42]
The Survival Kit	-	Bayesian mixed effect time-to-event models		[151]

2.5 Extensions

2.5.1 Modeling the baseline hazard function

Whereas the Cox PH model leaves the baseline hazard unspecified, we may apply a parametric model by specifying a baseline hazard (Table 2.1), either in the first stage of the two-stage approach, or within the one-stage approach. To allow for flexible shapes of the baseline hazard, we can apply spline functions. Particularly the approach of Royston and Parmar is useful, where the baseline cumulative hazard is modelled using restricted cubic splines, [61] and which has been extended to allow for random effects. [113]

Parametric models are especially suitable when absolute (rather than relative) risks for individual subjects (rather than for subpopulations) are of primary interest. It leads to smooth predicted survival curves and is well suited to deal with non-proportionality of hazards. For instance, researchers increasingly often aim to develop prediction models that can assess individual intervention benefits (or harms). [162] Most simply, one can specify an exponential (eq. 2.1.2) or a Weibull (eq. 2.1.3) distribution within the proportional hazards framework. The exponential distribution assumes a constant rate over time, whereas the Weibull distribution (a generalization of the exponential distribution) allows for accelerated failure times (AFT). [49] Other (but less common) generalizations of the exponential distribution that can be used for modeling the baseline hazard are the Gompertz, gamma, and piecewise constant distributions. [49, 113] Further, the log-logistic, log-normal and generalized gamma distributions may be used. [61, 113] Unlike PH models, the estimate of an intervention effect in AFT models is unaffected by unmeasured prognostic covariates. [56] Also in one-stage models a wide range of distributions for parametric PH and AFT models is available. [113]

2.5.2 Modeling non-proportional hazards

For short trials with a low event rate the proportionality of hazards across time may be reasonable (i.e. the hazard ratio for the intervention effect may be assumed constant over time), but as the number of events in different intervention groups diverges a selection of participants remains in the trial for whom proportionality in the unadjusted intervention effect is not realistic. [112, 73] If an intervention is protective, frail participants in the intervention group will be better protected against the outcome than frail participants in the control group. Hence, the proportion of frail participants at risk will decrease more quickly in the control group than in the intervention group. To account for this issue within studies we can include covariates in the model, whereas we can use a frailty model to account for this issue between studies.

Non-proportionality of hazards may also be present due to the intervention effect truly being dependent on time. For instance, an intervention (such as surgery or chemo-therapy) may cause an increased risk of a negative outcome at first, but have a protective effect in the long run. This can be modeled by an interaction effect between the intervention (or a covariate) and time [53] in the one-stage approach or in the first stage of the two-stage approach. To allow for flexible shapes of this time-

Table 2.7: Effect Measures for Time-to-Event Analysis

Measure	Definition	Ref.	No.
Hazard ratio	$\frac{\lambda(t\|\boldsymbol{X}_1)}{\lambda(t\|\boldsymbol{X}_0)}$, $\lambda(t\|\boldsymbol{X}_k) = -\frac{d\ \ln(S(t\|\boldsymbol{X}_k))}{dt} = \frac{f(t\|\boldsymbol{X}_k)}{S(t\|\boldsymbol{X}_k)}$	[50, 49]	2.7.1
Odds ratio	$\frac{O(t\|\boldsymbol{X}_1)}{O(t\|\boldsymbol{X}_0)}$, $O(t\|\boldsymbol{X}_k) = \frac{1-S(t\|\boldsymbol{X}_k)}{S(t\|\boldsymbol{X}_k)}$	[69]	2.7.2
RMSTD(t^*)	$\text{RMST}_1(t^*) - \text{RMST}_0(t^*)$, $\text{RMST}(t^*) = \int_0^{t^*} S_k(t)dt$	[168, 170]	2.7.3
Percentile Ratio	$q_k = \frac{k^{\text{th}} \text{ percentile of dist for group A}}{k^{\text{th}} \text{ percentile of dist for group B}}$	[114]	2.7.4

RMST = Restricted Mean Survival Time, D = Difference.

dependent effect, fractional polynomials or splines can be applied. [163, 164, 165]

Two methods have been developed for combining fractional polynomials or splines in the two-stage approach. The meta curve method directly meta-analyzes the curves estimated in the first stage. Though, this requires setting a reference level which may have an impact on the results. Alternatively, by using multivariate meta-analysis (section 2.5.3) the coefficients can be combined. This method only works when the same polynomials or splines have been fitted in each study, but that is not an issue when IPD are available. [166]

Alternatively, non-PH can sometimes be handled more naturally with models that assume proportionality on another scale. [73, 61] For instance, an intervention might temporarily reduce the hazards, but as time progresses and the effect wears off, hazards converge and thereby violate the proportional hazards assumption. This can be modeled with a proportional odds regression model such as the log-logistic (equation 2.1.7, Table 2.1), which assumes that covariates have a constant additive effect on the log odds of survival. [167, 62, 63, 69] In this model, the modeled hazard ratio naturally approaches 1 over time, whereas the odds remain proportional. [62]

As the implementation of TTE models with non-proportional hazards (e.g. with splines) may complicate the interpretation of regression parameters, alternate effect measures have been proposed to summarize intervention effects (Table 2.7). For instance, the restricted mean survival time (RMST, equation 2.7.3) until time t^* represents the area under the survival curve until time t^*. [168, 169, 170] The RMST can thus be calculated for different intervention groups, and subsequently be subtracted to assess the intervention effect. This difference represents the expected gain (or loss) in survival until time t^* for the intervention group, as compared to the control group. An advantage is that it provides a clinically meaningful summary of the survival differences between intervention groups.

The percentile ratio, an effect measure alternative to the more common haz-

ard ratio, was suggested by to make the interpretation of survival models more straightforward. [114] Briefly, the percentile ratio for an intervention is defined as the expected ratio for the time at which a certain fraction (given as 'k') of the participants will have an event in the intervention group as compared to the control group (equation 2.7.4). The percentile ratio is easiest to interpret for AFT models, as the percentile ratio does not depend on the percentile chosen in such models and always equals the acceleration factor. Two-stage MA methods for the percentile-ratio have also been developed. [171]

2.5.3 Modeling multiple outcomes

Throughout this manuscript, we have assumed that each patient in each trail is at risk of having a single type of event (i.e. the outcome of interest, e.g. all-cause mortality), until censoring takes place. Alternatively, patients may be at risk for different events, where one event (e.g. death) prevents the patient from having another event (e.g. liver failure or stroke). Unlike the survival function, relative intervention effects can then still be assessed by modeling cause-specific hazards, which involves the modeling of the time to each type of event in a separate model, where all alternative types of event are coded as censoring. [88, 172] It is vital to do this for every type of event, to gain a full understanding of the relative intervention effect with respect to competing events. [88] Whereas for all-cause-mortality there is a direct relation between the hazard and the survival curve, when modeling cause-specific hazards this is not the case, [173] meaning that this approach does not have a direct interpretation in terms of absolute survival probabilities for the outcome of interest. [87] Only when independence of the event of interest and the competing event can be assumed, the survival function can be estimated by recoding the competing outcome as censoring, though this assumption is often not realistic. [90]

Therefore, when prediction of the average time-to-event per intervention group is wanted, competing events must be modeled using more complex survival models (for an introduction see [88, 174]). In the two-stage approach, this can be analyzed with competing risk models in the first stage, whereas Bayesian hierarchical competing risk models have been developed for the one-stage approach, [91] which may also model recurrent events jointly with the competing risk. [159] Further, multi-state models can be used to model transitions to intermediate events. [175]

When multiple outcomes that do not compete are available across trials, these can be assessed jointly in the two-stage framework to improve the efficiency of the analyses. [176, 177] For instance, outcomes may have been assessed at multiple follow-up times, or be defined for multiple endpoints. In the first stage, estimates of the intervention effects and variances are obtained for each outcome in each trial. Bootstrapping is used to obtain the covariance between intervention effects for each pair of outcomes in the same trial. [177] In the second stage, the vectors of estimates (and matrices of variances and covariances) are synthesised using a multivariate meta-analysis model in the second stage. Hence, multivariate meta-analysis is particularly relevant to address outcomes or time-points in the IPD from some trials.

2.5.4 Modeling multiple interventions

The concepts of multivariate meta-analysis can also be used to compare more than two interventions. In a so-called network meta-analysis (NMA), direct and indirect evidence about the difference in effect of two or more treatments is combined across trials, to summarize the relative effects of all available interventions. This may improve precision of the intervention estimates and allows for comparison of interventions that have not been compared head-to-head. This method uses direct evidence (intervention effects estimated within trials) and indirect evidence (intervention effects estimated across trials), by assuming that both sources of evidence are exchangeable. [178, 179]. When direct and indirect evidence disagree, the network is said to be inconsistent and may be prone to bias or may cause heterogeneity of the estimated intervention effects. Such inconsistency can be caused by effect modification, which can be addressed by modelling interactions between the intervention and patient-level covariates. [160]

In the two-stage approach, an appropriate (e.g. Cox) survival model is first estimated in each trial, possibly adjusting for relevant prognostic factors and effect modifiers. Corresponding effect estimates (e.g.log hazard ratios) can then be pooled using traditional NMA methods. [178] In the one-stage approach, time-to-event NMA models can be estimated using Bayesian hierarchical models. [180, 181] Also, Bayesian one-stage IPD-NMA Royston-Parmar models have been implemented. [75]

2.5.5 Surrogate endpoints

Trials for measuring intervention efficacy tend to be expensive and require a lengthy follow-up to observe the clinical outcome. The cost and duration of a trial may be reduced if a more readily available outcome can be used. Validated surrogate endpoints can be used instead when the surrogate is well known or likely to predict clinical outcome. [182] These surrogate endpoints are to be validated on the trial and the participant level, where IPD form multiple trials are preferred. [183, 184] When response to intervention is used to predict survival, response must be modeled as a time-dependent covariate or a landmarking method must be used. [185] Alternatively, a joint model with the survival outcome and a continuous surrogate or a dichotomous surrogate can be used. [186, 187] For an overview and comparison of the performance of measures of surrogacy, see [188, 189].When few trials are available, the trial level surrogacy cannot reliably be estimated using AD alone. However, surrogacy can sometimes be estimated on the center level by splitting multi-center data by center. [184, 190] This requires IPD when center specific parameter estimates are not available. For a recent overview of methods for estimating surrogacy, see [190]. To include a surrogate directly in the modeling of the outcome, a joint model can be used. [186, 187] For the one-stage approach, joint models with up to three levels have also been developed. [155]

2.5.6 Missing data

In a meta-analysis of survival data, several types of missing data may occur. It is possible, for instance, that not all studies provide IPD and thus that only AD are

available for some of the studies. In such cases, it is recommended to combine the available IPD and AD, as otherwise estimated intervention effects may be prone to (data availability) bias and overly large standard errors. [191] Including AD in a two-stage meta-analysis approach is fairly straightforward, provided that the model used for generating the AD is compatible with the models for analyzing the available IPD. It is also possible to directly combine IPD and AD using a one-stage meta-analysis, although this requires more advanced models, such as Bayesian hierarchical regression. [192]

Another common type of missing data occurs when events of individual subjects are censored, e.g. due to loss of follow-up. Survival models such as the Cox PH model and the AFT model readily account for this censoring, provided that it is not related to the outcome, conditional on any participant-level characteristics in the model (i.e. non-informative). When the assumption of independent censoring is challenged, its implications can be evaluated by adopting multiple imputation methods. [193]

Finally, it is possible that subject-level covariates are missing for one or more studies. Although participant covariates are not commonly used when estimating relative intervention effects from RCTs, they are crucial in IPD-MA of time-to-event data because of selection differences across trials (see section 2.2). When relevant participant-level covariates are missing for some trial participants, it is generally recommended to apply multiple imputation. [194] Hereby, researchers should adjust for the event indicator and the Nelson-Aalen estimator of the cumulative hazard, [93, 95] and also account for the presence of clustering. The latter can be achieved by adopting imputation models with mixed effects, which also facilitates imputation of covariates that have not been measured in one or more studies. [94, 195, 96, 196, 197]

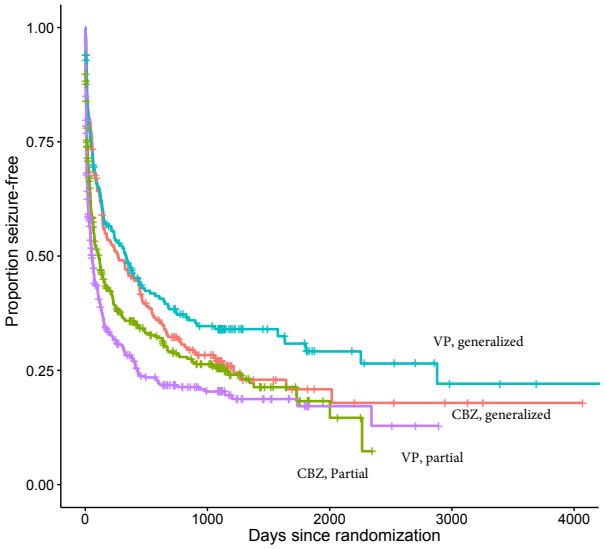
Although the assumptions needed for multiple imputation cannot always be tested or may not always be met, several simulation studies have shown that its use is usually superior to complete-case analysis or the use of missing data indicators. [92] However, caution is still warranted when analyzing imputed data sets from IPD-MA, as in the presence of between-trial heterogeneity these are inherently prone to some degree of incompatibility with the data generation mechanism. [198, 196] Further, because IPD-MA can only adjust for measured covariates and may therefore still be affected by unmeasured covariates, clustering of participants within trials should still be accounted for (section 2.4.3). [49]

2.6 Applied example

The efficacy of carbamazepine (CBZ) and valproate (VP) as interventions for epileptic seizures was compared in a systematic review and IPD-MA of RCTs. [199] IPD were obtained for a total of 1225 participants from five trials. In all these trials, one of the outcomes of interest was time to first epileptic seizure since randomization. Also, measured covariates were age at randomization, sex, type of epilepsy (partial-onset or generalized-onset), and the number of epileptic seizures before randomization. For illustrative purposes, we only consider the type of epilepsy. We use the `coxme` package of the `R` software,[200, 201] to fit the mixed effects Cox PH model. Our code is given in Supporting Information 2 (https://doi.org/10.1002/jrsm.1384).

As the two-stage method has been described extensively (see [171, 202]) we shall restrict our analyses to illustrate some key one-stage methods. First, to evaluate the relative effects of CBZ and VP, we adopted a Cox model, as this leaves the baseline hazard unspecified. We apply a one-stage model (eq. 2.5.2) with a log-normal frailty and random effects for the intervention estimated with penalized partial likelihood to account for the clustering of participants within trials and to allow for heterogeneous intervention effects across trials, respectively. We find no evidence against the hypothesis that the interventions are equally effective, with a summary hazard ratio of 1.08 for valproate (95% Confidence Interval (CI): 0.92 to 1.27, p = .37), versus the referent, carbamazepine.

Figure 2.2: Kaplan-Meier plot of Generalized and Partial Epileptic Seizure Patients Treated with Carbamazapine (CBZ) or Valproate (VP)



In the analysis of the effect of the intervention on the time to first epileptic seizure, we observed some statistical heterogeneity of the intervention effect. The standard deviations of the random intercept (i.e. frailty) and drug effect (i.e. random effect) equaled 0.139 and 0.099, respectively. In other words, the log hazard ratio of valproate versus carbamazapine varied with a standard deviation of .099 between trials. This random effect of the interventions translated to a Median Hazard Ratio (MHR) of 1.10, meaning that the median relative change in the effect on time-to first epileptic seizure when comparing two identical participants from two randomly selected different trials that were ordered by intervention effect was 1.10, calculated as $\exp\{\sqrt{2}\ 0.099\ \Phi^{-1}(0.75)\}$ (see section 2.4.3). In order to explain this heterogeneity in intervention effect, we added covariates and intervention-covariate interactions to the model (Figure 2.2). Partial epilepsy (vs generalized) was associated with a higher hazard rate ($\beta = 1.63$, 95% CI: 1.38 to 1.92, Table 2.8), meaning

that we have found evidence that epilepsy type is a prognostic factor of time to first epileptic seizure. However, we were unable to find evidence that epilepsy type interacted with the intervention ($\beta = 1.36$, 95% CI: 0.97 to 1.89), though it should be noted that the upper bound of the CI did not exclude clinically significant effects. We note that we obtained somewhat different results than the Cochrane review, [203] as we have used a different method for analysis. Further, the low power for tests for interaction effects is a notorious issue.

A recent investigation of the intervention-covariate interaction on the time to remission of epilepsy demonstrated that bias occurs when within-trial and across-trial information is not separated. [83] Such separation can be performed by centering the covariates, hence we have centered the covariates in in our analysis (Table 2.8). The possible bias that may occur when within-trial and across-trial information are amalgamated can be quantified by including the trial-mean in the model, [83] as we have done here (Table 2.8).

2.7 Discussion

Our search has identified a wide range of articles on topics regarding TTE IPD-MA, and is the first comprehensive review on this topic to our knowledge. However, the basics of the methodology regarding TTE data was excluded from our search as it did not concern MA or clustered data. Covering all methodological works regarding TTE data would have been an immense task. As such, we were forced to include relevant literature based on our own opinion to introduce this topic, and restrict our systematic search through Pubmed and Web of Science to works that simultaneously concerned IPD, meta-analysis and time-to-event data. We did not cover every article that covers these three topics, as this was not our aim. Instead, we our purpose was to achieve theoretical saturation, i.e. that an extended search would be unlikely to add important information.

The general consensus in the reviewed works was that the Cox model should be the default model of choice for TTE IPD-MA. Though, it is also criticized for not yielding a valid estimate of intervention effect when not all (un-)measured predictive covariates are accounted for, mostly on theoretical grounds. The literature is currently missing information on the impact of this issue in real life data, leading us to suggest that further research should focus thereon. As such, we have provided a comprehensive review of current methods for IPD-MA of TTE data.

Although the statistical properties of the meta-analysis estimators for the two-stage approach have been well studied and simulation studies have investigated the performance for meta-analysis of dichotomous and continuous outcome data, this is not the case for time-to-event data. Further, although aggregate data (i.e. estimates from the literature) can readily be included in the two-stage approach (provided that the models are specified the same), as well as in Bayesian one-stage models, there appears to be no method yet for doing so in a Frequentist model.

Table 2.8: Intervention, Covariates and Intervention-Covariate Interactions in a Multivariable Mixed Effects Cox Model

Variable	Variable Type	HR	95 % CI	p
VP (vs CBZ)	Intervention	1.05	0.86 to 1.28	0.65
Partial epilepsy (vs generalized), centered	Individual-level covariate	1.63	1.38 to 1.92	$< .001$
Partial epilepsy (vs generalized), trial mean	Trial-level covariate	1.47	0.99 to 2.19	0.06
Partial epilepsy (vs generalized), centered * VP (vs CBZ)	Intervention-covariate interaction	1.36	0.97 to 1.89	0.07

VP: Valproate, CBZ: Carbamazepine, HR: Hazard ratio, given by $exp(\beta)$, CI: Confidence interval. Standard deviations of random intercept (i.e. frailty) and random effect of VP (vs CBZ) equal 0.126 and 0.164, respectively. p-values are for Wald type tests of the null hypothesis that the log HR equals zero.
Covariates are centered within trials, to avoid ecological bias (see[83]).
Trial mean value for the covariate is entered in the analysis, to quantify the bias that would occur if centering of the covariate were not performed.

Another issue is to what extend one should try to borrow information across trials in the one-approach. In the two-stage approach, no information is borrowed (apart from the intervention effect and its uncertainty), as all parameters are naturally estimated per trial. To what extend one should account for this in the one-stage approach, by stratifying the baseline and covariate effects or by applying random effects and a frailty, deserves extra attention in the literature. For the meta-analysis of trials with adequate sample sizes, the safest choice is to stratify all included parameters as this accounts for all differences in baselines between trials. In a simulation study where IPD from a total of 600 participants from 3-20 trials were generated, both the frailty and the stratified baseline method worked well, [80] though exactly what sample sizes are necessary for this strategy, and especially for the stratification of covariates as well, has apparently not yet been identified.

2.8 Concluding remarks

We have discussed numerous models in this manuscript, the choice between which is not always straightforward. For this reason, we provide some recommendations below. First, intervention effect conditional on covariates and/or frailties have different interpretations from marginal ones (i.e. averaged over the entire sample and follow-up time), and yield different estimates. Before embarking on an IPD-MA, researchers should decide whether a conditional or a marginal effect is of interest. As assumptions may be satisfied on one scale but not the other, this may lead to a different choice of model.

Additionally, one can choose between one-stage and two-stage models. In the two-stage method participants within trials are compared, which inherently yields a conditional intervention effect and stratified baselines. The one-stage approach offers more possibilities as it allows for conditional intervention effects as well as marginal ones, and frailties for the baseline. When the same (or similar) model assumptions are made for these models and the same estimation methods are used, these two approaches generally lead to the same estimates of intervention effect. [48, 82] Though, the one-stage approach can have better convergence properties when the included studies are very small, [204, 82] or at least one of the studies has zero events.

Further, when a conditional effect is desired (in contrast to a marginal one), we recommend to apply random effects instead of common effects, as common effects models are only valid when no heterogeneity is present, which is unlikely in our experience. When a marginal effect is desired, only a correction for the variance is necessary. As described in section 2.2, when an intervention effect is present the estimated intervention effect in PH models may be time-dependent, depending on the distribution of prognostic factors that are not accounted for (even if balanced across intervention groups). This may lead to heterogeneity in intervention effects across trials that have different follow-up lengths. Further, differences in trial design and methodology and clinical procedures may contribute to the heterogeneity of the intervention effect. [47] Random effects models can account for heterogeneity of the intervention effect and lead to the same solution as common effect models when no heterogeneity is present. However, if a formal test of heterogeneity is desired, a

variety of tests can be used. For one-stage meta-analysis, the common effect model (without trial effects) is nested in the frailty model, and therefore a comparison of these models can be made using the log-likelihood ratio test. [49] Alternatively, a score test, [205, 206, 207] or a small sample test can be used. [208] A permutation test for testing of the presence of heterogeneity in time-to-event data was recently proposed, and a simulation showed that the method is more powerful and has a better type I error rate than likelihood ratio tests of a random effect. [209]

Finally, when comparing non-nested (e.g. PH versus AFT) models, more general methods are needed. In such cases, one may select the model with lowest value for Akaike's Information Criterion (AIC)[210, 211] or the Bayesian Information Criterion (BIC)[212, 210, 211]. Though, due to the correlated nature of participants within trials a correction for clustering should be made, which is not straightforward in the frequentist estimation framework as quantification of the number of degrees of freedom is difficult. For subject-specific inferences, the conditional (cAIC) can be used, whereas for inferences on the population level the marginal AIC can be used. [213, 79, 139, 214]

Further, one should be cautious regarding model selection. If one model is rejected, bias will appear in the estimated intervention effect and significance in a second model if the second model is not independent of the test that was used to reject the first, such as when a non-PH effect is included in the model after a statistical test indicated non-proportionality. [215] This bias can be alleviated by bootstrapping the model selection procedure. On the other hand, this bias does not occur when the second model is independent of the test used to reject the first model. [215]

Highlights

What is known?

- Time-to-event (survival) data can be analyzed with Cox Proportional Hazards regression, but proportionality of hazards should be tested.
- Individual participant data (IPD) from multiple randomized trials can be summarized by meta-analyzing the trial-specific estimates of the individual trials (studies) or by analyzing the pooled data with a mixed-effects model that accounts for between-trial heterogeneity in intervention effect and frailty of participants.

What is new?

- We summarize published guidance, statistical methods and software for survival analysis using IPD from multiple randomized clinical trials.
- We discuss how between-trial heterogeneity of intervention effects may appear and how its sources can be investigated.
- We illustrate the methods on real epilepsy data and provide R code.

Potential impact for other fields

- Meta-analysis is not only relevant in medical research, but also in other research areas.
- The methods naturally extend to meta-analysis of non-randomized studies, where treatment effect estimates need to be adjusted for confounding.

Acknowledgements

This work is financially supported by the Netherlands Organization for Health Research and Development grant 91617050 for TD and grant 91810615 for VJ and KM, and the European Union's Horizon 2020 Research and Innovation Programme under ReCoDID Grant Agreement no. 825746 for VJ and KM. The data that support the findings of this study are not publicly available, according to the conditions determined by the Epilepsy Monotherapy Trial Group, but are available on request from AM, by e-mailing A.G.Marson@liverpool.ac.uk. We would like to thank the editor and reviewers for thoughtful comments on the manuscript and suggesting items for the review.

Chapter 3

Developing more generalizable prediction models from individual participant data meta-analyses and large clustered data sets

Valentijn M.T. de Jong, Karel G.M. Moons, Marinus J.C. Eijkemans, Richard D. Riley and Thomas P.A. Debray
Submitted

Abstract
Prediction models often yield inaccurate predictions for new individuals. Although large data sets from individual participant data meta-analysis or electronic healthcare records may alleviate this, prevailing strategies for prediction model development generally do not account for heterogeneity between settings and populations. This limits the generalizability of developed models (even from large, combined, clustered data sets) and necessitates local revisions. We aim to develop methodology for producing more robust prediction models that require less tailoring when applied to different settings and populations.

We adopt Internal-External Cross-Validation to assess and reduce heterogeneity in a model's predictive performance during its development. We propose a predictor selection algorithm that optimizes the (weighted) average performance whilst minimizing its variability across the hold-out clusters (or studies). Predictors are added iteratively until the estimated generalizability is optimized. We illustrate this methodology by developing a new model for predicting the risk of atrial fibrillation and updating an existing one for diagnosing deep vein thrombosis. We used individual participant data from 20 cohorts (N = 10873) and 11 diagnostic studies (N = 10014), respectively. Meta-analysis of calibration and discrimination in each hold-out cluster shows that trade-offs between average performance and heterogeneity occurred.

Our methodology allows for the assessment of heterogeneity of prediction model performance during model development in multiple or clustered data sets, thereby informing researchers on predictor selection to minimize heterogeneity. This may improve the generalizability to different settings and populations, and reduce the need for tailoring the model. Our methodology has been implemented in the `R` package `metamisc`.

3.1 Background

Large combined clustered data sets are increasingly available, for example in so-called individual participant data meta-analyses (IPD-MA) projects (where the data are clustered by study) and in studies using large scale electronic healthcare records (where the data are clustered by region, hospital, practice, etc). [216] Such data sets are frequently used to develop prediction models, to predict a current health status to aid in diagnosis or a future health outcome to provide a prognosis which may inform clinical decision making. [3, 4, 5] Well known examples are PHASES, [217] INTERCHEST, [218] S_2TOP-BLEED, [219] and EuroSCORE, [220] all of which were developed using data from multiple centers or studies. Unfortunately, prediction model studies that are based on IPD-MA or electronic healthcare records (EHR) rarely account for the potential of between-cluster heterogeneity (e.g. EuroSCORE [220]). [221, 15] Sometimes, parameters that capture the baseline risk are stratified by cluster (e.g. INTERCHEST[218]), but then usually no guidance is provided on how to use the prediction model in new patients.

Although random effects models are generally recommended for dealing with the presence of clustering and heterogeneity, their implementation during prediction model development hampers the applicability of the estimated regression coefficients. In particular, random effects modelling does not indicate which parameter values (for the random intercept and predictor coefficients) should be used when the model is applied in new settings and populations. Typically, a single value (e.g. the mean) is used for these parameters when making predictions.

In general, a developed prediction model cannot generate accurate predictions in new patients when the true value for its parameters (e.g. the intercept term) varies across the targeted settings and populations, especially when the true value of certain parameters is zero or has a reversed sign in some clusters. This heterogeneity may arise from differences in observed and unobserved patient characteristics, differences in patients' treatment and management strategies, differences in predictor and outcome definitions and differences in measurement methods across clusters.

The impact of heterogeneity in predictive associations (i.e. the effects of predictors in the included model) has been well documented in the literature. [116, 25] Many developed prediction models perform poorer than anticipated and require local revisions prior to implementation. [11] These revisions may involve a simple intercept update, a recalibration of the linear predictor (i.e. rescale all regression coefficients by a single value), the re-estimation of all the individual regression coefficients, or even the inclusion of new predictors. [222, 223, 17, 224] Unfortunately, revisions are rarely generalizable to other settings and populations; several reviews have found that prediction model performance substantially varies across validation studies. [225, 19] Therefore, such revisions, including recalibration and predictor selection, are preferably performed during prediction model development.

The identification of heterogeneity is not possible when data are available from only a single setting or (sub)population. For this reason, the use of clustered data during prediction model development and its subsequent validation offers a critical opportunity to inspect whether this heterogeneity would actually be a concern when the model would be implemented in clinical practice. [226, 116, 221, 15, 26, 23, 20, 21, 227, 25] However, actually resolving the presence of heterogeneity (and

thus ensuring model predictions are accurate for all clusters) remains a difficult challenge for which limited guidance is available. [228] For this reason, we here explore an alternative approach that aims to reduce this heterogeneity and minimize the need for estimating setting-specific model parameters, to thereby improve its generalizability.

Recently, internal-external cross-validation (IECV) has been introduced to assess the presence of heterogeneity of a model's performance during its development. [226, 15, 23] IECV is a special case of cross-validation; available data are split non-randomly in a natural manner by iteratively taking each cluster (or study) as a hold-out sample. In each iteration, a model is developed on the retained clusters, and then the model is tested in the hold-out cluster. A key advantage of this is that it allows the transportability (i.e. the generalizability to other populations and settings) of the model to be assessed multiple times.

In this paper, we will first revisit the IECV framework for assessment of model performance in large clustered data sets (section 3.2). We then extend the IECV framework to inform predictor selection during prediction model development in section 3.3, in order to identify and reduce their impact on the model's performance within and across clusters in the large combined dataset. We then apply the methods in our motivating examples in section 3.4 and 3.5. Finally, we provide a discussion in section 3.6. Our methodology can be applied using the R package `metamisc`. [229]

3.2 Internal-External Cross-Validation for Model Validation

Resampling procedures allow the optimal use of the available data, as all data can be used for model development and subsequent evaluation. Traditionally in cross-validation procedures, the data is iteratively split into a development and validation set by randomly sampling without replacement. In each iteration, a model is estimated on the development sample and predictions are made for the random validation sample. The performance of these predictions in the validation samples is then averaged across iterations, thereby giving an estimate of the reproducibility of model performance.

When data are clustered across different studies or settings, traditional resampling procedures that do not account for clustering cannot directly be applied. [230] For this reason several extensions have been proposed that preserve the clustering within and the heterogeneity across the generated samples. In the so-called Internal-External Cross-Validation approach, the data is split by cluster, which may represent the studies from an IPD-MA or the centers in data from EHR. [226, 20, 23] A model is then iteratively fit in K-1 clusters (section 3.2.1) and its corresponding performance model performance is calculated in the remaining cluster (section 3.2.2). This is repeated K times, so that, provided that sufficient data are available in the development and validation clusters, a performance estimate and its standard error is available for each of the clusters. Thus, IECV is cross-validation where the hold-out samples are non-random, in the presence of between-cluster heterogeneity.

IECV therefore allows the study of a developed model's potential transportability multiple times. Note that if all patients are exchangeable across clusters, IECV corresponds to the traditional cross-validation and assesses model reproducibility (rather than transportability). [15]

In contrast to traditional cross-validation, estimates of the performance in the hold-out samples cannot simply be averaged, as the variation within and across clusters needs to be taken into account. This can be achieved by adopting a (fixed- or random-effects) meta-analysis of the performance estimates (section 3.2.2), [231] or by weighting the performance estimates by the number of events in each cluster. [232] As the data is split non-randomly, this allows the transportability (i.e. the generalizability to other populations and settings) of the model to be assessed.

3.2.1 Model fitting

The development phase of IECV may involve a one-stage or a two-stage IPD-MA approach. In the two-stage approach, the prediction model is fitted separately in each cluster. The model coefficients estimated in each of the development K-1 clusters are then combined using standard meta-analysis techniques. In the one-stage approach, a Generalized Linear Model (GLM) is estimated in each of the K development samples consisting of $K-1$ clusters. This model may account for clustering by including random intercepts and/or predictor effects. [25, 116, 232, 228] A disadvantage of the one-stage approach in IECV is that the data from each cluster needs to be used K-1 times to fit a model in the one-stage approach. On the other hand, in the two-stage IECV approach the data from each cluster only needs to be used for model fitting once, as the second stage comprises meta-analysis of different combinations of coefficients and their standard errors. The two-stage approach may therefore substantially reduce the necessary computational performance time. However, the two-stage approach may not be feasible when clusters are relatively small, as parameters then become difficult to estimate. For this reason, the two-stage approach appears beneficial when most clusters (studies) in the meta-analysis are not small, and we adopt this approach in our article.

Let $x_{p,k,j}$ be the value of a pre-specified predictor p, p = 1,...,P (or function thereof) measured in individual patients $j, j = 1, ..., N$ in cluster $k, k = 1, ..., K$. Then their outcomes $y_{k,j}$ may be modeled as follows:

$$y_{k,j} = f(\alpha_k + \sum_{p=1}^{P} \beta_{p,k} x_{p,k,j}), \tag{3.1}$$

where $f(...)$ is a link function, α_k is a cluster-specific intercept and $\beta_{p,k}$ is a cluster-specific coefficient. Here, we propose to estimate α_k and $\beta_{p,k}$ in each cluster separately. Subsequently, the estimates can be summarized using traditional meta-analytic methods. We here use univariate random effects meta-analysis, where each of the estimated coefficients are summarized separately:

$$\hat{\beta}^{\mathrm{MA}}_{p,(h)} = \frac{\sum_{k \neq h} w_{p,k} \hat{\beta}_{p,k}}{\sum_{k \neq h} w_{p,k}}, \tag{3.2}$$

where $w_{p,k}$ is the weight attributed to $\hat{\beta}_{p,k}$ estimated in cluster k, and $\hat{\beta}^{\mathrm{MA}}_{p,(h)}$ is the meta-analytic estimate of the coefficient estimated on data from all clusters except hold-out cluster h. In the random-effects model the $w_{p,k}$ are given by $\frac{1}{\mathrm{var}(\hat{\beta}_{p,k})+\tau^2}$, where τ^2 is the statistical heterogeneity estimate of the coefficient across clusters:

$$\begin{aligned} \hat{\beta}_{p,k} &\sim \mathcal{N}\left(\beta_{p,k}, \mathrm{var}\left(\hat{\beta}_{p,k}\right)\right) \\ \beta_{p,k} &\sim \mathcal{N}\left(\beta^{\mathrm{MA}}_{p,(h)}, \tau^2_{p,(h)}\right) \end{aligned} \tag{3.3}$$

A confidence interval (CI) for $\hat{\beta}^{\mathrm{MA}}_{p,(h)}$ is preferably constructed with the Hartung-Knapp approach: $\hat{\beta}^{\mathrm{MA}}_{p,(h)} \pm t_{Q-1,1-\alpha/2}\sqrt{var_{\mathrm{HK}}(\hat{\beta}^{\mathrm{MA}}_{p,(h)})}$, where $t_{Q-1,1-\alpha/2}$ is the upper $\alpha/2$ quantile of a t-distribution with $Q-1$ degrees of freedom, $var_{\mathrm{HK}}(\hat{\beta}^{\mathrm{MA}}_{p,(h)})$ is a modified variance estimate and $Q = K-1$ as one cluster is held out for validation. [105, 106, 107, 108, 109] The extent of heterogeneity of a predictor effect can be explored by quantifying a prediction interval (PI), which estimates the interval of probable predictor effects in a new individual cluster, and can be calculated approximately as $\hat{\beta}^{\mathrm{MA}}_{p,(h)} \pm t_{Q-2,1-\alpha/2}\sqrt{\hat{\tau}^2_{p,(h)} + var(\hat{\beta}^{\mathrm{MA}}_{p,(h)})}$. [110, 85] A wide prediction interval for the predictor effect indicates that the predictor effect may be very different in a new cluster, which makes it unlikely that the predictor will improve the model's predictions for individuals in a new cluster.

The random effects meta-analysis model is preferably estimated with REML or the Paule-Mandel method. [120, 121, 118, 122] When fewer than 10 clusters are included in the meta-analysis, or when some clusters are small or the outcome is rare, the heterogeneity cannot be reliably estimated by any currently available method. [122] The estimated coefficients could also be summarized using multivariate meta-analysis methods, [233, 231] which may be helpful in the presence of collinearity and missing parameter estimates. The necessary within-cluster covariances can then directly be estimated from the IPD set at hand. However, usually univariate and multivariate meta-analysis methods give very similar results when all of the parameter estimates of interest are available for all clusters, even when correlations are large. [179] In IECV, all parameters can be estimated from the data hand, meaning that univariate meta-analysis will usually suffice.

3.2.2 Assessing external model validity

In each iteration of the IECV, the developed model is validated in individuals from the hold-out cluster by applying the model (as developed in the other clusters) using the observed predictor values of individuals. If the developed model contains random (or stratified) intercept terms or predictor effects, this also requires choices about which parameter values are to be used when applying the developed model.

When comparing the risk predictions for the hold-out cluster with the observed outcomes, several performance measures such as the c-statistic, calibration slope, calibration intercept and/or mean square error can be calculated. [15, 21, 234] This process is repeated until each cluster has been used as a hold-out cluster once, yielding a set of performance statistics for each IECV iteration. The corresponding

estimates can then be pooled across the hold-out clusters using random effects meta-analysis methods, though some statistics and their standard errors may require transformation first. [231, 235, 228] Similar to the predictor effects, a prediction interval can then be constructed for the performance estimates, which provides an interval of likely values that the performance statistic will have in a new cluster.

Besides allowing one to obtain an average estimate of performance, meta-analysis is particularly helpful for investigating the presence of heterogeneity and any possible causes thereof. [13, 15, 231] Prediction model performance may vary across clusters due to imprecision or bias of the regression coefficients or performance estimates, or due to the variation in population characteristics. Disentangling these various sources of variation is necessary when inferring on the model's potential generalizability to different settings and populations.

Finally, if the average performance and heterogeneity of the performance of the prediction model are deemed adequate, that is it is considered likely that performance will be adequate in a new cluster, a so called global model may be developed by estimating the coefficients for the predictors on the data of all available clusters. In this final step no clusters are left out, in order to minimize the variability of the estimates of the coefficients. [23]

3.2.3 Motivating example: diagnosis of deep vein thrombosis

Patients with a deep vein thrombosis (DVT) have an increased risk of post-thrombotic syndrome and pulmonary embolism, which can be fatal. [236] In the majority of patients in whom DVT is suspected, no DVT is present on advanced (reference) testing. [237] For illustrative purposes, we here consider the diagnosis of DVT in patients that are suspected of having DVT and use the IPD of 10014 patients from eleven studies, [238] where each study is considered one cluster (Table 3.1 and 3.2). In each cluster separately, we estimated a binary logistic regression model with three pre-specified predictors: history of malignancy (yes/no), calf difference (difference in circumference of the calves $\geq$ 3 cm), recent surgery (yes/no). Preferably, a continuous predictor such as calf difference should not be dichotomized, as this leads to a loss of information. However, the continuous predictor was not available in the data at hand. As some clusters were small, we applied Firth's correction, [99] which yields unbiased Maximum Likelihood estimates for the coefficients and standard errors in small samples [239] and adjusted the intercept post-hoc by re-estimating it with an unpenalized GLM. [240] We then applied IECV and adopted a two-stage approach for prediction model development. The pooled regression coefficients (including the intercept term) from the development clusters were used for generating predictions in the hold-out cluster. Although Firth's correction still yielded estimates with high variance for the predictor coefficients in some clusters, this was mitigated by performing a meta-analysis of the regression coefficients.

Table 3.1: Clinical Characteristics of DVT Data

Outcome: DVT		No	Yes	Total
Sex	Female	5174 (83.8)	1001 (16.2)	6175
	Male	2943 (76.7)	896 (23.3)	3839
Malignancy	No	7600 (82.8)	1581 (17.2)	9181
	Yes	517 (62.1)	316 (37.9)	833
Recent surgery	No	7333 (82.4)	1569 (17.6)	8902
	Yes	784 (70.5)	328 (29.5)	1112
Leg trauma	No	5210 (77.1)	1544 (22.9)	6754
	Yes	2907 (89.2)	353 (10.8)	3260
Vein distension	No	7257 (82.5)	1538 (17.5)	8795
	Yes	860 (70.5)	359 (29.5)	1219
Calf difference > 3 cm	No	6160 (88.0)	843 (12.0)	7003
	Yes	1957 (65.0)	1054 (35.0)	3011
D-dimer abnormal	No	4392 (97.0)	137 (3.0)	4529
	Yes	3725 (67.9)	1760 (32.1)	5485
Age	Mean (SD)	58.8 (17.4)	61.1 (17.1)	10014
Duration of symptoms	Mean (SD)	22.8 (45.5)	27.0 (60.5)	10014

Results in Table 3.2 reveal that estimates for the predictor effects were very heterogeneous across the included clusters. For example, the coefficient for malignancy was 0.90 (standard error, SE: 0.33) in cluster 1 and 1.69 (SE: 0.22) in cluster 7. Similarly, the coefficient for calf difference was 0.98 (SE: 0.15) in cluster 2 and 1.68 (SE: 0.13) in cluster 4. As indicated in Table 3.3 this also resulted in heterogeneous model performance estimates across hold-out clusters. Although calibration was good on average, it was highly variable in individual clusters. For instance, whereas the summary calibration intercept equaled 0.03 (95% CI: -0.33 to 0.39), meaning that calibration in the large was very good on average, the calibration intercept's approximate 95% prediction interval (PI) ranged from -1.22 to 1.27, thereby indicating heterogeneity. Similarly, the calibration of the linear predictors was very good on average, as the calibration slope (also estimated with Firth's correction) equaled 1.00 (95% CI: 0.83 to 1.16), whereas the approximate 95% PI for the calibration slope ranged from 0.53 to 1.46. Further, the c-statistic equaled 0.68 (95% CI: 0.65 to 0.71) and was also substantially heterogeneous across clusters (approximate 95% PI: 0.60 to 0.75).

Table 3.2: Estimated Regression Coefficients for Predicting DVT in each of Eleven Clusters

Cluster	Intercept	Malignancy	Calf difference	Surgery
1	−2.46(0.14)	0.90(0.33)	1.17(0.19)	0.04(0.35)
2	−0.95(0.11)	0.31(0.24)	0.98(0.15)	0.17(0.25)
3	−2.92(0.44)	1.57(0.87)	1.59(0.50)	1.73(0.54)
4	−1.92(0.09)	0.63(0.16)	1.68(0.13)	0.83(0.17)
5	−2.27(0.16)	0.24(0.42)	1.03(0.20)	0.52(0.26)
6	−2.25(0.12)	1.23(0.30)	1.40(0.17)	0.51(0.21)
7	−3.18(0.13)	1.69(0.22)	1.41(0.19)	0.26(0.31)
8	−1.72(0.18)	1.02(0.58)	1.24(0.27)	0.78(0.51)
9	−2.01(0.11)	0.80(0.25)	1.25(0.14)	0.37(0.19)
10	−2.16(0.18)	1.04(0.46)	0.65(0.34)	0.79(0.35)
11	−2.30(0.19)	1.65(0.26)	1.32(0.23)	0.82(0.27)
Summary estimate	−2.17(0.18)	0.98(0.17)	1.27(0.08)	0.55(0.09)
Approximate 95% prediction interval	−3.33 : −1.01	0.08 : 1.88	0.86 : 1.67	0.20 : 0.90

Malignancy: history of malignancy, Calf difference: difference in circumference of calves ≥ 3 cm, Surgery: recent surgery. Summary estimates and prediction intervals for global model.

On overall, the IECV showed that the modeling strategy was unlikely to yield a prediction model with good generalizability. Substantial revision would be necessary to improve the model's average discrimination performance and to reduce the heterogeneity of its calibration and discrimination performance. A possible approach would be to refine the original modeling strategy by altering the set of included predictors and by considering interaction effects and/or non-linear terms. Subsequently, the revised model should be validated again, after which other revisions may be decided and so forth. It may be clear that this strategy is very time consuming and may lead to arbitrary choices in predictor selection. For these reasons we propose a formal framework for predictor selection in the context of heterogeneity of performance across clusters in the next section. We address methods that aim to reduce heterogeneity of performance, improve the average performance and a combination thereof. The code used to apply our methodology as presented in this manuscript is available on Github (https://github.com/VMTdeJong/SIECV-DVT).

Table 3.3: Internal-External Cross-Validation Performance Estimates and Standard Errors for the Predefined Model for Predicting DVT

Hold-out cluster for validation	Slope (SE)	Intercept (SE)	c-statistic (SE)
1	0.86(0.15)	−0.44(0.10)	0.65(0.02)
2	0.63(0.11)	1.06(0.08)	0.63(0.02)
3	1.49(0.35)	−0.24(0.22)	0.78(0.05)
4	1.18(0.10)	0.43(0.06)	0.72(0.01)
5	0.73(0.15)	−0.33(0.10)	0.65(0.02)
6	1.12(0.14)	−0.00(0.08)	0.70(0.02)
7	1.24(0.14)	−0.93(0.09)	0.71(0.02)
8	1.02(0.24)	0.51(0.13)	0.67(0.03)
9	0.92(0.11)	0.12(0.07)	0.68(0.02)
10	0.71(0.27)	−0.09(0.14)	0.62(0.04)
11	1.27(0.17)	0.15(0.11)	0.74(0.03)
Summary estimates	1.00(0.08)	0.03(0.16)	0.68(0.06)
Approximate 95% prediction interval	0.53 : 1.46	−1.22 : 1.27	0.60 : 0.75

Slope: Calibration Slope, SE: Standard Error, Intercept: Calibration Intercept.

3.3 Stepwise Internal-External Cross-Validation for Model Development

In the previous section, we described the purpose of IECV to assess the generalizability of a prediction model that is generated by a predefined modeling strategy. Here, we propose to extend IECV to optimize model generalizability *during its development*. We consider the situation that IECV will be used to expand an empty (intercept only) model by iteratively adding predictors, functions of predictors and interaction effects. The approach also readily generalizes to the expansion or reduction of a given model. In this Stepwise IECV (SIECV) for prediction model development models are estimated, validated in external data sets and updated in an iterative process, as follows.

Denote the data from the k^{th} cluster by S_k, and the data from a set of clusters excluding cluster h by $S_{(h)}$. Let $p, p = 0, 1, ..., P$ be indicators to denote the candidate predictors (or functions thereof), where $p = 0$ indicates none. The algorithm consists of up to I model adaptation cycles, where I generally equals P, the number of predictors available for inclusion in the model. Then, let $P_r(i)$ denote the set of candidate predictors for inclusion, where $P_r(1) = \{1, 2, ..., P\}$ and $P_r(0) = \{0\}$.

Further, let $M_{i,p}$ denote the models in cycle i with added predictor p in the stepwise process. Let $M_{i,p,(h)}$ denote a model estimated on data from all clusters excluding S_h. Let $\hat{Z}_{i,p,h}$ be an estimate of performance (i.e. a loss function) of model $M_{i,p,(h)}$ in cluster h, such as the mean squared error. Let $\hat{A}_{i,p}$ be the estimate of

a loss function (i.e. an aggregated loss function or an estimate of heterogeneity, further described in section 3.3.2) in cycle i for a model extended with predictor p, and let c indicate a predictor p that has minimal $\hat{A}_{i,p}$, such that $M_{i,c}$ is the model with best generalizability in cycle i. Then, the algorithm is defined as follows and starts at cycle $i = 0$:

1. For all p in $P_r(i)$:
   (a) Extend model $M_{i-1,c}$ with predictor p to generate new model $M_{i,p}$.
   (b) For $h, h = 1, ..., K$:
      i. Estimate the model $M_{i,p,(h)}$ on $S_{(h)}$, preferably while taking clustering within clusters into account.
      ii. Predict $\hat{y}_{i,p,h,j}$ for individual participants in hold-out sample S_k.
      iii. Estimate performance measure $\hat{Z}_{i,p,h}$ and its standard error $\widehat{\mathrm{SE}}(\hat{Z}_{i,p,h})$ for predictions $\hat{y}_{i,p,h,j}$ in S_h.
   (c) Estimate aggregated loss function $\hat{A}_{i,p}$ on $\hat{Z}_{i,p,1}, ..., \hat{Z}_{i,p,K}$ and $\widehat{\mathrm{SE}}(\hat{Z}_{i,p,1}), ..., \widehat{\mathrm{SE}}(\hat{Z}_{i,p,K})$.
2. Find the minimal $\hat{A}_{i,p}$ in this cycle. Denote this by $\hat{A}_{i,c}$ and its corresponding model by $M_{i,c}$.
3. The first condition that is satisfied:
   (a) If $i = 0$, continue to step 1.
   (b) Else, if $\hat{A}_{i,c} \geq \hat{A}_{i-1,c}$, the algorithm stops and $M_{i-1,c}$ is returned as the final model.
   (c) Else, if $i = I$, the algorithm stops and $M_{i,c}$ is returned as the final model.
   (d) Else, remove predictor c from the candidate predictor set $P_r(i)$, increment i by 1 and continue to step 1.

Finally, if the performance of model $M_{i,c}$ is deemed satisfactory, a so called global model is generated by estimating the coefficients for the predictors in $M_{i,c}$ on all available data. No clusters are left out in this final cycle, to reduce the variance of the estimates of the coefficients. [23]

This global model however, is at risk of overfitting as a result of small sample bias, unless the sample is sufficiently large and the event rate sufficiently high, even if no selection of predictors were applied. [241, 242, 243] To account for this, the prediction model could be fitted with penalized regression, such as Firth's regression. To reduce the variance of the estimated regression coefficients, the ridge penalty could be applied instead, or one could opt for a fully Bayesian approach.

By considering the candidate predictors for inclusion, however, the prediction model is at further risk of overfitting. [3, 58] A straightforward adjustment for overfitting could be achieved with the calibration slope and intercept. [244] In

step 1 (b) iii of the final cycle these could be estimated and then summary meta-analyses estimates could be computed. The final model coefficients (excluding the intercept) would then be multiplied by the summary calibration slope, whereas the summary calibration intercept would be added to the global model's intercept, thereby yielding a final model. Ideally, however, the entire model selection procedure is to be performed within an additional bootstrap or cross-validation procedure, [245] as this would account for any overfitting introduced by the SIECV itself. Alternatively, heuristic shrinkage, that shrinks the coefficients by a function of the number of predictors considered, may be applied. [246, 3, 58]

3.3.1 Extensions

Throughout this manuscript, we work from the perspective that an entirely new prediction model is to be developed. However, our proposed framework readily encompasses model redevelopment including the adding and removal of predictor terms. Selection of predictor effects may then also be performed with a backwards procedure starting with all candidate predictors and their transformations or interactions, rather than forwards. Though, this may yield issues in the estimation when many predictor effects are considered, especially when random effects are applied. Further, similar to IECV for model validation we may adopt a one- or two-stage approach (section 3.2.1) for model estimation.

3.3.2 Quantifying model generalizability

The SIECV algorithm requires specification of an aggregated loss function ($A_{i,p}$) that is to be minimized, in order to optimize generalizability of performance across clusters. Here, we consider parametric and non-parametric aggregated loss functions, that vary with respect to the importance they place on the average and heterogeneity of performance.

Ignoring heterogeneity

As a first step, we consider a naive estimator of predictive performance across hold-out data sets from different clusters, that ignores variation within and across clusters. This approach may be reasonable when the clusters are very large and of similar size, and when the clustering is negligible. The overall performance is then given by the mean performance across clusters. For instance, when optimizing the mean square error (MSE, or Brier score for categorical outcomes), we can apply the following aggregated loss function:

$$\widehat{A_{i,p}^{\mathrm{M}}} = \frac{1}{K}\sum_{h=1}^{K}\hat{Z}_{i,p,h} \tag{3.4}$$

Weighted meta-analysis

To incorporate the uncertainty of the predictive performance estimates into an aggregated loss function, it may be more appropriate to adopt a weighting procedure.

The meta-analysis framework (see section 3.2.1) therefore appears an appealing choice. A straightforward extension to equation 3.4 would be to apply the weighting procedure in described in equations 3.2 and 3.3. This allows to minimize the prediction error in an "average" cluster, but still does not attempt to optimize their stability across clusters. As a result, it is possible that developed models perform well on average, but require substantial local revisions before implementation. To reduce the need for local revisions, the aggregated loss function should account not only for the average performance, but also for its variation across clusters. For this reason, we propose an extension that combines both sources of error:

$$\widehat{A_{i,p}^{\mathrm{RE}_\lambda}} = \lambda \hat{Z}_{i,p}^{\mathrm{RE}} + (1-\lambda)\hat{\tau}_{i,p} \tag{3.5}$$

where λ is a hyperparameter that defines the impact of random effects meta-analysis summary estimate of performance $\hat{Z}_{i,p}^{\mathrm{RE}}$ and heterogeneity estimate $\hat{\tau}_{i,p}$ on aggregated loss function $\widehat{A_{i,p}^{\mathrm{RE}_\lambda}}$. This is a parameter that is to be chosen on beforehand, where its value should depend on the relative importance of average and heterogeneity of performance. In the simplest case we let $\lambda = 1$, such that the estimate for generalizability is given by the mean of the distribution of performance, $\widehat{A_{i,p}^{\mathrm{RE}_1}} = \hat{Z}_{i,p}^{\mathrm{RE}}$. Alternatively, if desired, we can set λ to 0, such that we can inform the selection of predictors solely based on the reduction in heterogeneity of performance, yielding $\widehat{A_{i,p}^{\mathrm{RE}_0}} = \hat{\tau}_{i,p}$. Finally, we consider the case where heterogeneity and average performance are given equal weighting by setting $\lambda = \frac{1}{2}$, such that $\widehat{A_{i,p}^{\mathrm{RE}_{1/2}}} = \frac{1}{2}\hat{Z}_{i,p}^{\mathrm{RE}} + \frac{1}{2}\hat{\tau}_{i,p}$.

This equation can be seen as an extension of the bias-variance decomposition of the MSE where we now have a summation of squared bias, within-cluster variance and between-cluster variance. If $\boldsymbol{p}$ are considered estimators for $\boldsymbol{y}$, then the MSE for $\boldsymbol{p}$ can be shown to be: $\mathrm{MSE}(\boldsymbol{p}) = \mathrm{var}(\boldsymbol{p}) + \mathrm{Bias}(\boldsymbol{p}, \boldsymbol{y})^2$. As $\hat{\tau}^2$ is the estimate of the between cluster variance of $\mathrm{MSE}(\boldsymbol{p})$, i.e. var_{bs}, the estimator $\widehat{A_{i,p}^{\mathrm{RE}_{1/2}}}$ estimator can be interpreted as the mean of:

$$\frac{1}{2}\lambda\Big(\mathrm{var}(\boldsymbol{p}) + \mathrm{Bias}(\boldsymbol{p}, \boldsymbol{y})^2\Big) + \frac{1}{2}\Big(1-\lambda\Big)\Big(\mathrm{var}_{bc}\big(\mathrm{var}(\boldsymbol{p}) + \mathrm{Bias}(\boldsymbol{p}, \boldsymbol{y})^2\big)\Big) \tag{3.6}$$

Variability of performance across data sets

In the meta-analysis approach, the evidence from small clusters is downweighted to attain an estimate of the mean of the distribution of performance. Yet this distribution might not be of central importance. Instead, all clusters might be considered of equal importance. Then we may instead apply a measure of variability directly to the performance estimates, for instance the standard deviation $\widehat{A_{i,p}^{\mathrm{SD}}} = SD(\hat{Z}_{i,p,1}, ..., \hat{Z}_{i,p,K})$.

Alternatively, if no assumptions can be made on the distribution of the predictive performance statistics, we may apply a non-parametric measure. For example, when

using Gini's Mean Difference we have [247, 248]:

$$\widehat{A^{\text{Gini}}}_{i,p} = \frac{2}{K(K-1)} \sum_{1 \leq h \leq v \leq K} \mid \hat{Z}_{i,p,v} - \hat{Z}_{i,p,h} \mid \tag{3.7}$$

3.4 Motivating example 2: Updating a model for diagnosing DVT

The prediction model developed in section 3.2.3 had a rather heterogeneous performance across validation clusters and was lacking in average discrimination performance. This heterogeneity of performance implies that although the outcome may be predicted well in individuals in some clusters, which may be helpful in diagnosis, it may be unsatisfactory for individuals in other clusters. The heterogeneity across the 11 clusters may be explained by differences in (measured and unmeasured) predictor distributions and true predictor effects. Therefore we here consider whether additional predictors and interaction effects might explain such differences. Whereas individual clusters may lack the sample size to detect nonlinear effects or may lead to highly variable predictor effects, this is more feasible in IPD-MA (and in large healthcare data bases).

Briefly, we considered the following ten additional candidate predictors to extend the model from section 3.2.3: sex, absence of leg trauma, absence of leg trauma x recent surgery (i.e. an interaction effect), vein distension, log of duration of symptoms, age/25 (i.e. divided by 25, to increase the absolute value of its coefficient), age/25 squared, age/25 x malignancy, abnormal d-dimer value and abnormal d-dimer x sex. As we developed this model for illustrative purposes only, we applied single imputation for missing data using a joint model with random effects. [249, 229]

We recommend that the inclusion of each candidate predictor (or transformation thereof) be carefully considered with respect to the improvements in generalizability of the model performance on the one hand, and the cost of measuring the predictor on the other. Here, we apply our methodology to illustrate how each of the strategies regarding heterogeneity of performance leads to different model specifications, and thereby to differing average and heterogeneity of performance. To assess the generalizability of prediction models that use these predictor functions, we follow the SIECV strategy for model development that we developed in section 3.3, apply the MSE (i.e. Brier score) to the predicted probabilities in the hold-out clusters and apply the aggregated loss functions (measures of heterogeneity) on the MSE estimates and standard errors thereof, to select predictors as outlined in section 3.3.2.

The six applied aggregated loss functions lead to models with four different predictor function specifications (Table 3.4), as the strategy that ignored clustering (A^{M}) when estimating generalizability of performance lead to the same model specification as the strategy that optimized the meta-analytic mean of performance (A^{RE_1}), and the meta-analysis strategy that placed equal importance on heterogeneity and average performance ($A^{\text{RE}_{1/2}}$) lead to the same model specification as the A^{SD} strategy. As the SIECV allows for the estimation of any performance statistic, we assessed discriminatory performance with the c-statistic, and calibration with

the calibration slope and intercept, for the final model for each aggregated loss function. Subsequently, we summarized the performance and heterogeneity thereof with univariate random effects meta-analyses.

Table 3.4: Estimated Regression Coefficients of Seven Models for Predicting DVT Estimated with (S)IECV

Predictor	None	A^{M}	A^{RE_1}	$A^{\mathrm{RE}_{1/2}}$	A^{RE_0}	A^{SD}	A^{SD}
Intercept	−2.17	−3.54	−3.54	−5.13	−5.00	−5.13	−3.99
Malignancy	0.98	0.76	0.76	1.64	1.68	1.64	2.05
Calf difference	1.26	1.13	1.13	1.38	1.34	1.38	1.07
Surgery	0.55	−0.04	−0.04	0.25	0.25	0.25	0.34
D-dimer positive		2.76	2.76	2.99	2.94	2.99	2.81
Age/25		−0.22	−0.22				
Vein distension		0.46	0.46				
Surgery x No leg trauma		0.68	0.68				
No leg trauma				0.95	0.96	0.95	
$(\mathrm{Age}/25)^2$				−0.02		−0.02	
Male				0.32	0.36	0.32	0.52
D-dimer positive x Male				−0.20	−0.24	−0.20	−0.28
Malignancy x Age/25				−0.32	−0.35	−0.32	−0.50

None: Model with no predictor selection, A^{M}: Mean performance; A^{RE_1}: Random effects meta-analytic estimate of mean of distribution of performance; A^{RE_0}: Random effects meta-analytic estimate of heterogeneity of distribution of performance; $A^{\mathrm{RE}_{1/2}}$: Sum of random effects meta-analytic estimates of mean and heterogeneity of distribution of performance; A^{SD}: Standard Deviation; A^{Gini}: Gini's mean difference.
Malignancy: history of malignancy, Calf difference: difference in circumference of calves ≥ 3 cm, Surgery: recent surgery, Age/25: Age divided by 25, Duration: duration of symptoms.
Empty cells indicate the predictor was not selected for inclusion in the corresponding model.
Summary predictor effects were estimated by the Dersimonian and Laird method, as REML did not converge for the estimation of some models. Although REML has better theoretical properties for the heterogeneity estimate, the difference for the summary effects (presented here) is limited.

In terms of calibration slopes (also estimated with Firth's correction), all strategies showed some overfit (summary calibration slope < 1), though to varying degrees (Table 3.5, Figure 3.1). Slopes < 1 imply that the estimated slopes were too large (the log odds ratios deviated too far from 0), which yielded predictions for individuals that were too extreme. The linear predictors in the A^{SD} strategy and the meta-analytic strategy that combined heterogeneity and average performance ($A^{\mathrm{RE}_{1/2}}$) were the worst calibrated (calibration slope of 0.85), and the A^{Gini} strategy the best (0.94).

Table 3.5: Meta-Analysis Summary Estimates of SIECV Performance of Six Strategies for Predicting DVT

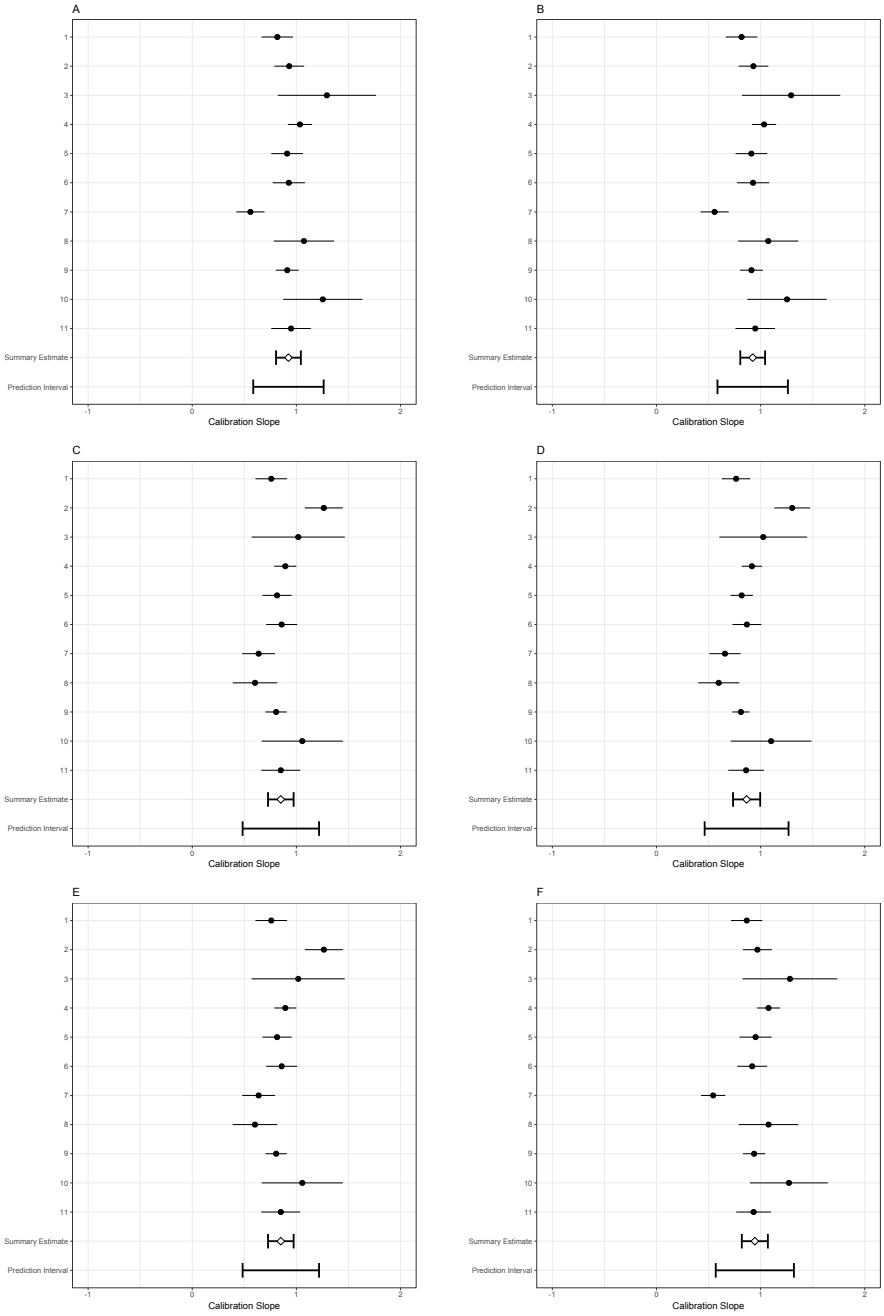
Measure	Strategy	A^{RE_1}	95% CI	95% PI
Calibration slope	None	1.00	0.83 : 1.16	0.53 : 1.46
	A^{M}	0.92	0.80 : 1.04	0.59 : 1.26
	A^{RE_1}	0.92	0.80 : 1.04	0.59 : 1.26
	$A^{\mathrm{RE}_{1/2}}$	0.85	0.73 : 0.97	0.48 : 1.22
	A^{RE_0}	0.87	0.73 : 1.00	0.46 : 1.27
	A^{SD}	0.85	0.73 : 0.97	0.48 : 1.22
	A^{Gini}	0.94	0.82 : 1.07	0.57 : 1.32
Calibration intercept	None	0.03	−0.33 : 0.39	−1.22 : 1.27
	A^{M}	−0.06	−0.46 : 0.35	−1.47 : 1.35
	A^{RE_1}	−0.06	−0.46 : 0.35	−1.47 : 1.35
	$A^{\mathrm{RE}_{1/2}}$	0.16	−0.25 : 0.58	−1.27 : 1.60
	A^{RE_0}	−0.04	−0.47 : 0.39	−1.51 : 1.43
	A^{SD}	0.16	−0.25 : 0.58	−1.27 : 1.60
	A^{Gini}	−0.20	−0.61 : 0.21	−1.63 : 1.23
c-statistic	None	0.68	0.65 : 0.71	0.60 : 0.75
	A^{M}	0.81	0.78 : 0.84	0.70 : 0.89
	A^{RE_1}	0.81	0.78 : 0.84	0.70 : 0.89
	$A^{\mathrm{RE}_{1/2}}$	0.81	0.79 : 0.84	0.73 : 0.88
	A^{RE_0}	0.81	0.79 : 0.84	0.71 : 0.89
	A^{SD}	0.81	0.79 : 0.84	0.73 : 0.88
	A^{Gini}	0.81	0.77 : 0.84	0.68 : 0.90

None: Model with no predictor selection, A^{M}: Mean performance; A^{RE_1}: Random effects meta-analytic estimate of mean of distribution of performance; A^{RE_0}: Random effects meta-analytic estimate of heterogeneity of distribution of performance; $A^{\mathrm{RE}_{1/2}}$: Sum of random effects meta-analytic estimates of mean and heterogeneity of distribution of performance; A^{SD}: Standard Deviation; A^{Gini}: Gini's mean difference. 95% CI: 95% confidence interval; 95% PI: the random effects meta-analysis approximate 95% prediction intervals lower and upper bound.

There was substantial heterogeneity in the estimated calibration slopes, especially for the predefined model with no predictor selection. For all strategies, the prediction interval for the calibration slope also included values > 1, which implies that for some (future) clusters the log odds ratios will probably not deviate from 0 enough and that predictions for individuals will probably be not extreme enough. The heterogeneity of the calibration slope decreased for all strategies, as compared to the predefined model with no added predictors. This means that for the resulting models there was a decreased need for extensive local updating.

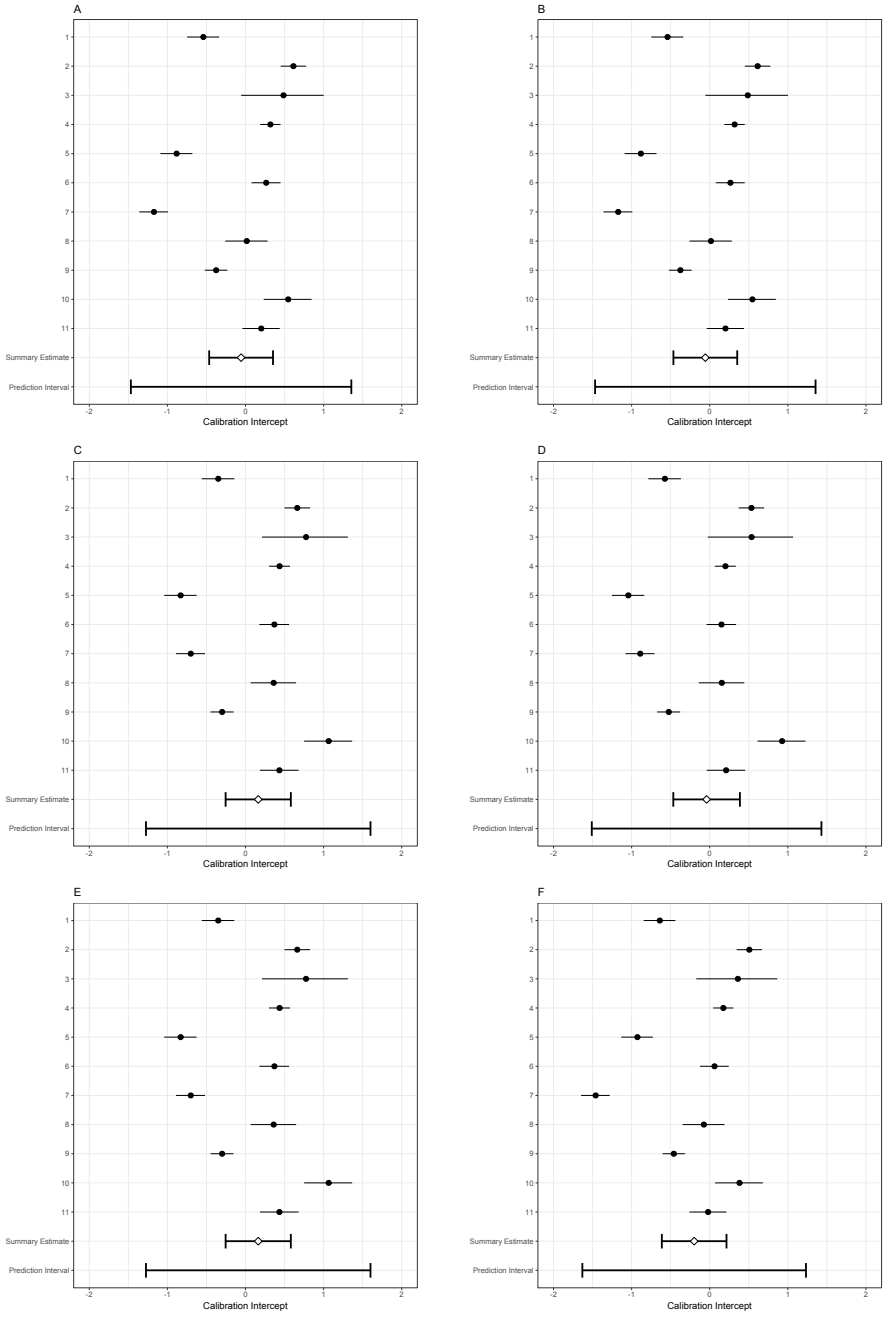
In terms of average calibration intercepts, all strategies achieved a reasonable calibration in the large, that is close to zero (Table 3.5, Figure 3.2). This means that on average the incidence was predicted accurately. On the other hand, the meta-analysis of the calibration intercepts showed that the heterogeneity of calibration in

Figure 3.1: Forest Plots of SIECV Estimates of Calibration Slopes of Six Strategies for Predicting DVT



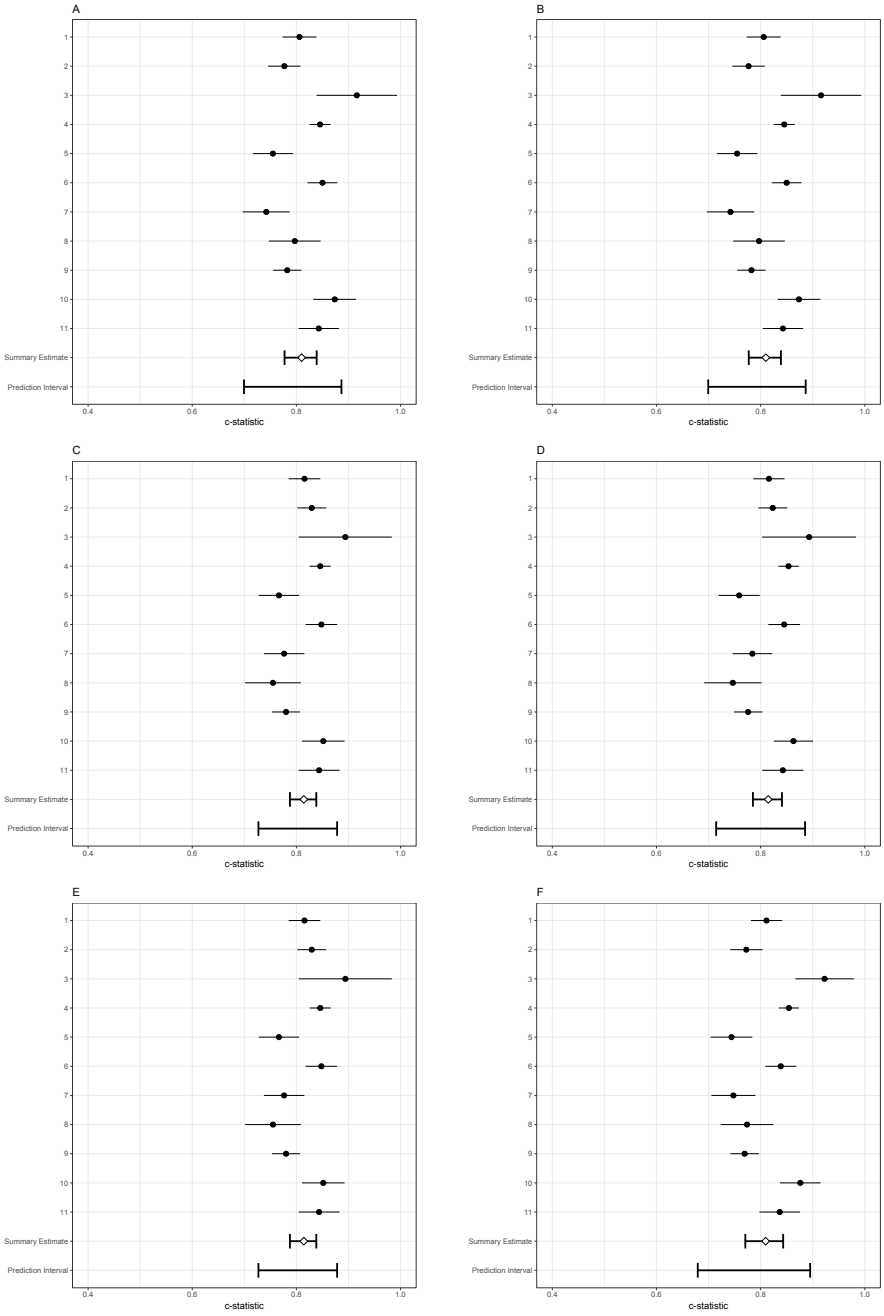
A: Mean performance, A^{M}; B: Random effects meta-analytic estimate of mean of distribution of performance, A^{RE_1}; C: Random effects meta-analytic estimate of heterogeneity of distribution of performance, A^{RE_0}; D: Sum of random effects meta-analytic estimates of mean and heterogeneity of distribution of performance, $A^{\mathrm{RE}_{1/2}}$; E: Standard Deviation, A^{SD}; F: Gini's mean difference, A^{Gini}.

Figure 3.2: Forest Plots of SIECV Estimates of Calibration Intercepts of Six Strategies for Predicting DVT



A: Mean performance, A^{M}; B: Random effects meta-analytic estimate of mean of distribution of performance, A^{RE_1}; C: Random effects meta-analytic estimate of heterogeneity of distribution of performance, A^{RE_0}; D: Sum of random effects meta-analytic estimates of mean and heterogeneity of distribution of performance, $A^{\mathrm{RE}_{1/2}}$; E: Standard Deviation, A^{SD}; F: Gini's mean difference, A^{Gini}.

Figure 3.3: Forest Plots of SIECV Estimates of c-statistics of Six Strategies for Predicting DVT



A: Mean performance, A^{M}; B: Random effects meta-analytic estimate of mean of distribution of performance, A^{RE_1}; C: Random effects meta-analytic estimate of heterogeneity of distribution of performance, A^{RE_0}; D: Sum of random effects meta-analytic estimates of mean and heterogeneity of distribution of performance, $A^{\mathrm{RE}_{1/2}}$; E: Standard Deviation, A^{SD}; F: Gini's mean difference, A^{Gini}.

the large had increased for all modeling strategies, as compared to the predefined model with no added predictors. Hence, a trade-off occurred between (heterogeneity of) calibration slopes and intercepts, and local updating of the intercept will remain necessary.

All strategies that applied SIECV achieved an internally-externally validated c-statistic value of 0.81 (Table 3.5 and Figure 3.3). There were small differences in heterogeneity of discrimination performance. The A^{SD} strategy and the strategies that included the meta-analytic estimate of heterogeneity had smaller values for the heterogeneity of the internally-externally validated c-statistic, than the strategies that focused on the mean performance alone (A^{M} and A^{RE_1}). Further, the A^{Gini} strategy, a non-meta-analytic strategy that focuses on heterogeneity of performance, yielded larger heterogeneity of discrimination performance.

As a final step, one must choose which modelling strategy is most likely to yield adequate performance when applied to individuals in a new cluster, if any. Although heterogeneity in the slopes had decreased substantially for all strategies, the prediction intervals still indicated that updating may be necessary. Further, the models resulting from all strategies are likely to need an intercept update. In terms of calibration, it may therefore not be advisable to develop a global model, that is a model developed on all available clusters (without leaving any out). Finally, although the discrimination for all models improved substantially, the diagnostic utility would have to be put into a clinical perspective.

3.5 Motivating example 3: Predicting atrial fibrillation

Patients with atrial fibrillation (AF) are at an increased risk for stroke. [250] Although stroke is usually not fatal, it often results in neurological deficiencies. [251] In patients with AF the incidence of stroke as well as the incidence of death from stroke can be greatly reduced by oral anticoagulation. [252]

For illustrative purposes, we here consider the development and validation of a binary logistic prediction model to estimate the probability that atrial fibrillation is present in an individual patient. Previously, Audigier et al prepared a simulated dataset to mimic the patients from 28 cohorts (clusters, from hereon) of the GREAT consortium. [253, 254] This dataset comprises a total of 11685 patients of which 3335 have AF.

Because some of the clusters are very small and may thereby cause estimation issues during model development or validation, we removed a total of 8 clusters in which fewer than 50 patients had the outcome or did not have the outcome. Missing values were imputed once using a joint model with random effects. [255, 229] We subsequently modeled the probability of the presence of atrial fibrillation in 10873 patients from the remaining 20 clusters (Table 3.6). To further prevent overfitting, we applied Firth's correction, [99] and re-estimated the intercepts with unpenalized maximum likelihood. [240]

Table 3.6: Clinical Characteristics of AF Data

Outcome: AF		No	Yes	Total
Gender	0	4583 (71.3)	1844 (28.7)	6427
	1	3059 (68.8)	1387 (31.2)	4446
BMI	Mean (SD)	27.3 (5.7)	27.4 (5.8)	27.4 (5.7)
Age	Mean (SD)	67.7 (14.1)	73.1 (13.1)	69.3 (14.0)
SBP	Mean (SD)	135.6 (32.2)	135.3 (32.1)	135.5 (32.2)
DBP	Mean (SD)	78.6 (18.3)	79.1 (18.4)	78.7 (18.3)
HR	Mean (SD)	88.0 (25.0)	96.4 (29.0)	90.5 (26.5)
BNP	Mean (SD)	3.0 (0.9)	2.9 (1.0)	2.9 (0.9)

We considered 7 candidate predictors, consisting of gender (binary) and 6 continuous predictors : body mass index (BMI), age, systolic blood pressure (SBP), diastolic blood pressure (DBP), heart rate (HR) and brain natriuretic peptide (BNP). BMI, age, and HR were divided by 25, and SBP and DBP by 100 to increase the absolute values of their coefficients. For the continuous predictors, we considered linear and quadratic terms and applied centering (within clusters) before application of the quadratic function. This was necessary to ensure that the coefficients are stabilized and positive coefficients for quadratic terms represent an increased probability of presence of AF for values that deviate from the mean value.

Here, we implement the proposed predictor selection procedures to illustrate their impact on average performance as well as on generalizability across the different clusters. We follow the SIECV strategy for model development as described in section 3.3, apply the MSE to the predicted probabilities in the hold-out clusters and apply the aggregated loss functions on the MSE estimates and standard errors thereof, to select predictors functions as we outlined in section 3.3.2.

The six applied aggregated loss functions lead to three different model specifications (Table 3.7), as the strategy that ignores clustering when quantifying generalizability (A^{M}) and the strategy that optimized the meta-analytic mean of performance (A^{RE_1}) lead to the same model specification. Further, both meta-analytic strategies that included heterogeneity of performance ($A^{\mathrm{RE}_{1/2}}$ and A^{RE_0}) lead to the same model, as well as the strategies that directly quantified heterogeneity of performance (A^{SD} and A^{Gini}). Again, we assessed performance with the calibration slope, calibration intercept and c-statistic, and summarized these and the heterogeneity thereof with univariate random effects meta-analyses.

Table 3.7: Estimated Regression Coefficients of Seven Models for Predicting AF Estimated with SIECV

Predictor	A^{M}	A^{RE_1}	$A^{RE_{1/2}}$	A^{RE_0}	A^{SD}	A^{Gini}
Intercept	−0.87	−0.87	−0.84	−0.84	−0.81	−0.81
Gender					−0.08	−0.08
Age/25	0.75	0.75	0.59	0.59	0.62	0.62
$(\text{Age}/25)^2$	−0.17	−0.17				
HR/25	0.29	0.29				
SBP/100	−0.59	−0.59				
$(\text{SBP}/100)^2$	0.28	0.28				
$(\text{BMI}/25)^2$	0.37	0.37	0.32	0.32	0.32	0.32
BNP^2			0.03	0.03	0.03	0.03

A^{M}: Mean performance; A^{RE_1}: Random effects meta-analytic estimate of mean of distribution of performance; A^{RE_0}: Random effects meta-analytic estimate of heterogeneity of distribution of performance; $A^{RE_{1/2}}$: Sum of random effects meta-analytic estimates of mean and heterogeneity of distribution of performance; A^{SD}: Standard Deviation; A^{Gini}: Gini's mean difference.

HR: heart rate, SBP: systolic blood pressure, BMI: Body mass index, BNP: brain natriuretic peptide.

Empty cells indicate the predictor was not selected for inclusion in the corresponding model.

Summary predictor effects were estimated by the Dersimonian and Laird method, as REML did not converge for the estimation of some models. Although REML has better theoretical properties for the heterogeneity estimate, the difference for the summary effects (presented here) is limited.

In terms of summary calibration slopes (estimated with Firth's correction and then pooled in a meta-analysis), all strategies were rather well calibrated, showing only minor overfit (Table 3.8, Figure 3.4). However, there was substantial heterogeneity of the calibration slopes. The approximate 95% prediction interval of the calibration slopes of the models for the A^{M} and A^{RE_1} were the widest, as the upper bound reached 2.20 and the lower bound was estimated at a -0.24. This negative value for the lower bound implies that the predictive effect for the models might be reversed in some clusters: individuals with AF received lower probabilities of AF than individuals without AF in these clusters. This means that for each of these models, there was still a need for extensive updating or model redevelopment.

Table 3.8: Meta-Analysis Summary Estimates of SIECV Performance of Six Strategies for Predicting AF

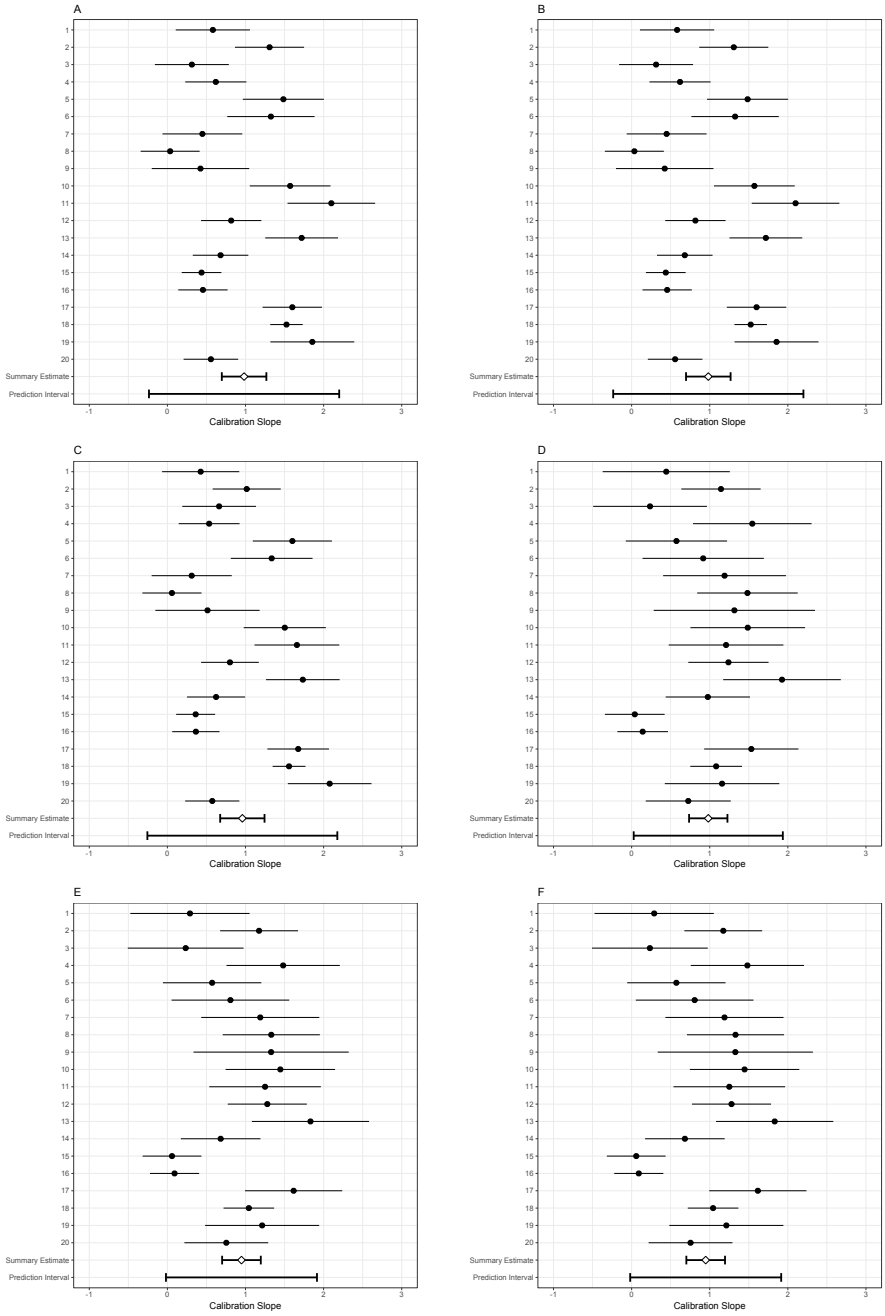
Measure	Strategy	A^{RE_1}	95% CI	95% PI
Calibration slope	A^{M}	0.98	0.70 : 1.27	−0.24 : 2.20
	A^{RE_1}	0.98	0.70 : 1.27	−0.24 : 2.20
	$A^{\mathrm{RE}_{1/2}}$	0.98	0.74 : 1.23	0.03 : 1.94
	A^{RE_0}	0.98	0.74 : 1.23	0.03 : 1.94
	A^{SD}	0.95	0.70 : 1.20	−0.02 : 1.91
	A^{Gini}	0.95	0.70 : 1.20	−0.02 : 1.91
Calibration intercept	A^{M}	0.04	−0.24 : 0.33	−1.25 : 1.34
	A^{RE_1}	0.04	−0.24 : 0.33	−1.25 : 1.34
	$A^{\mathrm{RE}_{1/2}}$	0.00	−0.28 : 0.29	−1.28 : 1.28
	A^{RE_0}	0.00	−0.28 : 0.29	−1.28 : 1.28
	A^{SD}	0.00	−0.28 : 0.29	−1.28 : 1.28
	A^{Gini}	0.00	−0.28 : 0.29	−1.28 : 1.28
c-statistic	A^{M}	0.62	0.58 : 0.65	0.46 : 0.75
	A^{RE_1}	0.62	0.58 : 0.65	0.46 : 0.75
	$A^{\mathrm{RE}_{1/2}}$	0.58	0.56 : 0.60	0.51 : 0.65
	A^{RE_0}	0.58	0.56 : 0.60	0.51 : 0.65
	A^{SD}	0.58	0.56 : 0.60	0.49 : 0.66
	A^{Gini}	0.58	0.56 : 0.60	0.49 : 0.66

A^{M}: Mean performance; A^{RE_1}: Random effects meta-analytic estimate of mean of distribution of performance; A^{RE_0}: Random effects meta-analytic estimate of heterogeneity of distribution of performance; $A^{\mathrm{RE}_{1/2}}$: Sum of random effects meta-analytic estimates of mean and heterogeneity of distribution of performance; A^{SD}: Standard Deviation; A^{Gini}: Gini's mean difference. 95% CI: 95% confidence interval; 95% PI: the random effects meta-analysis approximate 95% prediction intervals lower and upper bound.

In terms of average calibration intercepts, the calibration in the large was (near) perfect (Table 3.8 and Figure 3.5). The $A^{\mathrm{RE}_{1/2}}$, A^{RE_0}, A^{SD} and A^{Gini} strategies all achieved a calibration intercept of 0.00 (95% CI: -0.28 to 0.29), whereas those of A^{M} and A^{RE_1} were hardly different with 0.04 (95% CI: -0.24 to 0.33). Again, there was large heterogeneity in calibration in the large, as shown by the approximate 95% prediction intervals of the calibration intercepts. This means that for each of these models there was still a need for intercept updating they may be used.

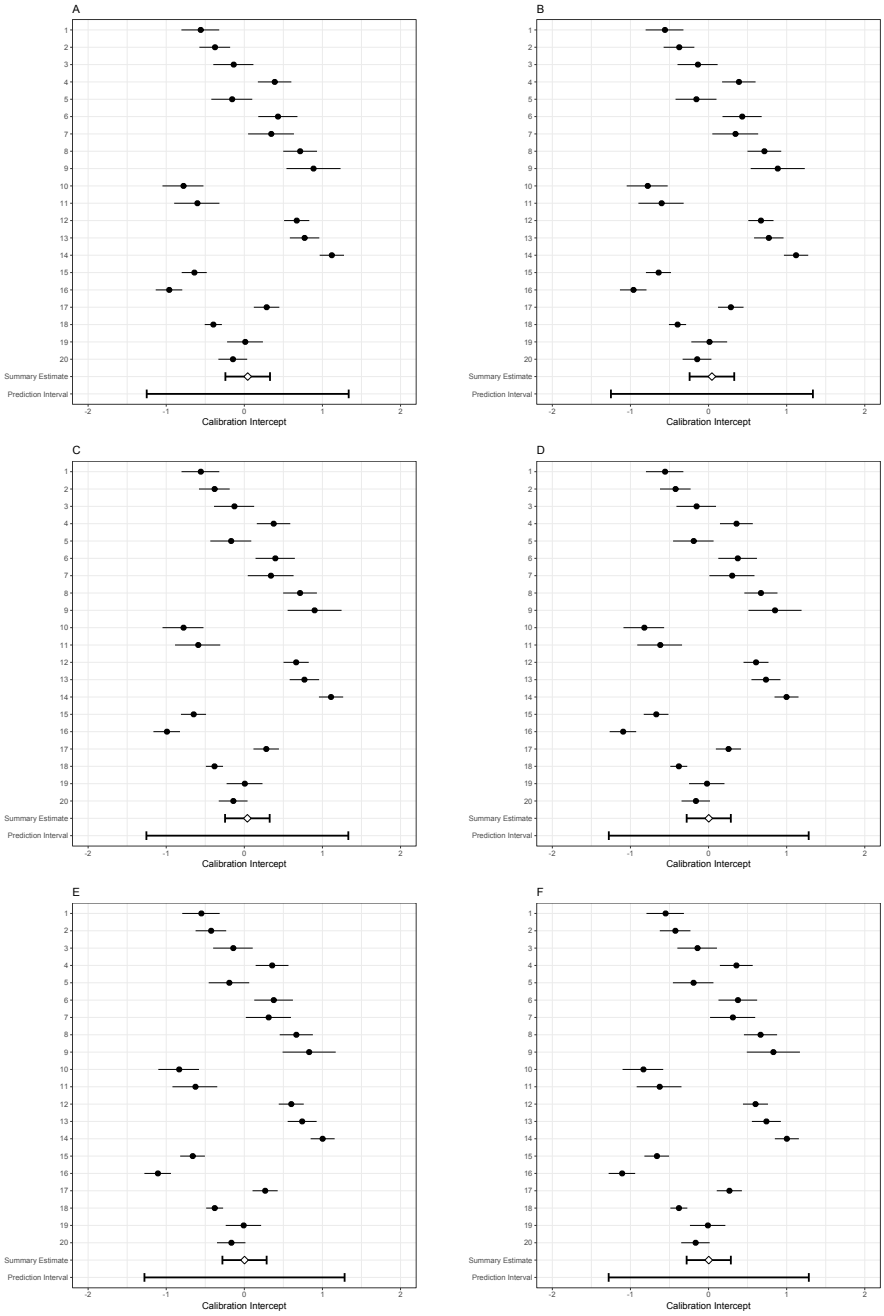
The A^{M} and A^{RE_1} strategies attained a somewhat better discrimination with c-statistics of 0.62 (95% CI: 0.58 to 0.65) than the other strategies, that all attained c-statistics of 0.58 (95% CI: 0.56 to 0.60), respectively (Table 3.8 and Figure 3.6). There was considerable heterogeneity in the c-statistics for all strategies. The discrimination was worse than random (c-statistic $<$ .50) in at least one cluster for each of the strategies. Indeed, the approximate 95% prediction interval shows it is most likely that this will occur in a new cluster for the A^{M} and A^{RE_1} strategies.

Figure 3.4: Forest Plots of SIECV Estimates of Calibration Slopes of Six Strategies for Modeling AF



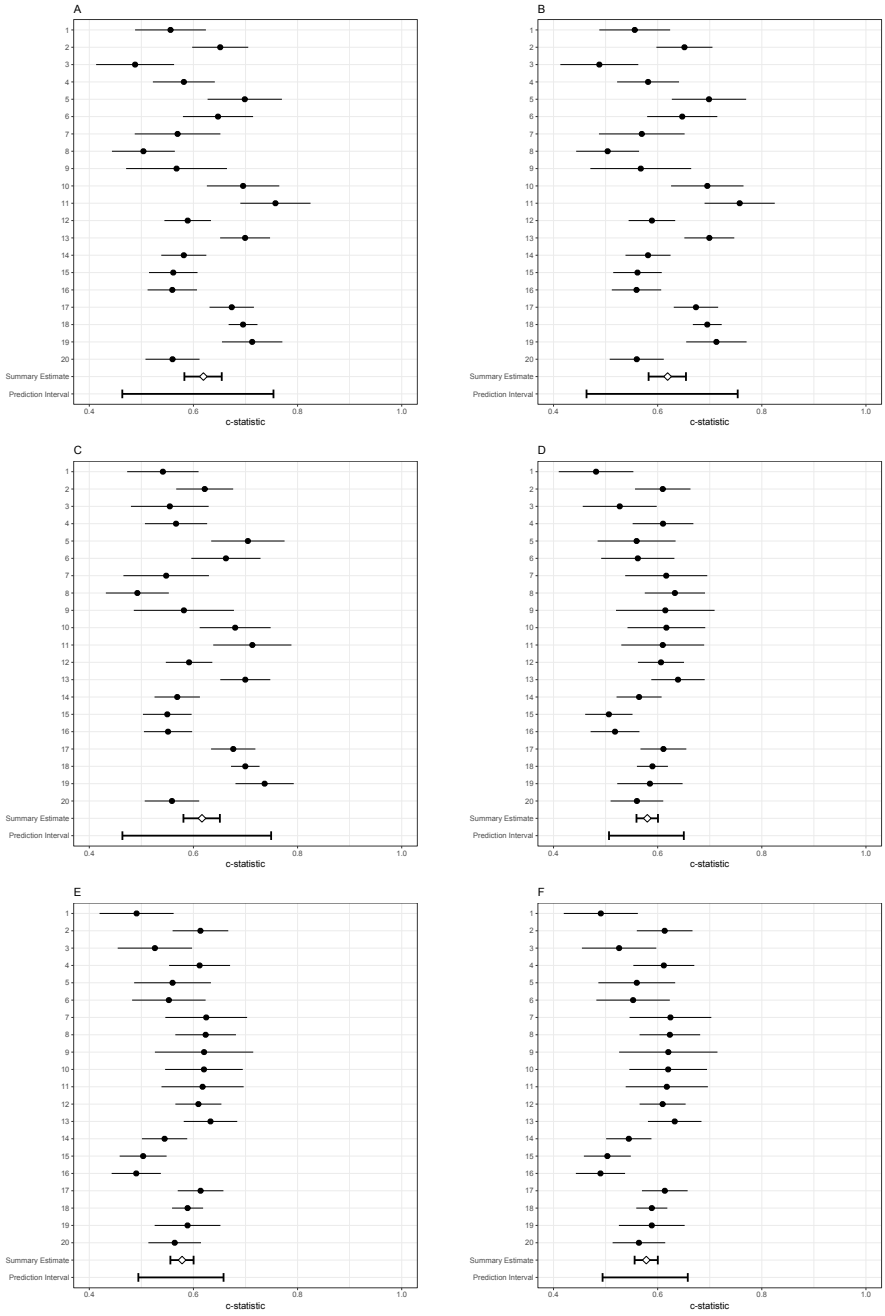
A: Mean performance, A^{M}; B: Random effects meta-analytic estimate of mean of distribution of performance, A^{RE_1}; C: Random effects meta-analytic estimate of heterogeneity of distribution of performance, A^{RE_0}; D: Sum of random effects meta-analytic estimates of mean and heterogeneity of distribution of performance, $A^{\mathrm{RE}_{1/2}}$; E: Standard Deviation, A^{SD}; F: Gini's mean difference, A^{Gini}.

Figure 3.5: Forest Plots of SIECV Estimates of Calibration Intercepts of Six Strategies for Modeling AF



A: Mean performance, A^{M}; B: Random effects meta-analytic estimate of mean of distribution of performance, A^{RE_1}; C: Random effects meta-analytic estimate of heterogeneity of distribution of performance, A^{RE_0}; D: Sum of random effects meta-analytic estimates of mean and heterogeneity of distribution of performance, $A^{\mathrm{RE}_{1/2}}$; E: Standard Deviation, A^{SD}; F: Gini's mean difference, A^{Gini}.

Figure 3.6: Forest Plots of SIECV Estimates of c-statistics of Six Strategies for Modeling AF



A: Mean performance, A^{M}; B: Random effects meta-analytic estimate of mean of distribution of performance, A^{RE_1}; C: Random effects meta-analytic estimate of heterogeneity of distribution of performance, A^{RE_0}; D: Sum of random effects meta-analytic estimates of mean and heterogeneity of distribution of performance, $A^{\mathrm{RE}_{1/2}}$; E: Standard Deviation, A^{SD}; F: Gini's mean difference, A^{Gini}.

For all of the considered strategies the SIECV shows that although calibration was adequate on average and that overall discrimination was modest, it is not very likely that any of the developed models will perform well at predicting AF in new clusters without local updating. The effects of the included predictors vary substantially across clusters and heterogeneity could not be resolved by considering non-linear terms.

3.6 Discussion

We proposed a methodology for improving the generalizabilty of prediction models developed across different settings and populations when data from IPD MA or electronic health record data sets are available. This methodology leverages the information from multiple clusters by iteratively using all but one cluster for model development and assessing performance in the remainder.

The overall predictive performance and its variation across clusters can then be used to quantify a model's generalizability across clusters, and to inform the selection of predictors. As we have demonstrated in our motivating examples, selection of predictors based on the proposed aggregated loss functions can lead to differing model specifications. Each model specification may perform differently in predicting the outcome for individuals in differing clusters, leading to differing average and heterogeneity of calibration and discrimination. Trade-offs may occur between discrimination and calibration as well as between average and heterogeneity of performance. These may be quantified in the SIECV algorithm during model development, by which the need for extensive or local model updating may be assessed immediately. For instance, before validation in the hold-out cluster, the intercept may be updated, which will inform the researcher on the generalizability of performance of the model after local updates.

Although it remains unlikely that SIECV can completely resolve the need for an intercept update, it may reduce the necessity of the local tailoring of prediction models (particularly with respect to the calibration slope), which is highly prevalent in the implementation of today's prediction models. In particular, we aimed to reduce the need for re-estimation of individual predictor effects. Evidently, the potential impact of SIECV strongly depends on the availability of patient-level covariates that may explain heterogeneity in predictive associations.

We discussed a variety of aggregated loss functions for quantifying the generalizability of a developed model. These ranged from producing an average performance estimate to quantifying its dispersion across clusters. Our framework also allows for assessing average and heterogeneity of performance across data sets (clusters) through meta-analysis, and formalizes the predictor function selection by both of these simultaneously. This requires specifying the relative importance of average performance and the heterogeneity thereof on beforehand. For the meta-analysis strategy, this either requires the prespecification of λ, or the finding of an optimal value for λ through a resampling method. When applied in the SIECV model development process, these measures lead to different model specifications with different average performance and heterogeneity, as each places different importance thereon. Therefore, the researcher will have to choose whether to optimize average calibration

(intercept close to 0, slope close to 1) and discrimination (high c-statistic, positive and/or negative predictive value for its clinical issue), or the heterogeneity of one or both of these across clusters.

3.6.1 Limitations and future directions

Our main focus was on informing the selection of predictors in an IPD-MA or the analysis of EHR, and not on the estimation of corresponding predictor effects. Hence, the impact of our methodology may be limited in cases where the number of available predictors is low or when available predictors are not related to heterogeneity in predictor-outcome associations. Though, even in cases where we have few predictors, it may still prove worthwhile to consider non-linear terms and interaction effects by the use of SIECV.

Improvements in one prediction model performance measure may come at a cost of those in another measure. For instance, an improvement in calibration may result in a deterioration in discrimination. In our motivating example, the predictor D-Dimer had a large predictive effect, and substantially contributed to discrimination performance. However, D-Dimer was a dichotomized predictor and was measured using different methods across clusters. [237, 256] This resulted in substantial heterogeneity in its regression coefficients and in heterogeneity of its diagnostic accuracy (see [257]).

In this manuscript, we have focused on improving the generalizability of a model's predictive performance through the addition or removal of predictors or functions of predictors. However, this may lead to instability in the estimation process and overfitting of the predictor coefficients if not enough data are available or the outcome is too rare. A future step may be to incorporate heterogeneity of performance into the estimation process. For instance, in penalized Maximum Likelihood estimation a penalty for heterogeneity of predictor coefficients or predictive performance across clusters could be applied. Such a penalty could readily produce more generalizable models, without the need for a stepwise selection process.

We have not performed a simulation study to assess the proposed methods. Such an endeavour might not be fruitful, as it should be obvious that each of the measures for predictor selection would lead to different estimates of both average and heterogeneity of performance, i.e. there is no gold standard. As each of the proposed methods serve a different goal, it is unclear what the evaluation criteria in such a simulation study would be. Nevertheless, it may still be helpful to assess under what circumstances certain generalizability measures may be preferred.

Whereas methodology has been developed on dealing with missing data in model validation in general (e.g. see [258]), it remains unclear how this would affect SIECV. In general, the use of multilevel imputation methods have been recommended in IPD-MA and other types of clustered data. [94, 195, 253, 96] These methods can account for variables that are sporadically missing within one or more clusters, but also impute variables that are not measured in certain clusters. [195] The adoption of multilevel imputation methods therefore seems paramount when adopting SIECV in datasets with missing values. Further, in multiple imputation it is imperative that the imputation model includes the all of the variables used in the analysis model, including the outcome, as well as any other predictive variables. [259] For

this reason, we recommend to include all candidate predictors of SIECV in the imputation model, and to allow for random effects for all of these.

3.6.2 Conclusion

The IECV methodology for model validation can inform the researcher on the need for updating a prediction model to adapt it to particular populations and settings, when IPD from multiple clusters or studies are available. Our SIECV methodology extends this framework to quantify and reduce the impact of any heterogeneity on prediction model performance. This can inform the researcher and may reduce the need for tailoring the prediction model to specific populations and settings.

Acknowledgements

We gratefully acknowledge the following authors for sharing of individual participant data from the deep vein thrombosis (DVT) studies: G.J. Geersing, N.P.A. Zuithoff, C. Kearon, D.R. Anderson, A.J. ten Cate-Hoek, J.L. Elf, S.M. Bates, A.W. Hoes, R.A. Kraaijenhagen, R. Oudega, R.E.G. Schutgens, S.M. Stevens, S.C. Woller, P.S. Wells and K.G.M. Moons. The DVT data that support the findings of this study are not publicly available, according to the conditions determined by the authors of the DVT studies, but are available on request from G.J. Geersing, by e-mailing G.J.Geersing@umcutrecht.nl. The atrial fibrillation data that support the findings of this study are openly available in The Comprehensive R Archive Network at https://cran.r-project.org/, [229] in R package `micemd`. [196]

This work is financially supported by the Netherlands Organization for Health Research and Development grant 91617050 for TD and grant 91810615 for VJ and KM, and the European Union's Horizon 2020 Research and Innovation Programme under ReCoDID Grant Agreement no. 825746 for VJ and KM.

Chapter 4

Propensity-based standardization to enhance the interpretation of predictive performance in external validation studies

Valentijn M.T. de Jong, Jeroen Hoogland, Thomas P.A. Debray, Tri-Long Nguyen

Abstract
Many prediction models perform worse when applied to new individuals, which may be due to a (lack of) representativeness of the validation sample for the prediction model. If the validation sample does not fully represent the model's intended target population, estimates of model performance in the validation set are misleading.

We consider the use of propensity score weighting methods to standardize predictive performance measures estimated in multiple validation samples that are obtained from different but related populations and settings, by weighting with respect to the covariate distribution of the target population and setting. We show how standardized measures for a model's discrimination and calibration can be derived. We illustrate our methods in a motivating example on the validation of eight different diagnostic prediction models for the detectiong of deep vein thrombosis (DVT) that may aid in the diagnosis of patients suspected of DVT in 12 external validation data sets. We applied random effects meta-analysis to analyze the estimates of prediction models' performance across these 12 external validation data sets.

The summary meta-analysis estimates of standardized and unstandardized discrimination performance indicate that, on average, discrimination was not substantially affected by differences in case-mix between the development and validation samples. The between-study heterogeneity estimates, however, indicate that differences between discriminatory performance in the individual validation studies can partially be attributed to differences in case-mix, rather than the use of invalid model coefficients. Further, the meta-analysis showed that the between-study heterogeneity for the calibration slopes was increased by standardization for all models. This demonstrates that there were differences in case-mix between the development and validation samples, and that the case-mix differences partially masked the differences in optimal coefficients between these samples. When standardization filtered out these differences in case-mix, heterogeneity in the calibration slopes became more apparent.

Propensity score-based standardization may help to facilitate the interpretation of (heterogeneity in) prediction model performance across multiple external validation studies and to guide model updating strategies or to accept that the validation sample does not reflect the target population of the developed model.

4.1 Introduction

Prediction models provide estimates of absolute risk that a particular health status is present (diagnosis) or will occur in the future (prognosis). The development of prediction models has seen a rapid growth in medicine. Unfortunately, many prediction models perform worse when tested in or applied to new individuals. [11, 15, 13, 260] Common reasons for inaccurate predictions are the violation of model assumptions, omission of important predictors, poor handling of missing data in the development or validation data and, in particular, overfitting of the developed model. [261, 242, 241, 243] However, the validity of model predictions may also be affected by the (lack of) representativeness of the validation sample in view of the development sample, differences in predictor and outcome definitions and measurements, and the presence of measurement error both. [262, 263, 16] Also, performance measures may provide misleading results as they are sensitive to variation in case-mix (i.e. being sensitive to differences in covariate distributions) between development and validation samples. [12] The latter implies that the measured deterioration in the prediction model performance measure should be attributed to the choice of validation sample and measure. Hence, model revision efforts can be rather futile.

It is widely advocated that when researchers develop a new prediction model, they explore whether its predictions are sufficiently accurate across different settings and (sub)populations (i.e. different validation samples). [20, 5] The model's predictive performance is then assessed in new samples that have not been used during its development (so-called external validation). [22, 3, 264, 9, 262, 263] To facilitate investigation of model performance, researchers developing a novel prediction model ideally collect individual participant data (IPD) sets from two or more settings, institutes or even predesigned studies. One is then used for model development and the other for external validation. Results from the validation study are then used to confirm whether the model is adequate or to recommend certain revisions prior to its implementation in practice. It may be clear that choosing which data set should be used for validation is not a trivial task. Arguments for choosing one or another may focus on sample size, availability of predictors and outcome, or representativeness of the study population. If the validation sample does not fully represent the target population, estimates of model performance may be misleading.

When pursuing an external validation study, changes in model performance with respect to the development study should be interpreted with caution. Decline in a prediction model performance measure does not necessarily imply that the model coefficients (e.g. predictor weights) are invalid. Likewise, adequate performance upon external validation does not necessarily imply that the model transports well to different settings and populations, as this requires some degree of consistent model performance across multiple validation samples with different case-mix. [263, 262] To disentangle the possible sources of variability in prediction model performance across multiple validation studies, it has been recommended to quantify the relatedness between the development and validation samples. [15] This allows for the isolation of changes in performance that can only be attributed to the use of invalid regression coefficients and thus to establish which type of model revisions may be

necessary.

Previously, Debray et al. proposed to develop a so-called membership model, [15] which calculates the probability that an individual belongs to a certain (development or validation) sample. The concordance index of this model then indicates their relatedness and can be used to identify whether the evaluation of a particular model's performance is likely to be affected by case-mix differences. [13, 15] In this article, we build on their framework and consider the use of propensity score weighting methods to standardize prediction model performance measures estimated in multiple samples validation that are obtained from different but related settings and populations, by weighting with respect to the covariate distributions. [265, 266]

Such standardization may help during the external validation of an existing model to improve the interpretation of performance differences with respect to the development sample and to identify the usefulness of specific model updating or revision strategies. This usage of propensity-based standardization can be conducted when individual participant data (IPD) from multiple samples are readily available, which appears particularly useful in an IPD meta-analysis or large electronic healthcare database context.

To our best knowledge, propensity scores are not yet used to assist external validation of clinical prediction models. This article explores the untapped value of propensity-based standardization in clinical prediction model studies that are based on large data sets with IPD from multiple studies or sources. In section 4.2 we present propensity-based standardization methods in the context of clinical prediction models as well as propensity-standardized validation measures, in section 4.3 we describe a motivating example with illustrative data on the diagnosis of DVT and finally section 4.4 provides a discussion of our results.

4.2 Propensity score standardization and clinical prediction models

Propensity score methods were initially proposed for the estimation of causal (e.g. treatment) effects in non-randomized data. [267] Clinical prediction models typically do not aim to provide a causal explanation [268, 269] and therefore do not (strictly) require the incorporation of treatment propensity scores. [270] Although it is possible to account for received treatments during the development and validation of prediction models, [271, 272] we propose a different use of propensity score methods. In particular, when IPD from multiple settings or populations (and thus at least one development and one validation sample) are available to the researcher, one can estimate the probability that a certain individual is a member of a certain sample. These propensities can then be used to standardize the available samples with respect to a specific target population. The advantage of this approach is that it facilitates interpretation of a particular model's performance estimates across different validation samples.

4.2.1 Standardization to membership propensity scores

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

$$m_{S_i}(j) = \Pr\left(S_i = j | X_i, Y_i\right) \tag{4.1}$$

where S_i is a random variable denoting the study sample of individual i, X_i contains the individual's predictor values, and Y_i the observed outcome. In an IPD-MA of J study samples, we have $S_i \in (1, 2, \ldots, J)$. Additionally, let s_i denote the study to which individual i actually belongs, such that the propensity score $m_{S_i}(S_i = s_i)$ quantifies the conditional probability of individual i being member of its originating sample. Further, in all cases, by definition, $\sum_j m_{S_i}(j) = 1$, for each i. As a result, the propensity score $m_{S_i}(j)$ can be estimated by a (multinomial) logistic regression model. The propensity score $m_{s_i}(j)$ can subsequently be used to construct the standardization weight with respect to sample j:

$$w_i(j, S_i) = \frac{m_{S_i}(j)}{m_{S_i}(S_i)} \tag{4.2}$$

For instance, consider we have 2 study samples: one validation sample obtained from a randomized trial (sample a) and one development obtained from an observational study (sample b). Although both samples may contain individuals from the model's target population, it may occur that the inclusion criteria for trial participants are too restrictive and therefore do not fully capture the (case-mix) diversity of the target population as reflected by the original development sample. Estimates of prediction model performance from sample a may therefore be misleading if no account is made for the case-mix differences with respect to sample b. For this reason, we can standardize individuals from sample a with respect to sample b, such that the weighted sample a better represents the target population. For individuals from sample a, we then assign the following weights:

$$w_i(b, a_i) = \frac{m_{a_i}(b)}{m_{a_i}(a_i)} \tag{4.3}$$

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

$$w_i(b, b_i) = \frac{m_{b_i}(b)}{m_{b_i}(b_i)} = 1 \tag{4.4}$$

These weights are derived from the standardization weighting methods described in the causal inference literature, commonly referred to as 'inverse probability weighting' or 'standardized mortality ratio' weighting. [265, 266] (In inverse probability weighting, the numerator is the propensity of belonging to the entire population, including all samples; that is, $\sum_j m_{S_i}(j)$ instead of $m_{S_i}(j)$.)

4.2.2 Validation of prediction models in standardized samples

Propensity score methods can then be used to standardize the predictive performance estimates in external validation samples with respect to the original devel-

opment sample. By removing the difference in case-mix, this approach may help to interpret performance estimates of prediction models in external validation studies with respect to the original development sample. Prediction model performance measure differences are then adjusted for differences in case-mix, which may help to identify causes of poor transportability that cannot directly be attributed to case-mix differences(e.g. invalid model coefficients). For instance, when data from an RCT are available for validating an existing model that was developed in a data from an observational study, it may be more difficult to discriminate between trial patients with and without the outcome due to the stricter inclusion criteria and thus reduced case-mix variability. [14, 273] The estimated discriminative performance in the RCT data would then be a biased estimate of the discriminative performance of the model in the model's intended target population. Propensity score methods may help to appreciate and even alleviate this issue by standardizing the validation samples according to the case-mix distribution of the development sample.

Below, we describe how measures of prediction model performance can be standardized with respect to differences in case-mix between samples. For all performance measures, we use the original development sample as target of standardization, such that any performance differences between the development and validation sample can be interpreted as a consequence of invalid model parameters in the latter (and therefore a lack of model transportability).

Standardized calibration-in-the-large

Calibration-in-the-large is preferably assessed with the calibration intercept, which can be estimated by fitting a logistic regression model (in case of binary outcome) to the validation sample and fixing the coefficient for the linear predictor of the prediction model under study at 1. [244] The resulting calibration intercept then corresponds to an overall correction that is to be applied to the prediction model. The standardized calibration intercept can be estimated based on weighted logistic regression. When externally validating a model previously developed on sample d in external validation sample v_i, the weights are then given by $w_i = m_{v_i}(d)/m_{v_i}(v_i)$.

Standardized calibration slope

Calibration of the linear predictor is preferably assessed with the calibration slope. It can be estimated by fitting a logistic regression model to the validation sample, where a single coefficient is estimated for the linear predictor of the prediction model. The resulting calibration slope then corresponds to a correction factor for the predictor coefficients of the prediction model. Similar to the standardized calibration intercept, the standardized calibration slope can be estimated by first weighting the observations according to the aforementioned weights w_i.

Standardized concordance statistic

Discrimination can be assessed with the concordance (c)-statistic (AUC). For a randomly selected patient $i, i \in (1, \ldots, N_+)$, with the outcome and a randomly selected patient $q, q \in (1, \ldots, N_-)$, without the outcome, the c-statistic estimates

the probability that patient i has the highest predicted probability p_i of the outcome. The c-statistic can be described as:

$$c = \frac{1}{N_+ N_-} \sum_{i=1}^{N_+} \sum_{q=1}^{N_-} I(p_i > p_q), \tag{4.5}$$

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

We propose to apply a weighting procedure to the c-statistic, similar to precedents. [274, 275] We propose to define weights of concordant pairs according to the propensity scores of the pairs. Assuming independence between members of a same pair, the propensity score of a pair is equal to the product of the propensity scores of the members of the pair. Accordingly, the weight of the pair is equal to the product of the weights of the members of the pair. Then, the standardized c-statistic is given by:

$$c_s = \frac{1}{N_+ N_-} \frac{1}{W} \sum_{i=1}^{N_+} \sum_{q=1}^{N_-} I(p_i > p_q) w_i w_q, \tag{4.6}$$

where $\sum_{i=1}^{N_+} \sum_{q=1}^{N_-} w_i w_q = W$ denotes the sum of all weights such that the standardized c-statistic is bounded from 0 to 1.

Alternatively, the standardized c-statistic may be obtained by the bootstrap. The individuals of the validation sample are then sampled with replacement with probability equal to their respective weights (rescaled to range from 0 to 1) and the (unstandardized) c-statistic is estimated on the resulting sample. The center and percentiles of the resulting propensity weighted distribution of c-statistics then estimate the standardized c-statistic and its confidence interval, respectively, similar to the percentile method for the bootstrap estimation of the unstandardized confidence interval. [276]

In the next section we present a motivating example, in which we estimate these standardized predictive performance metrics for an existing model at multiple external validations.

4.3 Motivating example: external validation of previously developed model

Deep vein thrombosis (DVT) increases a patient's risk of post-thrombotic syndrome and pulmonary embolism, which can be fatal. [236] In DVT suspected patients, often no DVT is present on advanced reference testing. [237] We here consider for illustrative purposes the development and validation of 8 different prediction models that could help in the diagnosis of DVT in patients that are suspected of having DVT and use the IPD of 10002 patients, of which 1864 have DVT, from thirteen different cross-sectional diagnostic studies across multiple countries. [238]

4.3.1 Methods

The data from one study (the development study, Table 4.1) were used to develop eight logistic regression prediction models for the probability that DVT is present. Coefficients for eight prespecified predictors were estimated: positive d-dimer test, calf difference > 3cm, oral contraceptive usage, male sex, no presence of leg trauma, vein distension, active malignancy, and recent surgery (Table 4.2). Prediction model 1 consisted of only the first predictor and prediction model 2 consisted of the first two, etc. Our aim is to investigate to what extent these 8 models generalize well across the 12 remaining validation samples and to what extent variability in their performance can be attributed to lack of transportability or rather to case-mix heterogeneity.

Table 4.1: Clinical characteristics of development data for a model for diagnosing DVT

Variable	Value	Count
Sex	female	828
	male	467
Oral contraceptive (OC)	not used	1167
	used	128
Presence of malignancy	no active malignancy	1214
	active malignancy	81
Recent surgery	no recent surgery (or bedridden)	1114
	recent surgery (or bedridden)	181
Absence of leg trauma	leg trauma present	197
	no leg trauma present	1098
Vein distension	no vein distension	1038
	vein distension	257
Calf difference	calf difference < 3 cm	739
	calf difference > 3 cm	556
D-dimer abnormal	D-dimer negative	398
	D-dimer positive	897
DVT	no DVT	1006
	DVT	289

We externally validated the 8 prediction models in the remaining 12 studies. We estimated for each prediction model the traditional unstandardized c-statistic, calibration slope and intercept as well as the standardized measures described in Section 4.2.2 in each external validation study, to disentangle invalid coefficients from differences in case-mix as causes of heterogeneity in prediction model performance.

Table 4.2: Coefficients of eight prediction models for diagnosing DVT

Model	Intercept	D-dimer	Calf	OC	Male	No trauma	Vein	Malig.	Surg.
1	-3.39	2.58							
2	-4.95	2.42	1.11						
3	-5.04	2.44	1.13	0.40					
4	-6.12	2.46	1.15	0.72	0.72				
5	-6.77	2.49	1.17	0.72	0.73	0.68			
6	-7.35	2.47	1.16	0.70	0.72	0.66	0.52		
7	-7.82	2.44	1.14	0.72	0.70	0.64	0.52	0.53	
8	-8.33	2.43	1.15	0.76	0.71	0.67	0.53	0.50	0.42

Empty cells indicate the coefficient for the respective predictor is assumed zero.
For predictor definitions see Table 4.1.

We then applied random-effects meta-analysis to summarize these models' estimated performance measures across the 12 validation studies and to investigate their generalizability across the different settings and populations. [21, 234] The meta-analyses were performed with REML and 95% confidence intervals were estimated by the method of Knapp and Hartung. [107] Approximate 95% prediction intervals were constructed with the t-distribution with 10 degrees of freedom. [110] The confidence intervals for the propensity-weighted c-statistic were obtained with 5000 resamples with replacement, with probability equal to the weights as defined in Section 4.2.

4.3.2 Results

Discrimination performance

The meta-analysis summary estimates indicate that, as expected, discrimination performance greatly improved as more predictors were added to the models. In particular, the pooled c-statistic for the prediction model only adjusting for d-dimer results was 0.67 (95% CI from 0.63 to 0.71), whereas the prediction model with 8 predictors yielded a pooled c-statistic of 0.77 (95% CI from 0.74 to 0.80). Further, we found that summary estimates for the c-statistic that were obtained via propensity standardization did not much differ from the crude (i.e. non-standardized) summary estimates (Table 4.3). This implies that on average, the discrimination of the developed prediction models is no different in the target population as compared to non-target populations. In other words, on average it is not affected by case-mix differences.

In terms of between-study heterogeneity, however, we observed substantial differences. For instance, for prediction model 1 (which only accounts for D-dimer results) the heterogeneity estimate τ for the unstandardized (logit) c-statistic was 0.30. The approximate 95% PI for the pooled c-statistic ranged from 0.51 to 0.80. These results appear to suggest that predictions from model 1 have limited transportability across the included validation studies and that the model may require local updating. However, when standardizing the validation studies, the heterogeneity estimate τ for model 1 decreased to 0.11 and the 95% PI become much

more narrow (0.59 to 0.71). These additional results reveal that between-study heterogeneity in the discriminatory performance of prediction model 1 can partially be attributed to differences in case-mix, rather than the use of invalid model coefficients (i.e. predictor weights).

The difference in τ values between the unstandardized and standardized c-statistic declines as the number of predictors is increased. For instance, for prediction model 8, we found $\hat{\tau} = 0.27$ for both the standardized and unstandardized c-statistic. Hence, it appears that models with more predictors are less sensitive to heterogeneity in case-mix and their variation in discrimination performance can mostly be attributed to the use of invalid model coefficients. This implies that these prediction models with more predictors may benefit from the re-estimation of regression coefficients, to improve predictions in local settings and populations.

Calibration performance (calibration slope)

Standardization increased the summary calibration slope from 1.12 (unstandardized) to 1.18 (standardized) for prediction model 1, which also indicates there was a case-mix difference between the development and validation samples. The value greater than 1 for the standardized slope indicates that on average larger prediction coefficients were necessary in the validation data. This finding became more apparent after standardisation, indicating that case-mix differences partially masked the slope differences in the unstandardized evaluation. Overall, the unstandardized and standardized summary calibration slopes approached 1 as the number of predictors increased. This implies that the added predictors accounted for the differences in coefficients between the development and validation samples. Though, the traditional calibration slope was closer to 1 than the standardized calibration slope, meaning that the differences in predictor coefficients and case-mix had counteracted each other.

There was greater heterogeneity across validation studies in the standardized calibration slopes than in the unstandardized ones. For instance, the unstandardized and standardized estimates of τ for the calibration slope for prediction model 8 were 0.06 and 0.21, respectively. Again, this indicates that there were differences in case-mix between the development and validation samples, and that the case-mix differences partially masked the differences in optimal coefficients between the development and validation samples. When the differences in case-mix were filtered out by means of standardization, heterogeneity in the calibration slopes became more apparent.

Calibration performance (Calibration-in-the-large)

Finally, the models were slightly miscalibrated-in-the-large. For instance, for prediction model 1 the summary calibration intercept equaled -0.30 and this was reduced (in absolute terms) to -0.08 for prediction model 8. Standardization reduced the calibration intercept to nearly zero, which means that poor summary calibration-in-the-large can entirely be attributed to case-mix differences. This was expected, since calibration-in-the-large mostly captures case-mix differences in outcome prevalence.

Table 4.3: Unstandardized and propensity-standardized random effects meta-analysis estimates of performance of eight prediction models in 12 external validation studies

Measure	Standardized	Statistic	1	2	3	4	5	6	7	8
c	No	Est	0.67	0.74	0.74	0.75	0.75	0.75	0.76	0.77
		CI	0.63 : 0.71	0.70 : 0.77	0.71 : 0.77	0.72 : 0.78	0.71 : 0.79	0.72 : 0.79	0.73 : 0.79	0.74 : 0.80
		PI	0.51 : 0.80	0.59 : 0.84	0.61 : 0.84	0.63 : 0.85	0.60 : 0.86	0.61 : 0.86	0.64 : 0.86	0.66 : 0.85
		τ	0.30	0.29	0.26	0.25	0.30	0.29	0.26	0.24
	Yes	Est	0.65	0.73	0.74	0.75	0.74	0.75	0.76	0.77
		CI	0.63 : 0.67	0.70 : 0.76	0.71 : 0.77	0.72 : 0.77	0.71 : 0.77	0.72 : 0.78	0.73 : 0.79	0.74 : 0.80
		PI	0.59 : 0.71	0.61 : 0.82	0.63 : 0.83	0.63 : 0.83	0.62 : 0.83	0.62 : 0.85	0.64 : 0.85	0.64 : 0.86
		τ	0.11	0.23	0.22	0.23	0.24	0.26	0.25	0.27
Intercept	No	Est	−0.30	−0.15	−0.13	−0.16	−0.13	−0.08	−0.09	−0.08
		CI	−0.78 : 0.18	−0.60 : 0.30	−0.59 : 0.33	−0.63 : 0.31	−0.61 : 0.35	−0.55 : 0.40	−0.55 : 0.37	−0.54 : 0.38
		PI	−2.03 : 1.42	−1.76 : 1.47	−1.79 : 1.52	−1.86 : 1.54	−1.85 : 1.59	−1.78 : 1.62	−1.76 : 1.57	−1.73 : 1.57
		τ	0.74	0.70	0.71	0.73	0.74	0.73	0.72	0.71
	Yes	Est	0.01	0.02	0.01	−0.01	−0.01	0.00	−0.01	−0.01
		CI	−0.15 : 0.16	−0.13 : 0.16	−0.13 : 0.16	−0.17 : 0.15	−0.18 : 0.15	−0.17 : 0.16	−0.17 : 0.16	−0.18 : 0.15
		PI	−0.52 : 0.53	−0.45 : 0.49	−0.47 : 0.49	−0.56 : 0.54	−0.58 : 0.55	−0.57 : 0.57	−0.57 : 0.56	−0.58 : 0.55
		τ	0.23	0.20	0.20	0.23	0.24	0.24	0.24	0.24
Slope	No	Est	1.12	1.10	1.09	1.02	1.02	1.02	1.03	1.02
		CI	1.01 : 1.24	1.01 : 1.18	1.01 : 1.18	0.95 : 1.10	0.95 : 1.10	0.94 : 1.10	0.95 : 1.11	0.95 : 1.10
		PI	0.86 : 1.39	0.90 : 1.30	0.89 : 1.30	0.83 : 1.21	0.86 : 1.19	0.85 : 1.19	0.88 : 1.18	0.87 : 1.17
		τ	0.11	0.08	0.08	0.08	0.06	0.07	0.06	0.06
	Yes	Est	1.18	1.14	1.15	1.03	1.04	1.04	1.06	1.06
		CI	1.00 : 1.36	0.97 : 1.31	0.99 : 1.31	0.89 : 1.16	0.91 : 1.18	0.89 : 1.19	0.91 : 1.21	0.89 : 1.22
		PI	0.71 : 1.66	0.63 : 1.64	0.70 : 1.61	0.61 : 1.44	0.61 : 1.48	0.57 : 1.51	0.59 : 1.53	0.55 : 1.56
		τ	0.20	0.21	0.19	0.18	0.18	0.20	0.20	0.21

Est: Summary estimate;
CI: 95% Confidence interval;
PI: Approximate 95% Prediction interval;
τ: Estimate of heterogeneity.

Heterogeneity in calibration intercepts across the 12 validation studies was also reduced by standardization for all eight prediction models. For instance, for prediction model 8 the unstandardized estimate for τ for the calibration intercepts equaled 0.71, whereas the standardized estimate equaled 0.24.

4.3.3 Summary

In conclusion, the standardized prediction model performance measures have provided greater insight into the differences in case-mix and optimal regression coefficients between the development and validation samples. Standardization disentangles the effects of (differences in) case-mix and regression coefficients, allowing one to assess a prediction model on the appropriateness of its originally estimated regression coefficients across different settings and populations, and thus to assess its (genuine) reproducibility at multiple occasions. Standardization may provide estimates of prediction model performance in external validation studies that are improved or worsened compared to unstandardized estimates, that may be due to any differences in case-mix between development and validation samples.

4.4 Discussion

We proposed a method for standardizing samples in which prediction models are to be validated. When combining the IPD from multiple studies, settings, institutes or populations, differences will often exist in their respective case-mix distributions, as well as their predictor-outcome associations. This, in turn, may lead to excessive heterogeneity in a prediction model's performance across the evaluated samples and thereby distort any inferences about the model's reproducibility. For instance, it may occur that the estimated coefficients of a prediction model remain valid when applied to new settings and populations, but that variation in case-mix distributions affect discrimination and calibration performance. In such cases, the prediction model may not benefit much from local tailoring or updating strategies. Conversely, prediction models with regression coefficients that do not generalize to other populations and settings are most likely to benefit from local revisions.

Standardization methods as shown in this study facilitate the interpretation of prediction model performance differences found in validation samples with respect to the (original) development sample. In particular, by standardizing validation samples with respect to the original development sample, it becomes possible to remove the impact of case-mix effects on prediction model performance estimates found in the validation sample. In other words, standardization allows one to interpret validation study results as if the case-mix distributions would remain unchanged as compared to the development sample. Any heterogeneity in prediction model performance estimates can then only be attributed to the use of invalid regression coefficients and thus to a lack of transportability of the original model.

Since case-mix differences can be found with regard to many variables (predictors and/or outcome), we propose a multivariable standardization approach, which has originally been described in the causal inference literature to balance covariate distributions across patient settings under different 'exposures'. [265, 266] Transpos-

ing this framework to clinical prediction model development and validation research, one can consider the memberships to the development or external validation settings as 'exposures'. A similar approach has been suggested recently to anticipate the external validity of results from RCTs [277] and to use a larger sample size by including propensity weighted external data to assess the intervention effect in a (single) trial. [278]

Although we observed considerable differences in heterogeneity of the prediction models' predictive performance between the standardized and unstandardized measures of performance in the motivating example of 12 external validation studies, the absolute differences in the summary estimates were minor. Overall, the validation samples were not very distinct from the development sample. Specifically, standardization of the external validation sets indicated that the summary discrimination estimates were virtually the same, the estimates for the summary calibration slopes of the prediction models were slightly worse and the estimated calibration-in-the-large was slightly better for a population and setting that was similar to the development sample, compared to the validation samples. Hence, standardization allowed us to interpret the summary estimates of the prediction models' performance in the 12 external validation studies in light of the target population and setting of the development sample. It showed that the calibration slopes of the prediction models were not optimal, indicating that recalibration of the slope may somewhat improve the prediction models' performance in the target population and setting.

In terms of heterogeneity of the prediction models' performance, we did observe considerable differences after standardization. There was less heterogeneity of discrimination after standardization, though, as the number of predictors increased this difference disappeared. As the number of predictors was increased, heterogeneity in discrimination performance resulting from differences in case-mix was partially resolved.

The heterogeneity for the calibration-in-the-large was reduced by standardization for all models. This was expected, since calibration-in-the-large mostly captures case-mix differences in outcome occurence. Conversely, the heterogeneity for the calibration slopes was increased by standardization for all models. This demonstrates that there were differences in case-mix between the development and 12 validation samples, and that the case-mix differences partially masked the differences in optimal regression coefficients between these samples. When standardization filtered out these differences in case-mix, heterogeneity in the calibration slopes became more apparent. Hence, standardization allowed us to interpret the external validation results for both summary measures and heterogeneity measures in the light of a different setting or population.

4.4.1 Limitations and future directions

Standardization using propensity score weighting methods can be performed in different ways. [265, 266] Though, each requires that IPD for the development and validation samples are available, such that the propensity score model can be estimated and applied. This propensity score model may also contain predictors that are not included in the prediction model. In our motivating example, we chose

weights that allowed the validation samples to resemble the (single) development sample in terms of case-mix. Another strategy could be to define weights such that the sample to be standardized approximates all available studies or settings taken together (i.e. 'entire population'), akin to the 'inverse probability weighting' described in the causal inference literature. [265] In fact, the choice between standardization weights should be made according to the target population, that is depending on whether the prediction model aims for a specific clinical setting or to a larger scale. Further studies are needed to compare these weighting methods. Further studies should also take into account other issues that may compromise model transportability, such as measurement error (for this, also see Chapter 5). [260]

Further, propensity scores might also be used to standardize samples for a specific target population during model development on a data set that consists of multiple, combined, data sets. In contrast to the here studied standard dichotomy between development and validation data sets, re-weighting the development data to match a specific target population increases the sample size available for model development in the target population.

For instance, in an IPD-MA with the aim to develop a prediction model, data from RTCs may be included. Due to strict eligibility criteria, data from these RCTs might not fully match the intended target population. Simply stacking every such available data set for model development purposes would then bias model parameters and deteriorate its predictive performance. Standardization may then help to estimate model parameters with respect to the target population and to assess its reproducibility in the targeted population.

4.4.2 Conclusion

Propensity score-based standardization helps to facilitate the interpretation of (heterogeneity in) prediction model performancem observed in (multiple) validation studies and to guide the need for prediction model updating strategies or to accept that the validation sample does not reflect the target population of the developed model. Further research may focus on the use of propensity score weighting during model development on heterogeneous data sets to enhance the reproducibility of prediction models.

Acknowledgements

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

This work is financially supported by the European Union's Horizon 2020 Research and Innovation Programme under ReCoDID Grant Agreement no. 825746.

Chapter 5

Adjusting for misclassification of a predictor in an individual participant data meta-analysis

Valentijn M.T. de Jong, Harlan Campbell, Thomas Jaenisch, Paul Gustafson, Thomas P.A. Debray

Abstract
A common problem in the retrieval and analysis of multiple data sources, such as individual-participant-data meta-analysis (IPD-MA) is the presence of measurement error. Misclassification of binary predictors arises when these study variables are not accurately measured. The presence of misclassification may introduce bias in estimates of parameters (including predictor effects), even when the error is entirely random. Although several methods for addressing misclassification during the development of a prediction model have been proposed, these do not account for the heterogeneity that is often present in individual participant data meta-analysis (IPD-MA).

We aim to develop statistical methods for addressing predictor misclassification in an IPD-MA, where the extent and nature of measurement error may vary across studies. With these methods we aim to facilitate unbiased estimation of adjusted and unadjusted predictor-outcome associations, as well as unbiased estimates of between-study heterogeneity.

We adopt a Bayesian estimation framework and present statistical methods that allow misclassification rates to be dependent on study-level and participant-level characteristics. We illustrate our methodology in a motivating example of the diagnosis of the dengue virus using two predictor variables. In this example, the gold standard measurement for one predictor variable is unavailable for half of the studies. Instead, these studies only measured a surrogate that is prone to misclassification. We evaluate our methodology in a simulation study and assess it for bias, root mean square error (RMSE), coverage and power in estimating a predictor-outcome association.

In the motivating example, our methods reduced the error in the estimates for the predictor-outcome association. In general, our methods yielded estimates with less error than an analysis that was naive with regard to measurement error and an analysis based on gold standard measurements alone. Estimates for heterogeneity of the predictor-outcome association were similar across all investigated methods.

Our simulations show that our framework can appropriately account for misclassification that is dependent on study- and participant-level information. By implementing a proposed misclassification model that models participant-level effects and heterogeneity between studies in the outcome and gold standard and surrogate measurement of the predictor, we obtained valid estimates of the predictor-outcome association, with less RMSE, greater power and similar coverage compared to an analysis that was restricted to observations for which gold standard measurements were available. Heterogeneity estimates were adequate for all studied models.

Our proposed framework can be used to address the presence of misclassification of a predictor variable in an IPD-MA. This framework requires that 1) some studies supply IPD for the surrogate predictor and the gold standard predictor and 2) misclassification is exchangeable across studies conditional on the observed covariates (and outcome). Further work is needed to address other types of misclassification.

5.1 Introduction

Individual participant data meta-analysis (IPD-MA) comprises the pooling and subsequent analysis of the individuals' raw data from multiple studies. As an IPD-MA synthesizes the evidence of all data available to answer a specific research question, it is generally seen as the highest standard of scientific evidence [279]. It may therefore come to no surprise that IPD-MA have become increasingly common to summarize the evidence from experimental and observational studies, and that their results can substantially impact clinical practice. Although IPD-MA are frequently conducted to study the efficacy of therapeutic interventions, they can also be used to investigate etiologic, diagnostic, and prognostic variables. For an IPD-MA to yield valid inference or optimal predictions, however, it is vital that the data is of the highest quality. In all fields, though mostly in observational ones, data may have been gathered with methods or instruments that are inaccurate (i.e. prone to measurement error).

Measurement error is any difference between the value that is observed for a variable and its true value. Measurement error may arise due to a variety of random or systematic causes, such as errors in measurement instruments, the reading of such instruments, poor recall memory, misunderstanding items on questionnaires and data entry and management. The presence of measurement error may introduce (upward or downward) bias in estimates of parameters, even when the error is entirely random and independent of other variables. [28, 34, 35]

Measurement error in categorical variables is referred to as misclassification. It is commonly believed that misclassification, if present, leads to attenuation of predictor-outcome (or exposure-outcome) associations.[280] As a result, researchers often interpret estimates as conservative and dismiss the need for more advanced analyses that account for ME. [281] However, attenuation only occurs when the misclassification is non-differential (that is, misclassification is independent of the outcome given the measured covariates), [27, 28, 30, 33, 35] the predictor has no more than two categories [31, 32] and all covariates are measured without error. [34] When a covariate is also measured with error, the bias that is introduced by including them in a multivariable regression analysis becomes much more difficult to quantify. [34] Further, extreme misclassification can reverse the sign of the observed association. [29]

In an individual participant data meta-analysis (IPD-MA), misclassification may be present in one or more studies. For instance, when the IPD from previously published studies are combined, a different (e.g. less accurate) measurement instrument for a certain predictor variable may have been used in some studies. If one of these instruments is prone to misclassification, this will result in a biased estimate for the corresponding predictor's effect. Therefore, in IPD-MA it is generally recommended to standardize measurements, and where possible to adjust for misclassification to reduce bias. [228]

In meta-analysis, methods must also account for the effects of clustering in individual studies [97] and should allow for heterogeneity of the effect of interest. Hence, methods that account for misclassification must do so as well. Further, it may occur that different measurements methods are used across studies. This directly implies that a gold standard measurement may be missing for entire studies. Applying a

traditional method that accounts for misclassification in IPD-MA therefore requires that the misclassification rate is transportable to other studies. This may be tenable when the measurement instruments, protocol, population and setting are the same in the included studies, but this would be a rare occasion in the context of IPD-MA. Hence, a method that accounts for possible heterogeneity across studies in misclassification as well as outcome prevalence and the predictor-outcome association should then be applied.

In this article, we consider a binary predictor in an IPD-MA that is prone to misclassification error. We distinguish between measurements that are obtained (or defined) according to the gold standard, and measurements that are made using an instrument that is prone to error (further referred to as the surrogate predictor). We subsequently discuss how valid inferences (at least to a certain degree) can be made while the gold standard measurements for the predictor are missing in some studies, using information on the surrogate predictor and the observed patient characteristics. We adopt a Bayesian estimation framework that extends previously proposed methods [282, 283, 284] for addressing misclassification in single studies and in aggregate data meta-analysis (AD-MA).

In section 5.2 we provide our motivating example of the diagnosis of the dengue virus. In section 5.3 we discuss existing methods for dealing with misclassification, and provide our extensions thereof. We apply these methods in section 5.4 and provide a discussion in section 5.6.

5.2 Motivating example: Diagnosing dengue

An estimated 100 million infections of dengue occur globally each year. [285] Although dengue infection is often asymptomatic, patients can present with various clinical symptoms ranging from mild febrile illness to hemorrhagic fever, organ impairment and hypovolaemic shock, and can be fatal. [286, 285] In its early phase, dengue can be difficult to distinguish from other febrile illnesses (OFI) such as influenza, measles, leptospirosis and typhoid due to the similarity of clinical symptoms, which include headache and rash. Therefore, the identification of laboratory and other clinical variables that aid in the differential diagnosis of dengue is imperative. [285] In this motivating example we focus on the strength of the association between muscle pain and dengue vs OFI. To assess the added diagnostic value of muscle pain in the differential diagnosis of dengue vs OFI, a multivariable logistic prediction model can be developed.

Here we show how potentially misclassifying the presence of muscle pain in some studies will affect its apparent association with presence of dengue vs OFI, in patients suspected of Dengue. We use simulated IPD for 10 studies (Table 5.1), that are based on real data gathered in three studies of the IDAMS consortium (Appendix 5.1, page 109). [285] The IPD were generated according to three scenarios with varying heterogeneity in the outcome model. In the first scenario we defined the heterogeneity parameters such that all studies have the same (true) prevalence of dengue conditional on the predictor and covariate and the same (true) predictor-outcome association of muscle pain, conditional on the covariate. In the second scenario we allowed for heterogeneity in the true prevalence of dengue conditional

on the predictor and covariate but not in the true predictor-outcome association, conditional on the covariate. In the third scenario we allowed for the presence of heterogeneity in both the true prevalence of dengue conditional on the predictor and covariate as well as the true predictor-outcome association of muscle pain, conditional on the covariate. In all scenarios we allowed the true prevalence of muscle pain and the true misclassification rates to vary across studies. The challenge is to account for this rate of misclassification that is heterogeneous across studies and depends on patient covariates, while simultaneously accounting for heterogeneity of the prevalence of dengue and heterogeneity in the muscle pain-dengue association. In the following sections we first provide a short overview of methods for accounting for misclassification in single studies and in AD-MA before we move on to accounting for these sources of heterogeneity in IPD, such in this IPD-MA of the muscle pain-dengue association.

Table 5.1: Characteristics of dengue data in scenarios 1, 2 & 3

Scenario	Outcome: dengue		Absent	Present	Total
1	Muscle pain[a]	Absent	1997 (55.6)	1597 (44.4)	3594
		Present	1267 (37.2)	2139 (62.8)	3406
	Muscle pain[b]	Absent	968 (54.5)	808 (45.5)	1776
		Present	655 (38.0)	1069 (62.0)	1724
	Muscle pain[c]	Absent	2029 (52.0)	1870 (48.0)	3899
		Present	1235 (39.8)	1866 (60.2)	3101
	Joint pain	Absent	2096 (52.2)	1920 (47.8)	4016
		Present	1168 (39.1)	1816 (60.9)	2984
2	Muscle pain[a]	Absent	2024 (53.7)	1742 (46.3)	3766
		Present	1187 (36.7)	2047 (63.3)	3234
	Muscle pain[b]	Absent	931 (51.0)	896 (49.0)	1827
		Present	611 (36.5)	1062 (63.5)	1673
	Muscle pain[c]	Absent	2195 (50.9)	2117 (49.1)	4312
		Present	1016 (37.8)	1672 (62.2)	2688
	Joint pain	Absent	2123 (50.7)	2067 (49.3)	4190
		Present	1088 (38.7)	1722 (61.3)	2810
3	Muscle pain[a]	Absent	2056 (56.3)	1593 (43.7)	3649
		Present	1231 (36.7)	2120 (63.3)	3351
	Muscle pain[b]	Absent	1076 (58.6)	760 (41.4)	1836
		Present	572 (34.4)	1092 (65.6)	1664
	Muscle pain[c]	Absent	2185 (52.5)	1975 (47.5)	4160
		Present	1102 (38.8)	1738 (61.2)	2840
	Joint pain	Absent	2184 (53.2)	1921 (46.8)	4105
		Present	1103 (38.1)	1792 (61.9)	2895

Data shown as counts (percentages).
[a] As if it were fully observed in all studies.
[b] As observed in five studies. Missing in the other five.
[c] Potentially misclassified measurement.

5.3 Methods

Many methods have been developed to adjust for misclassification of predictors in the analysis of a single study. These include regression calibration and multiple imputation based methods. Methods for adjusting meta-analyses of aggregate data for misclassification have also been proposed. We start by briefly summarizing these methods and their characteristics. More detailed information is available from Keogh et al. [287]

5.3.1 Adjusting for misclassification in single studies

Regression calibration

In regression calibration, the outcome is regressed on the expected value of the predictor, given the surrogate predictor and covariates. The expected value of the predictor can be estimated by regressing the predictor on the surrogate predictor and covariates for participants for whom all these variables have been measured. When modeling a continuous outcome with linear regression this approach may yield unbiased estimates of the predictor-outcome association. [34] However, regression calibration has been demonstrated to yield (somewhat) biased results when applied to logistic regression [34, 288, 287]. As regression calibration does not use the observed outcome for estimating the expected value of the predictor, it cannot account for differential misclassification.

Multiple imputation

In the multiple imputation approach, gold standard and surrogate measurements of the predictor are treated as separate covariates. Participants for whom the gold standard or surrogate measurement has not been applied are then considered to have missing values for the corresponding covariate(s). If there are sufficient participants for whom the surrogate and gold standard predictor are available, the missing predictors can be imputed.

Multiple imputation for measurement error (MIME) is an implementation of multiple imputation (MI), which was designed to deal with missing data. In MI using chained equations, each variable (or a transformation thereof) is iteratively regressed on all other variables. The estimated regression models are then used to impute missing values. In MIME, the estimated regression models are used to impute the missing gold standard measurement of the predictor for participants for whom only the surrogate is observed. MIME models typically include the outcome as covariate, which naturally accounts for differential error if the imputation model is correctly specified. However, it overestimates the uncertainty in the imputation of the true predictor. [288]

5.3.2 Adjustment for misclassification in a meta-analysis of contingency tables

Most meta-analyses are based on aggregate data. When the predictors are binary, the aggregate data for the predictor-outcome associations are often presented as

counts in contingency tables. Provided that contingency tables for the surrogate-gold standard predictor association are also available, one can adjust for the misclassificication in the surrogate predictor-outcome association that is unadjusted for covariates. [284]

Exchangeability in meta-analysis

As the rate of misclassification may differ across studies, Lian et al recently developed a model that accounts for clustering and heterogeneity and relaxed the assumption of transportability to exchangeability. [284] That is, the degree of misclassification is allowed to vary across studies by applying a random effect. The resulting coefficients for the misclassification model and for the predictor-outcome model need not come from the same studies if exchangeability can be assumed. This is advantageous, as it implies that studies in which misclassification was not investigated can be included in the analysis.

Although Lian et al's model does not assume that misclassification in the measured predictor is common across studies, the exchangeability assumption nevertheless requires that misclassification is independent of any patient-level covariates, given the value of the gold standard measurement of the predictor. [284] In particular, the exchangeability assumption implies that misclassification is assumed to depend solely on study-level variables. This is an important distinction, as misclassification that is non-differential given covariates, may be differential when these covariates are not taken into account. [35] Thus, if misclassification rates are different for the levels of the outcome and patient-level covariates can explain those differences, then these covariates must be taken into account.

However, extending these methods that rely on stratified contingency tables to the analysis of covariate-adjusted predictor-outcome associations may be impractical. It would require that studies provide contingency tables that are stratified for the outcome, gold standard measurement of the predictor, surrogate predictor and every adjustment variable. Clearly, this may be infeasible for a large number of variables.

5.3.3 Adjustment for misclassification in AD-MA

Alternatively, one may opt to adjust for misclassification in a meta-analysis of aggregate data, that is, using predictor-outcome associations (and standard errors) reported in the form of regression coefficients such as (log) risk or odds ratios that have been adjusted for covariates. If all of these reported estimates (including the standard errors) were appropriately adjusted for misclassification in their respective studies, one could analyze these with traditional meta-analysis methods. On the other hand, if the estimation of these covariate adjusted predictor-outcome associations did not include accounting for misclassification, then this would have to occur in the meta-analysis.

If IPD are available for the gold standard and surrogate measurements of the exposure, one might apply a misclassification model to adjust the reported predictor-outcome associations for misclassification, but in this would require the assumption

of exchangeability of misclassification across the included studies. [284] This assumption would clearly be violated in case the misclassification is dependent on participant-level covariates. For instance, in our motivating example the misclassification of muscle pain was associated with the participant-specific value of joint pain. If the measurement for joint pain is missing for a participant, then the information to estimate the expected value of the missing measurement of muscle pain is missing for that participant. In the case of AD-MA, this implies that the covariate joint pain would be missing for the entire study. Thus, any participant-specific misclassification would not be accounted for. In the next section we describe how the exchangeability assumption in meta-analysis models for misclassification can be relaxed IPD are available.

5.3.4 Adjustment for misclassification in a meta-analysis of individual participant data

We extend the methods of Nelson et al [283] and Lian et al [284] to incorporate participant level covariates in a one-stage IPD-MA for potentially misclassified binary predictors. As such, we allow the probability of misclassification to depend on study-level variables and on individual participant level covariates that are observed without error. Further, modeling of IPD allows us to estimate the adjusted (i.e. multivariable) predictor-outcome associations.

Let x_{ij} denote the gold standard measurement of the binary predictor for participant $i, i = 1, ..., I$ in study j, $j = 1, ..., J$. The surrogate predictor is given as x^*_{ij} and represents a possibly misclassified measurement of the predictor. We assume that x^*_{ij} has been observed for all participants in all studies, whereas x_{ij} has only been observed for some participants in some studies. Further, we assume that z_{ij} is a covariate without measurement error and that y_{ij} is a binary outcome.

Following the approach described by Richardson and Gilks [289], we specify three submodels to account for misclassification: a measurement model, a predictor model and an outcome model. In the measurement model, the surrogate predictor (i.e. the measurement of the predictor that is prone to misclassification) is predicted, conditional on the latent gold standard measurement of the predictor, to determine the extent of misclassification. The measurement model models the relation $x^*_{ij} \sim x_{ij}, z_{ij}$. In the predictor model, the latent gold standard measurement of the predictor is regressed on covariates that are measured without error, in order to predict the gold standard measurement of the predictor in participants for whom it is missing. Hence, the predictor models the relation $x_{ij} \sim z_{ij}$. Note that the predictor model is commonly referred to as the exposure model in etiological studies where the gold standard measurement of the exposure is missing. In the outcome model, the outcome is regressed on the latent gold standard measurement of the predictor and on covariates that are measured without error, to determine the predictor-outcome relationship. The outcome model models the relation $y_{ij} \sim x_{ij}, z_{ij}$. Although our model generalizes to multiple covariates, we restrict our notation to a single covariate for simplicity.

Common effects IPD-MA

We start with describing an IPD-MA misclassification model containing three submodels that assumes common effects across studies. Hence, all data are analysed as if they were measured in a single study. In this first model, the probability of misclassification only depends on the value of the gold standard measurement of the predictor. The measurement (sub)model is then given by:

$$\begin{aligned} x^*_{ij} &\sim \text{Bernoulli}(p^*_{ij}), \\ g(p^*_{ij}) &= \lambda \text{ if } x_{ij} = 1, \\ g(p^*_{ij}) &= \phi \text{ if } x_{ij} = 0, \end{aligned} \tag{5.1}$$

where $\lambda \sim N(0, \sigma^2_\lambda), \phi \sim N(0, \sigma^2_\phi)$ and $g(.)$ is a link function. For instance, one could choose the logit for $g(.)$, such that intercept parameters represent log odds and (predictor) coefficient parameters represent log odds ratios. This is equivalent to a measurement model proposed by Nelson et al, [283] as λ and ϕ are parameters that determine the estimated g(sensitivity) and $g(1 - \text{specificity})$, respectively. The above parametrization allows us to introduce covariates to the measurement model in subsequent steps. We leave the variance parameters unspecified, as fixed values may be supplied for these. One may also supply prior distributions for the variance parameters.

The predictor model aims to estimate the relationship between the gold standard measurement of the predictor and covariate(s). It is simultaneously applied to predict the probability that the predictor is present in participants for whom the gold standard measurement of the predictor status is missing. For participants for whom the gold standard measurement of the predictor status is missing, the expected value given covariates is imputed following this model. It is given by:

$$\begin{aligned} x_{ij} &\sim \text{Bernoulli}(p_{ij}), \\ g(p_{ij}) &= \gamma_0 + \gamma_1 z_{ij}, \end{aligned} \tag{5.2}$$

where $\gamma_0 \sim N(0, \sigma^2_{\gamma_0})$ and $\gamma_1 \sim N(0, \sigma^2_{\gamma_1})$. Thirdly, of course, we describe the model that is designed to assess the (adjusted) predictor-outcome association. This outcome model is given by:

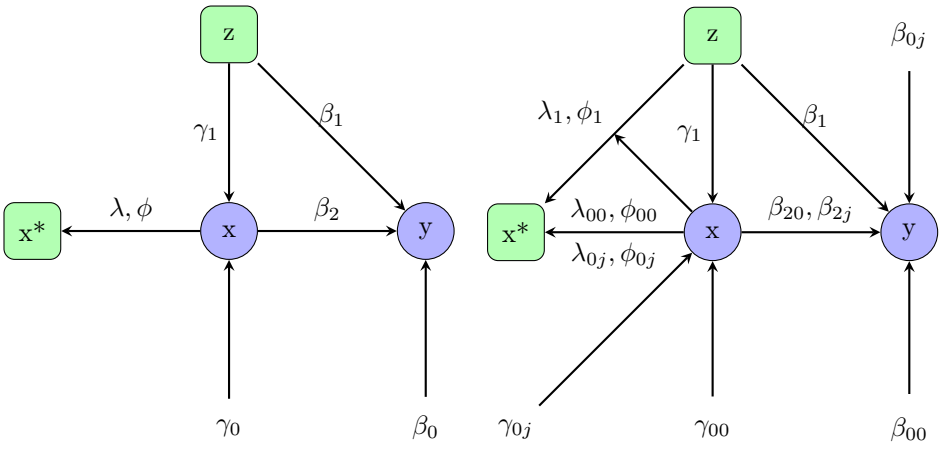
$$\begin{aligned} y_{ij} &\sim \text{Bernoulli}(\pi_{ij}), \\ g(\pi_{ij}) &= \beta_0 + \beta_1 z_{ij} + \beta_2 x_{ij}, \end{aligned} \tag{5.3}$$

where $\beta_0 \sim N(0, \sigma^2_{\beta_0}), \beta_1 \sim N(0, \sigma^2_{\beta_1}), \beta_2 \sim N(0, \sigma^2_{\beta_2}), \beta_0$ is an intercept, β_1 is the coefficient for the covariate and β_2 is the coefficient (log odds ratio) for the predictor of interest. Equations 5.1, 5.2 and 5.3 together make up the least complex misclassification model that we consider here and are illustrated in Figure 5.1. The likelihood of this model is given by the product of the likelihoods of the three submodels, including their priors:

$$p(\lambda,\phi)p(\gamma_0,\gamma_1)p(\beta_0,\beta_1,\beta_2)\prod_j\prod_i p(x^*_{ij}|x_{ij},\lambda,\phi)\prod_j\prod_i p(x_{ij}|z_{ij},\gamma_0,\gamma_1)$$
$$\prod_j\prod_i p(y_{ij}|x_{ij},z_{ij},\beta_0,\beta_1,\beta_2) \tag{5.4}$$

Although the implementation of aforementioned misclassification models is fairly straightforward in an IPD-MA, their justification becomes problematic when studies differ with respect to case-mix, baseline risk, predictor-outcome associations or the extent of misclassification. We therefore discuss how to adjust the submodels accordingly.

Figure 5.1: Diagrams of model equations 5.1, 5.2 and 5.3 (left) and 5.5, 5.8 and 5.11 (right)



Green squares: fully observed data, blue circles: at least partially observed data, not in boxes: parameters. Variance parameters omitted.

Accounting for between-study heterogeneity in the distribution of the predictor

A common situation in IPD-MA is the presence of heterogeneity in case-mix distributions. [97] In particular, when the distribution of the gold standard measurement of the predictor variable varies across studies and the predictor submodel does not account for this, then inadequate predictions will be be made for the unobserved gold standard measurements. We may model the varying prevalence of the gold standard measurement of the predictor x by applying random intercepts to the predictor model, replacing equation 5.2 with:

$$\begin{aligned} x_{ij} &\sim \text{Bernoulli}(p_{ij}),\\ g(p_{ij}) &= \gamma_{00} + \gamma_{0j} + \gamma_1 z_{ij}, \end{aligned} \tag{5.5}$$

where $\gamma_{00} \sim N(0, \sigma^2_{\gamma_{00}}), \gamma_1 \sim N(0, \sigma^2_{\gamma_1}), \gamma_{0j} \sim N(0, \tau^2_{\gamma_{0j}})$. The predictor model's contribution to the likelihood is then given by:

$$p\big(\gamma_{00}, \gamma_1, \tau^2_{\gamma_{0j}}\big) \prod_j \prod_i p\big(x_{ij} | z_{ij}, \gamma_{00}, \gamma_1, \tau^2_{\gamma_{0j}}\big) \tag{5.6}$$

Adjusting for between-study heterogeneity in misclassification

For various reasons, the extent of error in the measurement of the predictor may vary by study in an IPD-MA. This may be modeled by applying random intercepts in the measurement model, which can be interpreted as that the log-odds sensitivity and 1 - specificity vary by study. The measurement model is then given by:

$$\begin{aligned} x^*_{ij} &\sim \text{Bernoulli}(p^*_{ij}), \\ g(p^*_{ij}) &= \lambda_{00} + \lambda_{0j} \text{ if } x_{ij} = 1, \\ g(p^*_{ij}) &= \phi_{00} + \phi_{0j} \text{ if } x_{ij} = 0, \end{aligned} \tag{5.7}$$

where $\lambda_{00} \sim N(0, \sigma^2_{\lambda_{00}}), \phi_{00} \sim N(0, \sigma^2_{\phi_{00}}), \lambda_{0j} \sim N(0, \tau^2_{\lambda_{0j}})$, and $\phi_{0j} \sim N(0, \tau^2_{\phi_{0j}})$. Although it is common to assume a Normal prior distribution for regression coefficients, [33] the choice for a prior distribution for the variance parameters is less straightforward. A prior with too heavy tails will give too much prior weight a on high variance, whereas a prior with thin tails will put to much prior weight on a low variance. [290] We here consider $\tau^2_{\lambda_{0j}} \sim$ inverse-gamma$(\chi_\lambda, \xi_\lambda)$ and $\tau^2_{\lambda_{0j}} \sim$ inverse-gamma(χ_ϕ, ξ_ϕ), but would like to highlight that several alternatives have been proposed, such as the half-Cauchy and half-t distribution. [291]

Adjusting for participant-specific misclassification

A more complex situation arises when misclassification is related to participant-level covariates. For instance, recall of predictor values may be poorer in the elderly, the answering of questionnaires may be hampered by poor literacy and measurement instruments might be designed for specific subgroups of participants. Participant-specific misclassification is particularly problematic if the case-mix distributions vary across studies, as estimates of predictor-outcome associations will then be affected differently across studies. For this reason, the presence of such error can be accounted for by incorporating patient-level covariate effects in the measurement model:

$$\begin{aligned} x^*_{ij} &\sim \text{Bernoulli}(p^*_{ij}), \\ g(p^*_{ij}) &= \lambda_{00} + \lambda_{0j} + \lambda_1 z_{ij} \text{ if } x_{ij} = 1, \\ g(p^*_{ij}) &= \phi_{00} + \phi_{0j} + \phi_1 z_{ij} \text{ if } x_{ij} = 0, \end{aligned} \tag{5.8}$$

where $\lambda_{00} \sim N(0, \sigma^2_{\lambda_{00}}), \lambda_1 \sim N(0, \sigma^2_{\lambda_1}), \phi_{00} \sim N(0, \sigma^2_{\phi_{00}})$ and $\phi_1 \sim N(0, \sigma^2_{\phi_1})$. The contribution of the measurement model to the likelihood is then given by:

$$p\big(\lambda_{00}, \lambda_1, \phi_{00}, \phi_1, \tau^2_{\lambda_{0j}}, \tau^2_{\phi_{0j}}\big) \prod_j \prod_i p\big(x^*_{ij} | x_{ij}, \lambda_{00}, \lambda_1, \tau^2_{\lambda_{0j}}, \phi_{00}, \phi_1, \tau^2_{\phi_{0j}}\big) \tag{5.9}$$

Accounting for between-study heterogeneity in outcome prevalence

Commonly, in data from an IPD-MA and other clustered data sets the prevalence of the outcome varies by study. To account for this effect of clustering within studies, it is generally considered vital that random intercepts for the outcome are applied in an IPD-MA. [97] We can add these to the outcome model as follows:

$$\begin{aligned} y_{ij} &\sim \text{Bernoulli}(\pi_{ij}), \\ g(\pi_{ij}) &= \beta_{00} + \beta_{0j} + \beta_1 z_{ij} + \beta_2 x_{ij}, \end{aligned} \tag{5.10}$$

where $\beta_{0j} \sim N(0, \tau^2_{\beta_{0j}})$ and $\tau^2_{\beta_{0j}} \sim \text{inverse-gamma}(\chi_\beta, \xi_\beta)$.

Accounting for between-study heterogeneity in predictor-outcome associations

Further, the strength of the true predictor-outcome association might also vary by study. To model this, one may adopt a random effects model for the outcome, which does not assume there is a single predictor-outcome association. [117] Instead, it assumes there is a distribution of predictor-outcome associations and it estimates the center and variance of that distribution.

$$\begin{aligned} y_{ij} &\sim \text{Bernoulli}(\pi_{ij}), \\ g(\pi_{ij}) &= \beta_{00} + \beta_{0j} + \beta_1 z_{ij} + \beta_{20} x_{ij} + \beta_{2j} x_{ij}, \end{aligned} \tag{5.11}$$

where $\beta_{00} \sim N(0, \sigma^2_{\beta_{00}}), \beta_2 \sim N(0, \sigma^2_{\beta_2}), \beta_{20} \sim N(0, \sigma^2_{\beta_{20}}), \beta_{0j} \sim N(0, \tau^2_{\beta_{0j}}), \beta_{2j} \sim N(0, \tau^2_{\beta_{2j}})$, β_{20} is the center of the predictor-outcome association distribution and represents the overall association, β_{2j} is the study-specific predictor-outcome association and $\tau^2_{\beta_{2j}}$ is the variance of the distribution of predictor-outcome associations. The random effects assumption is commonly adopted in meta-analysis where sources of between-study heterogeneity cannot (fully) be explained using participant-specific information but need to be accounted for. It is also considered a rather safe assumption, as a random effects model will estimate the variance of the predictor-outcome association at near zero when that association does not vary in the sample. Conversely, a common effects model will lead to inadequate estimates when the common effects assumption does not hold. Equations 5.5 , 5.8 and 5.11 together are illustrated in Figure 5.1, and the contribution of the outcome model to the likelihood is then given by:

$$p\big(\beta_{00}, \beta_1, \beta_2, \tau^2_{\beta_{0j}}, \tau^2_{\beta_{2j}}\big) \prod_j \prod_i p\big(y_{ij} | x_{ij}, z_{ij}, \beta_{00}, \beta_1, \beta_2, \tau^2_{\beta_{0j}}, \tau^2_{\beta_{2j}}\big) \tag{5.12}$$

The models considered here are identifiable only if sufficient information is present in the data. [292, 33] For instance, to estimate equation 5.2 and 5.5 requires that the gold standard measurement of the predictor x_{ij} is observed for sufficient individuals. Strictly speaking, a single (large) study where the gold standard and surrogate measurements have been observed should be sufficient to estimate

the participant-level effects, though more studies would be necessary to estimate the study-level effects. For instance, in our motivating example x_{ij} is available for participants in half of the included studies.

Here we have assumed that the outcome y is available for every participant in every study of the IPD-MA. Though, if unavailable, it could be imputed following equation 5.3, 5.10 or 5.11. To ensure congeniality this imputation model must at least contain the predictor and covariates of the outcome model. [293]

Accounting for differential error

So far we have assumed the error in the measurement of the predictor is non-differential, that is that conditional on the gold standard measurement of the value of the predictor and on the perfectly measured covariates, the error in the measurement is unrelated to the outcome. In any other case the error is differential. An example of differential error is recall bias in a case-control (or case-referent) study, where individuals may overestimate (or underestimate) their predictor, as a result of a known outcome. The methods we described can be extended to allow for differential misclassification, by replacing equation 5.8 with:

$$\begin{aligned} x^*_{ij} &\sim \text{Bernoulli}(p^*_{ij}), \\ g(p^*_{ij}) &= \lambda_{00} + \lambda_{0j} + \lambda_1 z_{ij} + \lambda_2 y_{ij} \text{ if } x_{ij} = 1, \\ g(p^*_{ij}) &= \phi_{00} + \phi_{0j} + \phi_1 z_{ij} + \phi_2 y_{ij} \text{ if } x_{ij} = 0, \end{aligned} \tag{5.13}$$

where $\lambda_{00} \sim N(0, \sigma^2_{\lambda_{00}}), \lambda_1 \sim N(0, \sigma^2_{\lambda_1}), \lambda_2 \sim N(0, \sigma^2_{\lambda_2}), \phi_{00} \sim N(0, \sigma^2_{\phi_{00}}), \phi_1 \sim N(0, \sigma^2_{\phi_1}), \phi_2 \sim N(0, \sigma^2_{\phi_2}), \lambda_{0j} \sim N(0, \tau^2_{\lambda_{0j}}), \phi_{0j} \sim N(0, \tau^2_{\phi_{0j}})$,
$\tau^2_{\lambda_{0j}} \sim$ inverse-gamma$(\chi_\lambda, \xi_\lambda)$ and $\tau^2_{\phi_{0j}} \sim$ inverse-gamma(χ_ϕ, ξ_ϕ). This model bears much resemblance to (Bayesian) MI. The difference is that in MI no measurement model is specified and the surrogate measurement instead appears on the right hand side of the predictor model. That is, in the MI approach the surrogate is treated as just another variable, whereas in our approach it is treated as a surrogate of the gold standard. Although this model this model accounts for a form of differential error, it still assumes that the influence of covariates is the same for each level of the outcome and that the random intercept across studies is common for the levels of the outcome. Alternatively, it may be considered more likely that the nature of the misclassification differs entirely for participants with and without the outcome. Similar to differential misclassification in a single study, [294] this may be accounted for by stratifying the measurement model for the outcome of interest.

$$\begin{aligned} x^*_{ij} &\sim \text{Bernoulli}(p^*_{ij}), \\ g(p^*_{ij}) &= \eta_{00} + \eta_{0j} + \eta_1 z_{ij} \quad \text{if } x_{ij} = 1, y_{ij} = 1, \\ g(p^*_{ij}) &= \theta_{00} + \theta_{0j} + \theta_1 z_{ij} \quad \text{if } x_{ij} = 0, y_{ij} = 1, \\ g(p^*_{ij}) &= \psi_{00} + \psi_{0j} + \psi_1 z_{ij} \quad \text{if } x_{ij} = 1, y_{ij} = 0, \\ g(p^*_{ij}) &= \omega_{00} + \omega_{0j} + \omega_1 z_{ij} \quad \text{if } x_{ij} = 0, y_{ij} = 0, \end{aligned} \tag{5.14}$$

where $\eta_{00} \sim N(0, \sigma^2_{\eta_{00}}), \eta_1 \sim N(0, \sigma^2_{\eta_1}), \theta_{00} \sim N(0, \sigma^2_{\theta_{00}}), \theta_1 \sim N(0, \sigma^2_{\theta_1}), \psi_{00} \sim N(0, \sigma^2_{\psi_{00}}), \psi_1 \sim N(0, \sigma^2_{\psi_1}), \omega_{00} \sim N(0, \sigma^2_{\omega_{00}}), \omega_1 \sim N(0, \sigma^2_{\omega_1}), \eta_{0j} \sim N(0, \tau^2_{\eta_{0j}})$,

$\theta_{0j} \sim N(0, \tau^2_{\theta_{0j}}), \psi_{0j} \sim N(0, \tau^2_{\psi_{0j}}), \omega_{0j} \sim N(0, \tau^2_{\omega_{0j}}), \tau^2_{\eta_{0j}} \sim$ inverse-gamma(χ_η, ξ_η), $\tau^2_{\theta_{0j}} \sim$ inverse-gamma$(\chi_\theta, \xi_\theta), \tau^2_{\psi_{0j}} \sim$ inverse-gamma(χ_ψ, ξ_ψ) and $\tau^2_{\omega_{0j}} \sim$ inverse-gamma$(\chi_\omega, \xi_\omega)$. In case the error is assumed to be restricted to participants with (or without) the outcome, equation 5.14 could easily be simplified by letting $x^*_{ij} = x_{ij}$ for these cases.

5.4 Motivating example: application of methods to dengue IPD-MA

To illustrate the impact of misclassification on observed predictor-outcome associations in an IPD-MA, we apply several modeling strategies to estimate the muscle pain-dengue association in patients suspected of dengue. Hereto, we generated three scenarios for a dengue IPD-MA using real data on dengue as described in section 5.2. In all scenarios we allowed the true prevalence of muscle pain and the true misclassification rates to vary across studies. In the first scenario we defined the heterogeneity parameters such that all studies have the same (true) prevalence of dengue conditional on the predictor and the covariate and the same (true) predictor-outcome association of muscle pain, conditional on the covariate. In the second scenario we allowed for heterogeneity in the true prevalence of dengue conditional on the predictor and the covariate but not in the true predictor-outcome association, conditional on the covariate. In the third scenario we allowed for the presence of heterogeneity in both the true prevalence of dengue conditional on the predictor and covariate as well as the true predictor-outcome association of muscle pain, conditional on the covariate.

We aim to highlight the ability of the methodology we have presented here to restore this association and its uncertainty, while simultaneously accounting for the clustering of participants within studies and allowing for heterogeneity in the muscle pain-dengue association.

5.4.1 Methods

We apply eleven Bayesian binary logistic modeling strategies to estimate the muscle pain-dengue association and its heterogeneity across studies. First, we model the full data with a mixed effects model as if the gold standard measurement was observed for all participants in all studies. In reality, this would not be possible as the gold standard would not be observed for some participants, but here it serves as a comparison with the models that are restricted to the observed data. Second, we apply a mixed effects model on the subset of the data for which the gold standard measurement of the exposure was observed, that is, we apply a so-called complete case analysis. Third, we apply a naive mixed effects modeling strategy, in which we take the surrogate measurement as a proxy for any participant for whom the gold standard measurement is not observed. Finally, we apply the 8 models described in section 5.3.4. These models range from not accounting for heterogeneity and accounting for the simplest form of misclassification to accounting for heterogeneity in all submodels and for a differing extent and nature of misclassification. Although

many more combinations of the submodels exist, for brevity we chose to apply them in the order as outlined, which results in eight full models for accounting for misclassification. We note that some alternative specifications would not be sensible, as the predictor model needs to contain at least the variables that are included in the outcome model.

We estimated all the models with a Gibbs sampler with two independent chains. After 1000 adaptation and 1000 warm up samples, 25000 samples for the estimation of the parameters were performed in each chain. To reduce autocorrelation, we thinned the samples by a factor 5. The presented estimates are based on the remaining 2 * 5000 samples.

5.4.2 Results

In each of the scenarios (see Section 5.2), all models yielded positive estimates with 95% credibility intervals that excluded zero, which in each case may lead to the conclusion that muscle pain is positively associated with dengue. However, we observed considerable differences between the point estimates and estimated 95% credibility intervals of the different models, especially for the common muscle pain-dengue association.

Scenario 1: homogeneous conditional baseline prevalence and predictor-outcome associations across studies

In the first scenario, the estimated association (log-odds ratio) between muscle pain and dengue in the full data was 0.82 (95% CI: 0.67 : 0.98, Table 5.2). The complete case analysis (0.64, 95% Credibility Interval: 0.41 : 0.87) and especially the naive analysis (0.47, 95% CI: 0.34 : 0.60) underestimated this association. The misclassification methods were able to restore the muscle pain-dengue association to various degrees. The model comprising equations 5.8, 5.5 & 5.3, which was the correctly specified model, estimated the log odds ratio for the association at 0.72 (95% CI: 0.54 : 0.90). Surprisingly, the underspecified misclassification models estimated the association with similar or even less error. The overspecified (i.e. models with excess parameters) misclassification errors estimated the association with a larger error, though the errors were still smaller than the naive and complete case analyses.

Table 5.2: Multivariable log odds ratio and heterogeneity estimates (95% Credibility Interval) for presence of muscle pain for diagnosing dengue in scenario 1

Model	β_{20}(95%CI)	$\tau_{\beta_{2j}}$(95%CI)
Full data	0.82 (0.67 : 0.98)	0.05 (0.02 : 0.14)
Complete cases	0.64 (0.41 : 0.87)	0.06 (0.02 : 0.23)
Naive	0.47 (0.34 : 0.60)	0.06 (0.02 : 0.16)
Equations 5.1, 5.2 & 5.3	0.74 (0.55 : 0.93)	
Equations 5.1, 5.5 & 5.3	0.72 (0.53 : 0.91)	
Equations 5.7, 5.5 & 5.3	0.75 (0.56 : 0.93)	
Equations 5.8, 5.5 & 5.3	0.72 (0.54 : 0.90)	
Equations 5.8, 5.5 & 5.10	0.71 (0.54 : 0.90)	
Equations 5.8, 5.5 & 5.11	0.71 (0.53 : 0.91)	0.05 (0.02 : 0.16)
Equations 5.13, 5.5 & 5.11	0.70 (0.52 : 0.90)	0.05 (0.02 : 0.16)
Equations 5.14, 5.5 & 5.11	0.66 (0.45 : 0.88)	0.05 (0.02 : 0.15)

The center of the distribution was estimated by the median of the posterior distribution. Empty cells for $\tau_{\beta_{2j}}$(95%CI) indicate it is assumed to equal zero in the respective model.

All models estimated the between-study heterogeneity of the muscle pain-dengue association very well, as the estimates were very similar to the reference estimate of 0.05 (95% CI: 0.02 : 0.14) in the full data. The exception was the 95% CI of the complete case analysis, which was wider (0.02 : 0.23) than the 95% CI for the other models. This is unsurprising as it uses only a subset of the available data.

Scenario 2: heterogeneous baseline prevalence across studies

In this second scenario, the estimated association (log-odds ratio) between muscle pain and dengue in the full data was 0.76 (95% CI: 0.61 : 0.92, Table 5.3). Again, the complete case analysis (0.66, 95% CI: 0.42 : 0.89) and naive analysis (0.56, 95% CI: 0.42: 0.70) underestimated this association. The misclassification models all estimated the common muscle pain-dengue association with less error than the naive and complete case analysis. The model comprising equations 5.8, 5.5 & 5.10 (i.e. the correctly specified model) estimated the association at 0.74 (95% CI: 0.58 : 0.91), which was nearly identical to the estimates by the analysis on the full data. Also, all misclassification models had narrower 95% Credibility Intervals than the complete case analysis.

All considered models estimated the (lack of) between-study heterogeneity in the muscle pain-dengue association adequately. In the analysis on the full data this heterogeneity was estimated at 0.07 (95% CI: 0.02 : 0.22). Again, the 95% CI for the complete case analysis was the widest (95% CI: 0.02 : 0.33).

Table 5.3: Multivariable log odds ratio and heterogeneity estimates (95% Credibility Interval) for presence of muscle pain for diagnosing dengue in scenario 2

Model	β_{20}(95%CI)	$\tau_{\beta_{2j}}$(95%CI)
Full data	0.76 (0.61 : 0.92)	0.07 (0.02 : 0.22)
Complete cases	0.66 (0.42 : 0.89)	0.08 (0.02 : 0.33)
Naive	0.56 (0.42 : 0.70)	0.06 (0.02 : 0.19)
Equations 5.1, 5.2 & 5.3	0.75 (0.58 : 0.92)	
Equations 5.1, 5.5 & 5.3	0.69 (0.53 : 0.86)	
Equations 5.7, 5.5 & 5.3	0.76 (0.59 : 0.93)	
Equations 5.8, 5.5 & 5.3	0.73 (0.57 : 0.90)	
Equations 5.8, 5.5 & 5.10	0.74 (0.58 : 0.91)	
Equations 5.8, 5.5 & 5.11	0.75 (0.57 : 0.94)	0.09 (0.02 : 0.27)
Equations 5.13, 5.5 & 5.11	0.72 (0.54 : 0.91)	0.08 (0.02 : 0.25)
Equations 5.14, 5.5 & 5.11	0.67 (0.48 : 0.88)	0.07 (0.02 : 0.23)

The center of the distribution was estimated by the median of the posterior distribution. Empty cells for $\tau_{\beta_{2j}}$(95%CI) indicate it is assumed to equal zero in the respective model.

Scenario 3: heterogeneous baseline prevalence and predictor effects across studies

In this final scenario, the analysis on the full data yielded a muscle pain-dengue association of 0.87 (95% CI: 0.60 : 1:14), whereas the complete case analysis estimated it at 1.02 (95% CI: 0.67 : 1.38, Table 5.4) This neatly illustrates that the error in the muscle pain-dengue association estimated by complete case analysis is caused by an increased variance rather than bias, as the estimate by the complete case analysis is now increased with respect to the analysis on the full data, whereas in the other scenarios it was underestimated. As expected, the naive analysis underestimated the association yet again, at 0.60 (95% CI: 0.31 : 0.89).

Three of the misclassification models' point estimates were further away from the point estimate by the full data than the complete case analysis' point estimate, which highlights that applying a misclassification model is not guaranteed to reduce the error in the point estimate. Yet, these were all underspecified models that did not account for the various forms of heterogeneity. The correctly specified model, comprising equations 5.8, 5.5 & 5.11 estimated the muscle pain-dengue association at 0.79 (95% CI: 0.48 : 1.11), which was close to the estimate on the full data. The overspecified models yielded similar estimates.

Except for the complete case analysis, all models that estimated the between-study heterogeneity for the muscle pain-dengue association yielded adequate estimates for this variance. The complete case analysis underestimated the amount of between-study heterogeneity, whereas the underspecified misclassification models (wrongly) assumed it to be equal to 0.

Table 5.4: Multivariable log odds ratio and heterogeneity estimates (95% Credibility Interval) for presence of muscle pain for diagnosing dengue in scenario 3

Model	β_{20}(95%CI)	$\tau_{\beta_{2j}}$(95%CI)
Full data	0.87 (0.60 : 1.14)	0.32 (0.18 : 0.61)
Complete cases	1.02 (0.67 : 1.38)	0.23 (0.05 : 0.73)
Naive	0.60 (0.31 : 0.89)	0.37 (0.21 : 0.69)
Equations 5.1, 5.2 & 5.3	0.81 (0.63 : 0.99)	
Equations 5.1, 5.5 & 5.3	0.58 (0.41 : 0.75)	
Equations 5.7, 5.5 & 5.3	1.09 (0.92 : 1.28)	
Equations 5.8, 5.5 & 5.3	1.04 (0.88 : 1.22)	
Equations 5.8, 5.5 & 5.10	0.97 (0.73 : 1.20)	
Equations 5.8, 5.5 & 5.11	0.79 (0.48 : 1.11)	0.35 (0.19 : 0.67)
Equations 5.13, 5.5 & 5.11	0.80 (0.48 : 1.10)	0.35 (0.18 : 0.68)
Equations 5.14, 5.5 & 5.11	0.82 (0.48 : 1.14)	0.34 (0.17 : 0.67)

The center of the distribution was estimated by the median of the posterior distribution. Empty cells for $\tau_{\beta_{2j}}$(95%CI) indicate it is assumed to equal zero in the respective model.

5.4.3 Summary

Overall, the results of this motivating example on the association between muscle pain and dengue highlight the impact of misclassification on a predictor-outcome association. The misclassification models estimated the predictor-outcome association with less error (where the full data is taken as reference) than both the complete-case and naive approaches, with the exception for some models that were underspecified in scenario 3. This suggests that even in these scenarios for relatively small IPD-MAs, the more complex (possibly overspecified) models seem more suitable than the simpler (possibly underspecified) models.

In general, the models provided adequate estimates of the heterogeneity of the muscle pain-dengue association. The exception was the complete case analysis, which yielded different point estimates due the fact that these estimates were based on different data and which yielded wider credibility intervals due to the fact that these interval estimates were based on less data. In conclusion, the misclassification methods that accounted for heterogeneity in the various submodels gave the best available estimates of the muscle pain-dengue association and its heterogeneity.

5.5 Simulation study

We performed a simulation study to assess the impact of misclassification on estimated predictor-outcome associations and the heterogeneity thereof in an IPD-MA and to assess the validity of our methodology. We aim to highlight the bias that occurs in a predictor-outcome association when misclassification is not accounted for

and the ability of the methodology we have presented here to provide (possibly) unbiased estimates of these associations while propagating the uncertainty induced by misclassification and the various forms of heterogeneity, to facilitate valid inference.

5.5.1 Simulation methods

In each repetition of the simulation we applied four models on the simulated data. First, we performed analyses on the full data as if the gold standard predictor was observed for all (simulated) participants, which may serve as a comparison in the interpretation of the results. Second, we applied a model on the complete cases, that is only on the participants for whom the gold standard predictor was observed. Third, we applied a naive model in which the surrogate measurement of the muscle pain was used for participants for whom the gold standard measurement was not available. Finally, we applied the misclassification model given by equations 5.8, 5.5 & 5.11.

The data were simulated with the same data generating mechanism as in scenario 1 of the motivating example considering the diagnosis of dengue: there was heterogeneity in the distribution of the predictor of interest (muscle pain), but not in the true prevalence of dengue conditional on the predictor and covariate and not in the true predictor-outcome association, conditional on the covariate. We analyzed the estimates for the common predictor-outcome association for each model in terms of percentage bias and root mean square error (RMSE) relative to the true association, (statistical) power and coverage probability of the 95% Credibility Interval. We performed one thousand replications of the simulation in R 3.5.2. [229]

We estimated all the models with a Gibbs sampler with two independent chains using JAGS 4.3.0. After 1000 adaptation and 1000 warm up samples, 25000 samples for the estimation of the parameters were performed in each chain. To reduce autocorrelation, we thinned the samples by a factor 5. The presented estimates are based on the remaining 2 * 5000 samples.

5.5.2 Simulation results

As expected, the analyses on the full (observed and unobserved) data yielded practically unbiased estimates of the muscle pain-dengue association, had a nominal coverage rate and had the lowest RMSE and highest power of the compared analyses (Table 5.5). In practice, of course, the unobserved data will be unavailable. The naive method of substituting the surrogate predictors for the missing gold standard predictors yielded biased estimates. This increased the RMSE and reduced the power and coverage, giving it the worst performance of the compared methods in terms of all the assessed measures except for power.

Table 5.5: Simulation results for the estimated common predictor-outcome association

Model	Mean	RMSE	% Bias	Power	Coverage
Full data	0.80	0.21	2.49	0.97	0.94
Complete cases	0.80	0.29	2.98	0.75	0.96
Naive	0.48	0.34	-37.98	0.80	0.62
Equations 5.8, 5.5 & 5.11	0.79	0.24	1.66	0.92	0.96

RMSE: Root mean square error.
Coverage: Coverage probability of the 95% Credibility Interval.

The complete case analyses fared much better. By restricting the analyses to data measured without error, practically unbiased estimates were produced and the nominal coverage rate was retained. Due to the reduced sample size, however, the variance of the estimates increased, which increased the RMSE and reduced the power. Finally, the misclassification model was able to restore the muscle pain-dengue association, yielding practically unbiased estimates and retaining nominal coverage rates. As this model uses all observed data, the variance of the estimates was the lowest of the three feasible models, which resulted in the lowest RMSE and highest power. In conclusion, the misclassification model provided the best estimates of the muscle pain-dengue association.

5.6 Discussion

As measurement error or misclassification may cause bias in estimated predictor-outcome associations, standard errors and between-study heterogeneity in IPD-MA, it is essential to account for this. We have unified methods for misclassification in meta-analysis in a one-stage Bayesian meta-analysis framework. Our methodology allows for incorporation of covariates on the individual participant level to facilitate valid inference regarding therapeutic and etiologic effects, and added diagnostic and prognostic value. This modeling of the individual participant outcome, predictor and covariate values occurs via three submodels: one for modeling the measurements, one for modeling the (gold standard) predictor and one for modeling the outcome of interest. By doing so, individual level effects are accounted for in each part of the analysis. This, in turn, restores the association between the predictor and the outcome.

In our motivating example data sets, the association between muscle pain and dengue could be estimated with less error by applying the proposed misclassification models with individual participant covariate effects. These models account for the potential between-study heterogeneity in the prevalence of dengue and yielded adequate estimates of between-study heterogeneity of the muscle pain-dengue association.

In our simulations, we considered that baseline outcome prevalence conditional on covariates and predictor effects are homogeneous across studies, and compared

the performance of several models. We found that complete case analysis performed reasonably well, as it yielded unbiased estimates and adequate coverage of the true association. Though its estimates had increased variance, leading to considerably reduced statistical power. In practice, the feasibility of restricting the analysis to patients with complete data for the (gold standard) predictor will depend on the remaining sample size. If this number is low, the variance of the resulting estimates will be large and power negligible. In the extreme case, gold standard measurements are entirely unavailable for participants for whom the outcome is available, making this method impossible. In addition, the validity of a complete case analysis may become challenging when patients (or studies) for which only surrogate predictors are available differ with respect to covariates that are not part of the outcome model.

The simulations also demonstrated that our proposed methodology for misclassification models was able to provide (approximately) unbiased estimates of the muscle pain-dengue association. This is because the misclassification in the predictor was correctly specified and because sufficient data were available to estimate all model parameters. Note, however, that all of the applied models were overspecified in terms of heterogeneity parameters, as the true baseline prevalence of dengue and the true muscle pain-dengue association were, in fact, homogeneous across studies conditional on the predictor and covariate. By treating the presence of between-study heterogeneity as unknown (i.e. allowing for estimates greater than 0), the variance was increased for all models, which increased the RMSE and reduced the statistical power. As this affected all the models in the simulation equally, it had no impact on the validity of the comparison between model performance in the simulation.

In general, overspecification should not induce bias in the estimates, provided that the sample contains enough information to estimate all parameters. Nor should it affect the coverage as the models appropriately account for the uncertainty. However, we stress that if we had applied an underspecified misclassification model, we would expect to have observed (some) bias in the estimates for the muscle pain dengue-association, as well as less favourable statistical properties in terms of RMSE and coverage and depending on the nature of the misspecification also in terms of statistical power. After all, although the misclassification was non-differential given covariates, once those covariates are removed from the model the misclassification may become differential. [34] Therefore, it may be a sensible approach to apply a misclassification model that accounts for differential misclassification whenever sufficient data are available to reliably estimate such a model.

5.6.1 Limitations and future directions

Although we recommend the implementation of misclassification models, an alternative strategy is to implement models that require fewer assumptions. Two such methods, RC and MIME, do not specify measurement models and require fewer distributional assumptions and are therefore described as functional methods [33] or reclassification methods [295] In contrast, in structural methods such as ours, a predictor model is specified, which when analyzed with Bayesian method allows for the appropriate propagation of uncertainty. Though, this requires assumptions on the distribution of the gold standard measurement of the predictor and its surrogate

measurement. [33] However, we focused on the scenario where the predictor is a binary variable that is potentially misclassified, which is common in epidemiology. This binary variable is assumed to follow a Bernoulli distribution, so specification of a predictor model does not add a major assumption [34] aside from congeniality, which is also required for RC and MIME. Although both of these methods have been applied to account for misclassification in single studies, neither has yet been adapted to the heterogeneous setting that is IPD-MA. This would require the specification of multiple heterogeneity parameters. We suggest that further research may focus on integrating these into the IPD-MA framework.

In case the predictor is a continuous variable which has been transformed into a binary variable at a specific cut-off point, alternative assumptions are needed for modeling the distribution of the predictor and its measurement error (see e.g. [33]). Our method could be further extended in case multiple surrogate predictor measurements are available for some or each participant, by specifying a measurement model for each surrogate measurement.

In the simulation study, we generated the data from only a single data generating mechanism and applied only one misclassification model as this simulation was intended as a *proof of concept*, not to assess the relative performance of all the described models in a variety of scenarios. All of the methods discussed here require covariates that predict the value of the gold standard measurement of the predictor to be fruitful. If the available covariates are not predictive of the missing gold standard predictor or the surrogate predictor, only noise would be added by including individual participant covariate effects in the predictor and measurement models, respectively.

Due to the influence of misclassification on predictor-outcome associations and the presence of between-study heterogeneity, and the increase in parameters that are required to account for this, a larger amount of data are necessary than in an IPD-MA where misclassification is absent. This should be especially the case for the more complex misclassification models. In our simulation study however, we simulated data from one thousand individual participants spread over 10 studies and the results show that this was enough to obtain approximately unbiased estimates of the predictor-outcome association, with slightly reduced accuracy compared to the model on the full data. In a typical IPD-MA, where the sample size is often much larger, there should be enough information to estimate the more complex misclassification models.

5.6.2 Conclusion

In an IPD-MA, the gold standard measurement of a predictor may be entirely unavailable for all participants in some studies, or unavailable for some participants in all studies, leaving the researcher with only surrogate measurements for these participants. If ignored, this induces bias in the estimated parameters for predictor-outcome associations and other parameters of interest, which must be accounted for. Our Bayesian methodology can be applied to participant level data to reduce the error in the estimate of the predictor-outcome association compared to a analyses restricted to participants for whom the gold standard measurement is observed, while appropriately propagating uncertainty for all parameters. This may provide

unbiased estimates of the predictor-outcome association, its coverage of the true effect and its heterogeneity across studies, provided that the model is specified correctly.

5.7 Acknowledgements

This work is financially supported by the European Union's Horizon 2020 Research and Innovation Programme under ReCoDID Grant Agreement no. 825746. We thank the IDAMS consortium for providing aggregate data on the diagnosis of dengue.

Appendix 5.1: Dengue data

The IDAMS consortium [285] provided aggregate data on muscle pain, joint pain and dengue vs other febrile illness (OFI) stratified by three sites (n = 700, 700 and 500), as well as a common association between muscle and joint pain. The IDAMS consortium has collected other clinically important variables, which we do not consider here.

From this data point estimates for the intercept and log odds ratio for the predictor model (equation 5.5) and subsequently the intercept for the outcome model (equation 5.11) were estimated with `optim` in R. [229] As heterogeneity estimates are unreliable in only three sites/studies, we chose suitable values for these for three different scenarios. In the first scenario we set the heterogeneity parameters such that the studies all had identical true incidences of dengue conditional on muscle and joint pain and identical true predictor-outcome (muscle pain-dengue) associations conditional on the covariate, joint pain. In the second scenario we allowed for heterogeneity in the true incidence of dengue conditional on muscle and joint pain but not in the true predictor-outcome association conditional on the covariate, joint pain. In the third scenario we allowed there to be heterogeneity in both the true incidence of dengue conditional on muscle and joint pain as well as the true predictor-outcome association conditional on the covariate, joint pain. We generated the three scenarios with different simulation seeds, so that unique data sets were generated.

5.7.1 Parameters

The overall prevalence of joint pain was 0.414. We set the standard deviation for the prevalence of joint pain to 0.1 on the logit scale. The fixed effects for the predictor model were estimated at $\gamma_{00} = -1.70, \gamma_1 = 4.26$. We set $\sigma_{\gamma_{0j}}$ to 0.25. The fixed effects for the outcome model were $\beta_{00} = -0.26, \beta_1 = -0.06, \beta_2 = 0.78$. We generated data sets for three different scenarios with differing heterogeneity in the outcome model. In scenario 1: $\sigma_{\beta_{0j}} = 0$ and $\sigma_{\beta_{2j}} = 0$. In scenario 2: $\sigma_{\beta_{0j}} = 0.25$ and $\sigma_{\beta_{2j}} = 0$. And in scenario 3: $\sigma_{\beta_{0j}} = 0.25$ and $\sigma_{\beta_{2j}} = 0.15$.

The study specific parameters used to generate the data that was used in the analyses reported in this paper were then generated as follows. We set the number

of studies to 10. Study-specific parameters for intercepts and log odds ratios were sampled from normal distributions with their corresponding point and heterogeneity estimates. Prevalences were sampled on the inverse logit scale and then converted to prevalences using the logit function.

5.7.2 Individual participant data

Sampling of individual observations was performed using the parameter estimates as follows. We set the sample size to a value similar to the those of the IDAMS consortium: 700 per study, giving a total of 7000 patients. Data for the covariate joint pain were sampled first according to the study-specific prevalences. Then the predictor model (equation 5.5) was applied to sample the muscle pain status. Then the outcome model (equation 5.11) was applied to sample dengue status.

The misclassified predictor was generated by equation 5.8, with the following parameter values: $\lambda_{00} = 3, \sigma^2_{\lambda_{0j}} = 1, \lambda_1 = -2, \phi_{00} = -3, \sigma^2_{\phi_{0j}} = 1$ and $\phi_1 = 2$. The resulting sensitivity and specificity for the sampled true and misclassified muscle pain variables were respectively 0.81 and 0.90 in the full sampled data in scenario 1, 0.78 and 0.96 in scenario 2 and 0.76 and 0.92 in scenario 3. Finally, for the naive and the misclassification methods, for the first five studies the true values for muscle pain were removed, so that only the potentially misclassified values were available for those studies.

Chapter 6

Sample size considerations and predictive performance of multinomial logistic prediction models

Valentijn M.T. de Jong, Marinus J.C. Eijkemans, Ben van Calster, Dirk Timmerman, Karel G.M. Moons, Ewout W. Steyerberg, Maarten van Smeden. Sample size considerations and predictive performance of multinomial logistic prediction models. *Statistics in Medicine*. 2019;38(9):1601–19. DOI: 10.1002/sim.8063

Abstract
Multinomial Logistic Regression (MLR) has been advocated for developing clinical prediction models that distinguish between three or more unordered outcomes. We present a full-factorial simulation study to examine the predictive performance of MLR models in relation to the relative size of outcome categories, number of predictors and the number of events per variable. It is shown that MLR estimated by maximum likelihood yields overfitted prediction models in small to medium sized data. In most cases, the calibration and overall predictive performance of the multinomial prediction model is improved by using penalized MLR. Our simulation study also highlights the importance of events per variable in the multinomial context as well as the total sample size. As expected, our study demonstrates the need for optimism correction of the predictive performance measures when developing the multinomial logistic prediction model. We recommend the use of penalized MLR when prediction models are developed in small data sets, or in medium sized data sets with a small total sample size (i.e. when the sizes of the outcome categories are balanced). Finally, we present a case study in which we illustrate the development and validation of penalized and unpenalized multinomial prediction models for predicting malignancy of ovarian cancer.

6.1 Introduction

Prediction models are developed to estimate probabilities that conditions or diseases are present (diagnostic prediction) or will occur in the future (prognostic prediction). [296, 297] Most prediction models are developed to estimate the probability for two mutually exclusive diagnostic or prognostic outcomes (events versus non-events). [3, 298] However, for real diagnostic and prognostic questions there are often more than two diseases or conditions that need to be assessed. For instance, the presence of various alternative diseases must be considered when dealing with real patients (i.e., the so-called differential diagnosis). [299, 4] Similarly, there are often also more than two possible prognostic outcomes considered in patients diagnosed with a certain disease (e.g., progression free survival, disease free survival and death as outcome categories). Biesheuvel et al. [298] recognized that the polytomous nature of prediction questions should be taken into account more often in the development of prediction models, suggesting the use of multinomial logistic regression (MLR). While the use of MLR is still relatively rare, applications of MLR for risk prediction are found in a variety of medical fields, such as in predicting the risk of several modes of operative delivery, [300] predicting the risk of three prognostic outcomes of elderly after hospitalization, [301] the differential diagnosis of four types of ovarian tumors [302] and the differential diagnosis of three bacterial infections in children. [303]

So far, the operational characteristics of MLR models in relation to development data characteristics have not been evaluated. In contrast, the relevance of data characteristics for prediction models' out-of-sample performance has been clearly demonstrated for prediction models with binary and time-to-event outcomes. [304, 305] For these models, minimal sample size criteria have been suggested, supported by simulation studies, and a minimum of roughly 10 events per predictor variable (EPV) has been advocated for the development of these binary or time-to-event prediction models. [306, 245, 305, 3, 307, 308, 309, 297] For situations where $EPV < 20$, "shrinkage" of the regression coefficients has been recommended to reduce the chances of overfitting. [246, 304, 3] It is unclear to what extent these rules-of-thumb also apply to the polytomous case of MLR.

In this study we focus on the predictive performance of MLR models that are developed in small to medium sized data sets (multinomial $EPV \leq 50$). We study the effects of the number of multinomial events per variable (EPV_m), relative outcome sizes (frequencies) and number of predictors. In a sensitivity analysis we assess the effects of correlations between the predictors and the type of predictors. We compare the performance of MLR estimated by Maximum Likelihood (ML) and two popular penalized estimation methods that perform shrinkage of the regression coefficients (lasso and ridge regression [310, 311]). This article is structured as follows. In the next section, we describe the estimation methods and we provide a brief overview of predictive performance measures for multinomial logistic regression models. In section 6.3 and 6.4 we present our simulation study, and in section 6.5 we present our case study of predicting malignancy of ovarian cancer. Finally, a discussion is provided in section 6.6.

6.2 Multinomial logistic regression model

6.2.1 MLR Model

Let y_{ij} denote the presence ($y_{ij} = 1$) or absence ($y_{ij} = 0$) of multinomial outcomes $j, j = 1, \ldots, J$, for observation $i, i = 1, \ldots, N$. Let $\boldsymbol{x}_i$ denote observation i's R-dimensional vector of the predictor variables, $r = 1, \ldots \ldots, R$. We further assume that $\sum_j y_{ij} = 1$. Taking J as the reference outcome, the MLR for predicting the probabilities $\pi_{ij}(\boldsymbol{x}_i)$ for outcomes $j = 1, \ldots, J-1$ can then be defined by the multinomial logit [312]:

$$\pi_{ij}(\boldsymbol{x}_i) = \frac{\exp(\alpha_j + \boldsymbol{\beta}_j' \boldsymbol{x}_i)}{1 + \sum_{h=1}^{J-1} \exp(\alpha_h + \boldsymbol{\beta}_h' \boldsymbol{x}_i)}, \tag{6.1}$$

where $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jR})'$ denotes the coefficients for the j^{th} linear predictor, except its intercept α_j. For the reference outcome, $\pi_{iJ}(\boldsymbol{x}_i) = 1/\left(1 + \sum_{h=1}^{J-1} \exp(\alpha_h + \boldsymbol{\beta}_h' \boldsymbol{x}_i)\right)$. Hereafter, we refer to $\pi_{ij}(\boldsymbol{x}_i)$ simply as the risk of outcome j. ML estimation of model 6.1 proceeds by maximizing the log-likelihood $l(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{j=1}^{J} \sum_{i=1}^{N} y_{ij} \log \pi_{ij}(\boldsymbol{x}_i)$.

Penalized MLR

ML is known to produce parameter estimates $\hat{\boldsymbol{\beta}}$ that yield too extreme predictions in new samples, when estimated in small samples. [3] In this paper we therefore also apply lasso [310] (least absolute shrinkage and selection operator) and ridge estimation[313, 314, 311, 315]. Both of these approaches to shrinkage work via a penalty function and are directly applicable to MLR models. By shrinking the ML estimates $\hat{\boldsymbol{\beta}}$ towards the null-effect ($\boldsymbol{\beta} = 0$), both lasso and ridge produce probability estimates that tend to be less extreme (further away from the boundaries of 0 and 1) than the probabilities one would obtain with ML MLR. A slightly modified multinomial logit function is convenient for penalization, as the penalization removes the necessity to put restrictions on the reference category [316]: $\pi_{ij}(\boldsymbol{x}_i) = \exp(\alpha_j^* + \boldsymbol{\beta}_j^{*\prime} \boldsymbol{x}_i) / \sum_{h=1}^{J} \exp(\alpha_h^* + \boldsymbol{\beta}_h^{*\prime} \boldsymbol{x}_i)$.

The penalized MLR models are estimated by maximizing the penalized log-likelihoods $l(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) = \sum_{j=1}^{J} \sum_{i=1}^{N} \{y_{ij} \log \pi_{ij}^*(\boldsymbol{x}_i)\} - \lambda_1 \sum_{j=1}^{J} \sum_{r=1}^{R} |\beta_{jr}^*|$ and $l(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) = \sum_{j=1}^{J} \sum_{i=1}^{N} \{y_{ij} \log \pi_{ij}^*(\boldsymbol{x}_i)\} - \lambda_2 \sum_{j=1}^{J} \sum_{r=1}^{R} \beta_{jr}^{*2}$, for lasso and ridge, respectively. A consequence of the lasso's penalty is that coefficients can be shrunk to (exactly) zero, thereby removing a predictor variable from the equation. Estimation occurs via pathwise coordinate descent, which starts at large λ_1 and λ_2 values, such that all of the $\boldsymbol{\beta}^*$ are zero. The λ_1 and λ_2 values are then iteratively decremented, allowing the $\boldsymbol{\beta}^*$ vectors to increasingly deviate from zero. Maximization of the penalized log-likelihood proceeds by performing partial Newton steps, leading to a path of solutions. For every value of both λ_1 and λ_2 a $\boldsymbol{\beta}^*$ vector is attained. [316] In this study, the optimal λ_1 and λ_2 parameters (i.e., tuning parameters), for lasso and ridge respectively, are estimated by a search over a grid of possible values, selecting the values for λ_1 and λ_2 that minimize Deviance in 10-fold cross-validation. [316]

6.2.2 Predictive performance measures

As not all predictive performance measures for binary outcomes directly generalize to multinomial outcomes, we provide details of the multinomial predictive performance measures that were used in our study in this section, and an overview in Table 6.1.

Table 6.1: Multinomial Prediction Performance Measures.

Aspect	Measure	Interpretation
Discrimination	PDI	PDI $= 1/J$: no discriminative performance. PDI $= 1$: perfect discrimination.
Calibration	Calibration slope	Calibration slope < 1: overfitting. Calibration slope > 1: underfitting.
Overall performance	Brier score	Brier score $= 0$: Perfect predictive performance. Brier score $= 2$: completely imperfect predictive performance.
	Nagelkerke R^2	Nagelkerke $R^2 = 0$: 0% explained variation. Nagelkerke $R^2 = 1$: 100% explained variation.

Discrimination

The discriminative ability of prediction models with a binary outcome is commonly expressed by the concordance probability or c-statistic, [317] and by the c-index for time-to-event models. [318] We consider a generalization of the c-statistic to multinomial outcomes: the polytomous discrimination index (PDI). [319] The PDI is an estimator for the probability of correctly identifying a randomly selected case in a set of cases consisting of one case from each outcome category. [319] The PDI takes on the value 1 for perfect discrimination and $1/J$ for random discrimination. The PDI can be interpreted as the probability that the outcome of a randomly selected individual in a set of J different cases is correctly identified. [319]

The PDI is defined as follows. Let $q_h, q_h = 1, \ldots, n_h$, denote the observations with outcome h, and $\pi_{ij \in q_h}(\boldsymbol{x}_i)$ denote the predicted risk of outcome j for individuals with outcome h. First, the outcome specific components of the PDI are computed, denoted by PDI_h. For each possible set of J cases with a different observed outcome, determine whether the predicted risk for outcome h is highest for a case with observed outcome h. The value on an outcome specific component PDI_h equals the proportion of sets for which this is true, and can be interpreted as the probability that a randomly selected individual with outcome h is correctly identified as such in a set of J randomly selected cases. Second, the PDI is given by the average of the outcome specific PDI_h components. Formally, [319]

$$\text{PDI}_h = \frac{1}{n_1 \cdots n_J} \sum_{q_1=1}^{n_1} \cdots \sum_{q_J=1}^{n_J} C_h \left(\pi_{ij \in q_1}(\boldsymbol{x}_i), \ldots, \pi_{ij \in q_J}(\boldsymbol{x}_i) \right), \tag{6.2}$$

where C_h is an indicator function taking on the value 1 if $\pi_{ij \in q_h}(\boldsymbol{x}_i) > \pi_{ij \in q_j}(\boldsymbol{x}_i)$, for all $q_j \neq q_h$, or $1/t$ in case of ties, where t is the number of ties in

$\pi_{ij\in q_1}(\boldsymbol{x}_i), \ldots, \pi_{ij\in q_J}(\boldsymbol{x}_i)$, or else 0. By taking the mean of outcome specific components, the PDI is obtained: $\text{PDI} = \frac{1}{J}\sum_{h=1}^{J}\text{PDI}_h$.

Calibration slope

Calibration slopes are a measure of the calibration of a prediction model's linear predictors lp_{ij}, $\text{lp}_{ij} = \alpha_j + \boldsymbol{\beta}_j'\boldsymbol{x}_i$. For computation of the calibration slopes, we followed the approach of Van Hoorde et al., [320] who extended the recalibration framework of the binary logistic model [244, 321] to multinomial outcomes:

$$\log\left(\frac{P(y_i = j)}{P(y_i = Q)}\right) = \gamma_j + \theta_j \text{lp}_{i,j} \tag{6.3}$$

where γ_j is the calibration intercept for outcome category j, $\text{lp}_{i,j}$ is the linear predictor of outcome category j versus the referent Q (which need not be the same as the referent in equation 6.1) for observation i, and θ_j is the calibration slope for outcome category j versus the referent Q. Estimates of $\theta_{j\neq Q}$ are obtained with unpenalized MLR, whereas θ_Q naturally equals zero and is disregarded.

As ML perfectly calibrates the coefficients to the development sample, it will always attain a calibration slope of 1 there. We assess out-of-sample calibration, where a slope < 1 is evidence of overfitting, and a slope > 1 is evidence of underfitting. [321] As the value of the multinomial calibration slopes depend slightly on the choice of the reference category, [320] we computed all possible calibration slopes with each category as the reference once.

6.2.3 Overall performance

The overall performance measures quantify the distance between the predicted and observed outcomes and thus capture both the discrimination and calibration of the model. [3] The Brier score quantifies the squared distance between the observed outcomes and the predicted probabilities. [322] It can take values from 0 for perfect predictions to 2 for completely inaccurate predictions. The Brier score for a MLR model is defined by:

$$\text{Brier score} = \frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{N}(\pi_{ij}(\boldsymbol{x}_i) - y_{ij})^2. \tag{6.4}$$

The Nagelkerke R^2 estimates the proportion of explained variation in a discrete outcome variable [323]: it equals 0 for no explained variation and 1 for a complete explanation of the variation. [323] Let $l(0)$ and $l(\hat{\boldsymbol{\beta}})$ be the log-likelihood for an intercept-only MLR model and the MLR model under scrutiny, respectively. Then:

$$R^2_{\text{Nagelkerke}} = \frac{1 - \exp(\frac{2}{N}[l(\hat{\boldsymbol{\beta}}) - l(0)])}{1 - \exp(\frac{2}{N}l(0))}. \tag{6.5}$$

6.3 Simulation study - methods

6.3.1 Main simulation settings

For ease of presentation we focused our simulations on the simplest extension of the binary logistic regression model by studying the MLR for $J = 3$ outcome categories. Sixty-three Monte Carlo simulation scenarios were investigated by fully crossing the following simulation factors:

- Multinomial EPV: $3, 5, 10, 15, 20, 30$ and 50 events per predictor.
- Relative frequencies of the 3 outcome categories. Levels: $1 : (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$; $2 : (\frac{2}{20}, \frac{9}{20}, \frac{9}{20})$; $3 : (\frac{8}{10}, \frac{1}{10}, \frac{1}{10})$.
- Number of predictors (R): $4, 8$ and 16.

In binary logistic regression, the number of events per variable (EPV) is defined by the ratio of the number of observations in the smallest of two outcome categories divided by the number of estimated regression coefficients, excluding the intercept. [324] In parallel, we define EPV_m by ratio of the smallest number of observations in the multinomial outcome categories divided by the effective number of regression coefficients excluding the intercepts. The number of effective regression coefficients excluding the intercept is given by: $(J - 1)R$. Further, for categorical predictors with G categories the number of effective regression coefficients per predictor is $(J - 1)(G - 1)$.

Predictor covariate vectors were drawn from multivariate normal distributions with the covariance matrix an identity matrix. For the development of clinical prediction models, predictor variables may be selected based on expert knowledge, [3, 58] in which case variables with varying predictive impact may be present, and true noise variables predictors (regression coefficient of data generation mechanism of exactly zero) may be infrequent. This simulation was designed to mimic this situation and therefore did not include noise predictors. For the scenario with $R = 4$, $\boldsymbol{\beta}_1 = \{-0.2, -0.2, -0.5, -0.8\}$ and $\boldsymbol{\beta}_2 = \{0.2, 0.2, 0.5, 0.8\}$, corresponding to small (± 0.2), medium (± 0.5) and large (± 0.8) predictor effects of category 1 and 2 versus the referent category. [325] For simulation scenarios with 8 and 16 predictors, predictor effects were similarly distributed, i.e. $\frac{1}{2}$ small, $\frac{1}{4}$ medium and $\frac{1}{4}$ large effects. The true intercepts for each linear predictor were approximated numerically (Appendix A, https://doi.org/10.1002/sim.8063). Outcome data were sampled from a multinomial distribution, where the probability of drawing each outcome was computed by applying the multinomial logit function (equation 6.1) on the generated covariate vectors.

6.3.2 Sensitivity analyses

In the sensitivity analyses, we studied the effect of additional factors on the predictive performance of MLR. In each of these scenarios, EPV_m in the development data sets was fixed to 10, frequencies of outcome categories were equal and the number of predictors was set to 4. The factors that were varied were:

- Correlations between predictors. Levels: $0; 0.2; 0.3; 0.5; 0.7$ and 0.9.

- Type of predictors. Levels: Continuous (standard normal) and binary (with relative frequency 1/2).

6.3.3 Development and validation data sampling procedure

Two-thousand replications per simulation scenario were performed. For each replication a development data set was generated (total sample size per scenario is given in Tables 6.2 - 6.4), as well as an independent (external) validation data set of size $N = 30,000$. On each development data set, MLR models were estimated by ML (section 6.2.1), lasso and ridge (section 6.2.1). For these models, the apparent discrimination predictive performance and apparent overall predictive performance (Table 6.1) were calculated on the development data. Further, the out-of-sample predictive performance (all measures in Table 6.1) of the fitted models were evaluated on the validation data sets. Similar to earlier EPV studies, [326] EPV_m and N were fixed for each simulation data set by sampling covariate and outcome data until these criteria were met, while disregarding oversampled data.

6.3.4 Software

Simulations and analyses were carried out in `R 3.2.2`. [201] For the fitting of ML the `mlogit` [327] and `maxLik` [328] packages were used. For the fitting of ridge and lasso the `glmnet` package was used. [316] In a pilot study (data not shown), the sequence of default λ values generated for ridge MLR showed to be insufficient. This issue was alleviated by extending the sequence with smaller values. The models rarely failed to converge in general. On overall, in $< 0.01\%$ of the main analyses at least one of the models did not converge, whereas in the scenario with highest non-convergence this was 0.3%. In the sensitivity analyses all models converged. Our simulation code and aggregated data are available via GitHub (`https://github.com/VMTdeJong/Multinomial-Predictive-Performance`).

6.4 Results

6.4.1 Calibration

Calibration slopes could not be computed for the lasso in 0.04% of the simulations, when all predictor coefficients were shrunk to exactly zero. We report the results of the two multinomial calibration slopes where category 3 was taken as reference for simplicity of interpretation (Figure 6.1and Table 6.2). The distribution of calibration slopes estimated on the validation data sets was right skewed for some simulation scenarios. This was especially the case for the penalization methods, due to extensive shrinkage of coefficients to values very close to zero in a few simulation replications. Therefore, we report the medians of the calibration slopes as an overall measure of calibration.

As expected, the calibration slopes estimated on the validation data approached 1 (perfect calibration) as EPV_m increased for all methods (Figure 6.1and Table 6.2). Calibration slopes for ML were consistently smaller than 1 for all scenarios

with low EPV_m, demonstrating overfit. For penalized MLR, we observed a different calibration pattern than for ML. In scenarios where both EPV_m and total sample size were low, the calibration tended to be in the opposite direction for the two calibration slopes for the same model. That is, one of the two multinomial calibration slopes tended to be larger than 1 (indicating underfit) while the other tended to be smaller than 1 (indicating overfit). However, both lasso and ridge were on overall better calibrated than ML, as the calibration slopes approached the value of 1 more quickly than for ML. Further, in most scenarios the calibration slopes of ridge MLR approached the perfect value of 1 more quickly than those of lasso MLR.

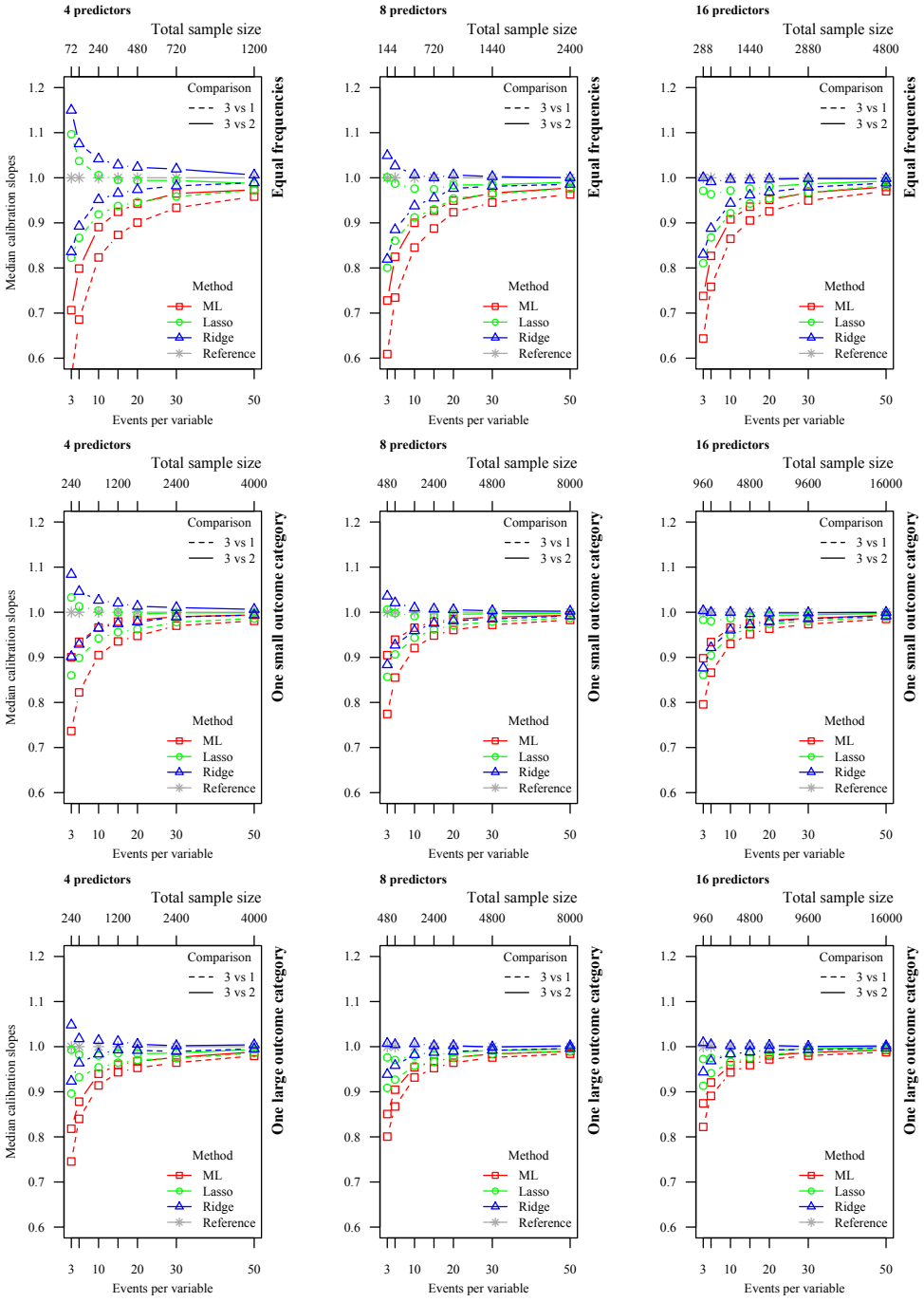
Median calibration slopes for all methods tended to be closer to 1 when there was one large or one small outcome than when the outcome categories were equal in size, when EPV_m was kept constant. Further, the median calibration slopes for one pair of outcomes (categories 3 and 2) improved, when only the remaining outcome category (category 1) increased in size. Additionally, calibration slopes for all methods were closer to optimal as the number of predictors increased, while EPV_m was kept constant. As the number of predictors and the relative frequencies of the outcome categories modify the total sample size, calibration slopes tended to be closer to 1 as the total sample size increased. Finally, calibration slopes were closer to 1 as the model strength of the data generating mechanism increased, as quantified by the reference PDI and Brier scores.

Table 6.2: Median Multinomial Calibration Slopes for ML, Lasso and Ridge.

				Maximum Likelihood		Lasso		Ridge	
RF	R	EPV_m	N	Slope 3 vs 1	3 vs 2	3 vs 1	3 vs 2	3 vs 1	3 vs 2
$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$.	4	3	72	0.556'	0.707'	0.823*	1.097*	0.836*	1.150*
		5	120	0.686*	0.799'	0.867*	1.037*	0.892*	1.075*
		10	240	0.823'	0.891'	0.919*	1.006'	0.951*	1.042'
		15	360	0.873'	0.924'	0.937*	0.995'	0.965'	1.028'
		20	480	0.901'	0.943	0.946'	0.994'	0.974'	1.023'
		30	720	0.933'	0.965	0.959'	0.994	0.982'	1.019
		50	1200	0.958	0.973	0.972	0.987	0.989'	1.006
	8	3	144	0.609'	0.728'	0.801'	1.001'	0.819*	1.049*
		5	240	0.734'	0.825'	0.860'	0.987*	0.885'	1.026'
		10	480	0.845'	0.900'	0.912'	0.976	0.937'	1.006'
		15	720	0.888	0.927	0.930'	0.975	0.955'	1.000
		20	960	0.924	0.949	0.954	0.984	0.976	1.006
		30	1440	0.945	0.966	0.964	0.985	0.981	1.003
		50	2400	0.963	0.977	0.976	0.990	0.985	1.000
	16	3	288	0.644'	0.738	0.811'	0.971'	0.830'	1.000'
		5	480	0.758'	0.827	0.868'	0.964	0.888'	0.991'
		10	960	0.865	0.908	0.921	0.972	0.943	0.996
		15	1440	0.905	0.936	0.943	0.976	0.962	0.995
		20	1920	0.926	0.951	0.954	0.980	0.968	0.997
		30	2880	0.949	0.967	0.967	0.987	0.979	0.998
		50	4800	0.970	0.980	0.983	0.993	0.988	0.998
$\frac{2}{20}, \frac{9}{20}, \frac{9}{20}$.	4	3	240	0.737*	0.900'	0.860*	1.033'	0.901*	1.084*
		5	400	0.822*	0.934	0.899*	1.013'	0.929*	1.046'
		10	800	0.905'	0.966	0.942'	1.004	0.965'	1.026
		15	1200	0.935'	0.979	0.956'	1.000	0.975'	1.020
		20	1600	0.948'	0.983	0.962'	0.996	0.978'	1.013
		30	2400	0.970'	0.990	0.978'	0.999	0.989	1.010
		50	4000	0.981	0.994	0.986	0.999	0.994	1.006
	8	3	480	0.774'	0.905	0.857'	1.006'	0.884'	1.036'
		5	800	0.855'	0.939	0.906'	0.998	0.927'	1.021
		10	1600	0.921'	0.966	0.944	0.991	0.959'	1.009
		15	2400	0.948	0.979	0.962	0.994	0.975	1.007
		20	3200	0.961	0.984	0.971	0.995	0.981	1.006
		30	4800	0.972	0.989	0.979	0.997	0.985	1.003
		50	8000	0.983	0.993	0.988	0.998	0.992	1.002
	16	3	960	0.796	0.898	0.861	0.983	0.876	1.003
		5	1600	0.866	0.934	0.904	0.980	0.921	1.000
		10	3200	0.930	0.966	0.948	0.987	0.960	1.000
		15	4800	0.951	0.976	0.966	0.991	0.972	0.997
		20	6400	0.964	0.982	0.974	0.993	0.979	0.998
		30	9600	0.974	0.987	0.981	0.994	0.986	0.999
		50	16000	0.985	0.993	0.990	0.997	0.992	1.000
$\frac{8}{10}, \frac{1}{10}, \frac{1}{10}$.	4	3	240	0.745*	0.818'	0.896*	0.993*	0.923*	1.048*
		5	400	0.840*	0.878'	0.932*	0.982'	0.964*	1.017*
		10	800	0.914'	0.940'	0.954'	0.980'	0.984'	1.014'
		15	1200	0.943'	0.960'	0.964'	0.985	0.993'	1.012'
		20	1600	0.953'	0.967	0.970'	0.984	0.992'	1.005'
		30	2400	0.965	0.976	0.974	0.986	0.990	1.002
		50	4000	0.980	0.988	0.985	0.994	0.995	1.004
	8	3	480	0.801'	0.851	0.909'	0.976'	0.938'	1.007'
		5	800	0.867'	0.905	0.927'	0.970	0.958'	1.004'
		10	1600	0.932	0.954	0.958	0.980	0.982	1.007
		15	2400	0.952	0.967	0.967	0.981	0.987	1.002
		20	3200	0.964	0.976	0.976	0.988	0.990	1.002
		30	4800	0.976	0.984	0.984	0.992	0.992	0.999
		50	8000	0.984	0.990	0.990	0.995	0.996	1.002
	16	3	960	0.822	0.874	0.913	0.973	0.944	1.009
		5	1600	0.891	0.921	0.941	0.975	0.968	1.004
		10	3200	0.943	0.959	0.966	0.984	0.984	1.002
		15	4800	0.959	0.973	0.975	0.988	0.988	1.002
		20	6400	0.972	0.981	0.983	0.992	0.993	1.003
		30	9600	0.980	0.987	0.989	0.995	0.993	1.000
		50	16000	0.988	0.993	0.993	0.998	0.997	1.002

Each multinomial calibration slope consisted of 2 slopes, where category 3 was taken as reference. RF: relative frequencies of the outcome categories. R: Number of predictors. EPV_m: multinomial events per variable. N: total sample size. SE are obtained by taking the SD of 10^5 bootstraps. SE are indicated as follows: omitted $< .0025 \leq$ ' $< 0.005 \leq$ * ≤ 0.012.

Figure 6.1: Median calibration slopes for ML, lasso and ridge.



Perfect calibration (1) has been included as reference. Horizontal axis: number of predictors varied. Vertical axis: relative frequency varied. Solid lines: category 3 vs 2. Dashed lines: category 3 vs 1.

6.4.2 Discrimination

The values of all out-of-sample PDI (i.e. estimated on validation data) were consistently lower than the within-sample PDI (i.e. estimated on development data), reflecting over-optimism of the within-sample PDI statistic, due to overfitted prediction models (Figure 6.2and Table 6.3). As EPV_m increased, both the within- and out-of-sample PDI approached the true values of the data generating mechanism. In situations where the outcome categories were unequally sized, out-of-sample PDI was better than where outcome categories were equally sized, while EPV_m was kept constant. The PDI of all models improved slightly as the number of predictors increased, while EPV_m was kept constant. Out-of-sample discrimination, as well as within-sample discrimination, were nearly equivalent for ML, ridge and lasso.

6.4.3 Overall performance

The results of the Brier score (Figure 6.3 and Table 6.4) were similar to the results of Nagelkerke R^2 (Figure 1 and Table 1 of Appendix B, https://doi.org/10.1002/sim.8063). The out-of-sample Brier scores were consistently higher than the within-sample Brier scores, again reflecting over-optimism of the within-sample statistics. As EPV_m increased, both the within-sample and out-of-sample Brier score approached that of the data generating mechanism. In situations where the outcome categories were unequally sized, out-of-sample Brier scores were better than where outcome categories were equally sized. Though, the Brier scores were marginally worse as the number of predictors increased.

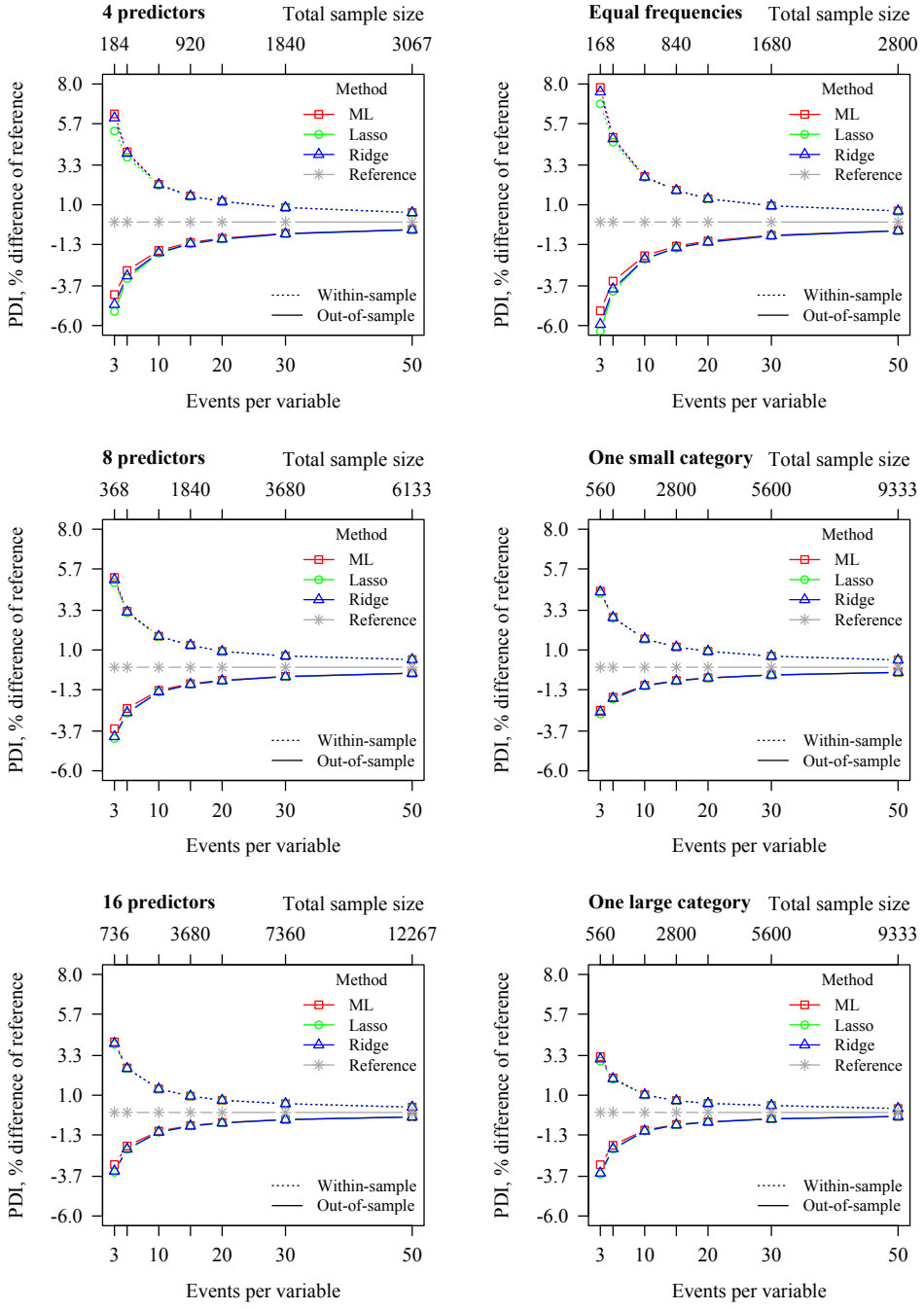
Out-of-sample Brier scores were slightly better for ridge and lasso than for ML in situations with low EPV_m (Figure 6.3 and Table 6.4). Within-sample Brier scores were closer to out-of-sample Brier scores for lasso and ridge than for ML in situations with low EPV_m, reflecting a decrease in optimism of the within-sample statistics, by the application of penalization.

Table 6.3: Percentage Difference between PDI of ML, Lasso and Ridge and the Reference.

					Within-sample			Out-of-sample		
RF	R	Ref.	EPV_m	N	ML	Lasso	Ridge	ML	Lasso	Ridge
$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$.	4	0.59	3	72	9.28*	7.21*	8.98*	-5.72'	-7.36'	-6.47'
			5	120	6.00*	5.37*	5.88*	-3.86	-4.70'	-4.38
			10	240	3.21*	3.07*	3.16*	-2.29	-2.58	-2.53
			15	360	2.18'	2.13'	2.16'	-1.65	-1.82	-1.79
			20	480	1.65'	1.63'	1.64'	-1.32	-1.43	-1.42
			30	720	1.15'	1.14'	1.15'	-0.91	-0.96	-0.96
			50	1200	0.85'	0.85'	0.85'	-0.61	-0.63	-0.63
	8	0.65	3	144	8.00*	7.36*	7.73*	-5.18	-6.22	-6.01
			5	240	4.88*	4.69*	4.80'	-3.51	-4.05	-3.97
			10	480	2.75'	2.72'	2.74'	-1.95	-2.15	-2.13
			15	720	2.00'	1.99'	1.99'	-1.41	-1.51	-1.51
			20	960	1.34'	1.34'	1.34'	-1.10	-1.16	-1.16
			30	1440	0.96	0.96	0.96	-0.78	-0.81	-0.81
			50	2400	0.64	0.64	0.64	-0.50	-0.51	-0.51
	16	0.72	3	288	6.41'	6.09'	6.20'	-4.65	-5.56	-5.43
			5	480	4.04'	3.96'	3.99'	-3.01	-3.43	-3.38
			10	960	2.06	2.05	2.05	-1.68	-1.82	-1.81
			15	1440	1.44	1.44	1.44	-1.18	-1.24	-1.24
			20	1920	1.11	1.11	1.11	-0.92	-0.96	-0.97
			30	2880	0.76	0.76	0.76	-0.64	-0.66	-0.66
			50	4800	0.47	0.47	0.47	-0.40	-0.41	-0.41
$\frac{2}{20}, \frac{9}{20}, \frac{9}{20}$.	4	0.58	3	240	5.20*	4.92*	5.15*	-2.79	-3.03	-2.76
			5	400	3.51*	3.41*	3.50*	-1.98	-2.10	-1.99
			10	800	1.95'	1.93'	1.95'	-1.25	-1.30	-1.26
			15	1200	1.47'	1.46'	1.47'	-0.93	-0.96	-0.94
			20	1600	1.18'	1.18'	1.18'	-0.74	-0.76	-0.75
			30	2400	0.75	0.75	0.75	-0.55	-0.56	-0.56
			50	4000	0.46	0.46	0.46	-0.37	-0.38	-0.38
	8	0.64	3	480	4.42'	4.33'	4.38'	-2.47	-2.66	-2.60
			5	800	2.86'	2.83'	2.85'	-1.75	-1.86	-1.84
			10	1600	1.69'	1.69'	1.69'	-1.05	-1.09	-1.09
			15	2400	1.14	1.13	1.13	-0.77	-0.79	-0.79
			20	3200	0.92	0.92	0.92	-0.62	-0.63	-0.63
			30	4800	0.64	0.64	0.64	-0.44	-0.45	-0.45
			50	8000	0.45	0.45	0.45	-0.29	-0.29	-0.30
	16	0.70	3	960	3.73'	3.70'	3.70'	-2.27	-2.48	-2.46
			5	1600	2.42	2.41	2.41	-1.52	-1.62	-1.62
			10	3200	1.34	1.34	1.34	-0.88	-0.91	-0.92
			15	4800	0.97	0.97	0.97	-0.64	-0.66	-0.66
			20	6400	0.72	0.72	0.71	-0.50	-0.51	-0.51
			30	9600	0.56	0.56	0.56	-0.35	-0.35	-0.36
			50	16000	0.35	0.35	0.35	-0.23	-0.23	-0.23
$\frac{8}{10}, \frac{1}{10}, \frac{1}{10}$.	4	0.62	3	240	4.38*	3.75*	4.05*	-4.08'	-5.11'	-5.05'
			5	400	2.72*	2.54*	2.63*	-2.60	-3.03	-2.98
			10	800	1.45'	1.41'	1.42'	-1.42	-1.53	-1.53
			15	1200	0.91'	0.90'	0.90'	-0.98	-1.03	-1.04
			20	1600	0.76'	0.76'	0.76'	-0.76	-0.79	-0.79
			30	2400	0.63'	0.63'	0.63'	-0.52	-0.54	-0.54
			50	4000	0.33	0.33	0.33	-0.33	-0.34	-0.34
	8	0.70	3	480	3.30'	3.13'	3.22'	-3.06	-3.53	-3.46
			5	800	2.06'	2.01'	2.03'	-1.93	-2.12	-2.10
			10	1600	1.00'	0.99'	0.99'	-1.03	-1.08	-1.08
			15	2400	0.74	0.74	0.74	-0.69	-0.71	-0.71
			20	3200	0.53	0.53	0.53	-0.55	-0.57	-0.57
			30	4800	0.39	0.39	0.39	-0.37	-0.38	-0.38
			50	8000	0.26	0.26	0.26	-0.23	-0.23	-0.23
	16	0.79	3	960	2.27'	2.22'	2.25'	-2.18	-2.41	-2.36
			5	1600	1.37	1.36	1.37	-1.34	-1.43	-1.41
			10	3200	0.73	0.73	0.73	-0.71	-0.73	-0.73
			15	4800	0.48	0.48	0.48	-0.48	-0.49	-0.49
			20	6400	0.34	0.34	0.34	-0.36	-0.36	-0.36
			30	9600	0.24	0.24	0.24	-0.24	-0.25	-0.25
			50	16000	0.12	0.12	0.12	-0.15	-0.15	-0.15

Reference values obtained with the data generating mechanism. All SE of reference $< 10^{-4}$. RF: relative frequencies of the outcome categories. R: Number of predictors. EPV_m: multinomial events per variable. ML: Maximum Likelihood. N: total sample size. PDI: polytomous discrimination index. SE are indicated as follows: omitted $< 0.05 \leq$ ' $< 0.10 \leq$ * ≤ 0.22.

Figure 6.2: Percent difference in PDI between reference and ML, lasso and ridge.



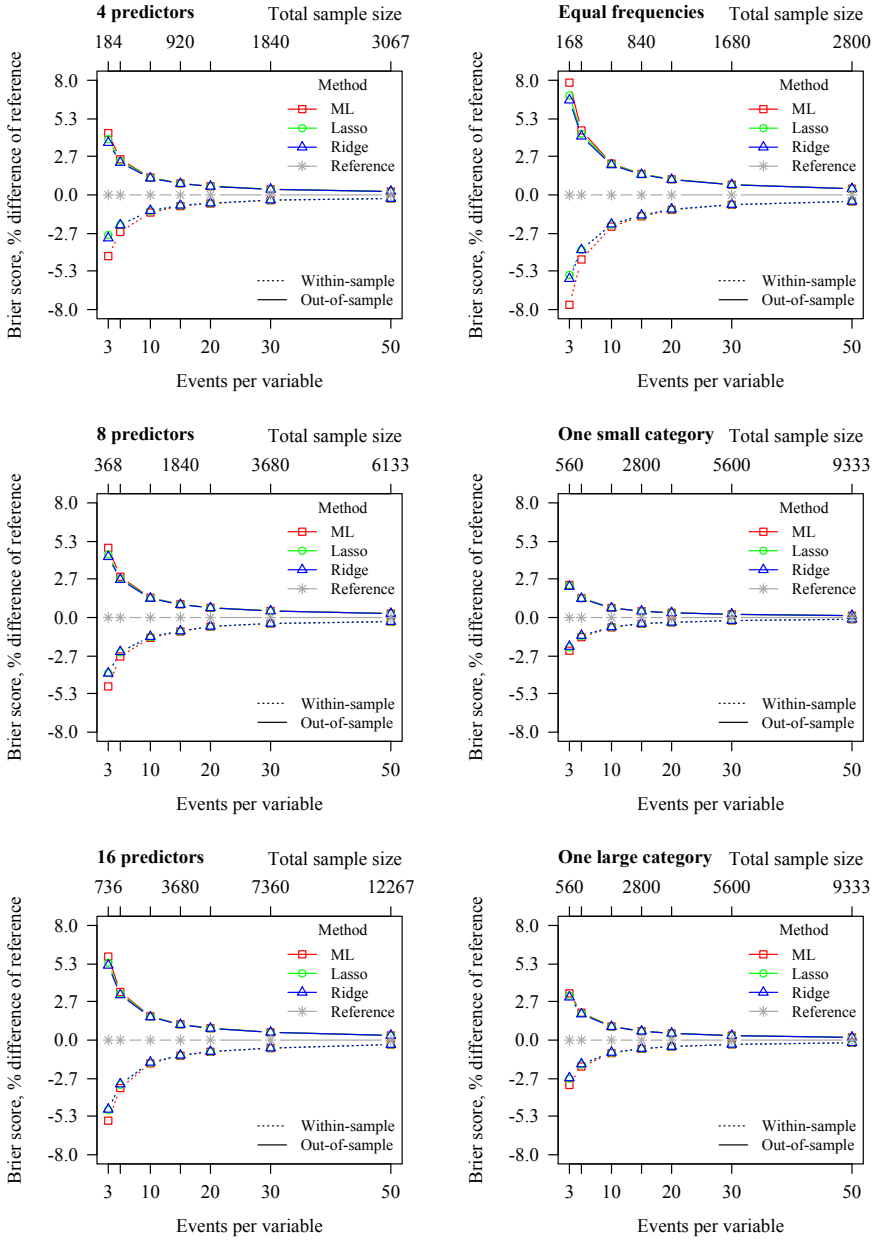
Zero (i.e. no difference with the data generating mechanism) has been included as reference. Left: stratified by number of predictors, frequency marginalized out. Right: stratified by frequency, number of predictors marginalized out. Dotted lines: within-sample PDI. Solid lines: out-of-sample PDI.

Table 6.4: Percentage Difference between Brier scores of ML, Lasso and Ridge and the Reference.

					Within-sample			Out-of-sample		
RF	R	Ref.	EPV_m	N	ML	Lasso	Ridge	ML	Lasso	Ridge
$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$.	4	0.55	3	72	-6.89*	-3.99*	-4.58*	7.01'	6.14'	5.70'
			5	120	-4.14*	-3.14*	-3.26*	4.00'	3.76	3.56
			10	240	-1.96'	-1.75'	-1.73'	1.94	1.90	1.84
			15	360	-1.26'	-1.18'	-1.16'	1.29	1.29	1.25
			20	480	-0.92'	-0.88'	-0.86'	0.97	0.97	0.94
			30	720	-0.59	-0.58	-0.57	0.62	0.62	0.61
			50	1200	-0.43	-0.43	-0.42	0.38	0.38	0.38
	8	0.50	3	144	-7.62*	-5.84*	-5.97*	7.72'	6.75'	6.52'
			5	240	-4.30*	-3.74*	-3.72*	4.47	4.24	4.10
			10	480	-2.28'	-2.16'	-2.13'	2.15	2.12	2.06
			15	720	-1.60'	-1.56'	-1.53'	1.45	1.44	1.40
			20	960	-0.99'	-0.97'	-0.96'	1.06	1.07	1.04
			30	1440	-0.69	-0.68	-0.67	0.72	0.72	0.71
			50	2400	-0.45	-0.45	-0.44	0.43	0.43	0.43
	16	0.44	3	288	-8.70*	-7.29*	-7.29*	9.01'	8.16	7.87
			5	480	-5.19'	-4.75'	-4.69'	5.15	4.92	4.76
			10	960	-2.47'	-2.39'	-2.34'	2.53	2.50	2.45
			15	1440	-1.66'	-1.62'	-1.60'	1.68	1.67	1.64
			20	1920	-1.25	-1.23	-1.21	1.27	1.26	1.25
			30	2880	-0.82	-0.81	-0.81	0.84	0.83	0.83
			50	4800	-0.49	-0.49	-0.49	0.50	0.50	0.50
$\frac{2}{20}, \frac{9}{20}, \frac{9}{20}$.	4	0.42	3	240	-2.07*	-1.74*	-1.66*	1.95	1.96	1.87
			5	400	-1.30'	-1.18'	-1.13'	1.16	1.17	1.14
			10	800	-0.61'	-0.58'	-0.56'	0.57	0.58	0.57
			15	1200	-0.37'	-0.36'	-0.35'	0.37	0.38	0.37
			20	1600	-0.34'	-0.34'	-0.33'	0.28	0.29	0.28
			30	2400	-0.15	-0.15	-0.15	0.19	0.19	0.19
			50	4000	-0.08	-0.08	-0.07	0.11	0.11	0.11
	8	0.36	3	480	-2.28*	-2.09*	-2.03*	2.23	2.19	2.11
			5	800	-1.26'	-1.20'	-1.17'	1.35	1.34	1.31
			10	1600	-0.68'	-0.67'	-0.66'	0.67	0.67	0.66
			15	2400	-0.39	-0.39	-0.38	0.44	0.44	0.44
			20	3200	-0.29	-0.29	-0.28	0.33	0.33	0.33
			30	4800	-0.18	-0.18	-0.18	0.22	0.22	0.22
			50	8000	-0.14	-0.14	-0.14	0.13	0.13	0.13
	16	0.31	3	960	-2.70'	-2.56'	-2.52'	2.77	2.70	2.62
			5	1600	-1.60'	-1.56'	-1.54'	1.63	1.62	1.58
			10	3200	-0.81	-0.80	-0.79	0.81	0.80	0.79
			15	4800	-0.56	-0.55	-0.55	0.54	0.53	0.53
			20	6400	-0.41	-0.41	-0.40	0.40	0.40	0.40
			30	9600	-0.34	-0.34	-0.34	0.26	0.26	0.26
			50	16000	-0.16	-0.16	-0.16	0.16	0.16	0.16
$\frac{8}{10}, \frac{1}{10}, \frac{1}{10}$.	4	0.32	3	240	-2.61'	-2.06'	-2.06'	2.73	2.51	2.42
			5	400	-1.56'	-1.38'	-1.36'	1.59	1.54	1.48
			10	800	-0.76	-0.73	-0.71	0.80	0.80	0.77
			15	1200	-0.48	-0.47	-0.46	0.52	0.53	0.52
			20	1600	-0.39	-0.38	-0.37	0.40	0.40	0.39
			30	2400	-0.28	-0.28	-0.27	0.26	0.26	0.26
			50	4000	-0.18	-0.18	-0.18	0.15	0.15	0.15
	8	0.29	3	480	-3.11'	-2.69'	-2.64'	3.21	3.08	2.97
			5	800	-1.81'	-1.69'	-1.64'	1.90	1.87	1.81
			10	1600	-0.86'	-0.84'	-0.81'	0.93	0.93	0.92
			15	2400	-0.63	-0.63	-0.61	0.61	0.61	0.60
			20	3200	-0.45	-0.45	-0.44	0.47	0.47	0.47
			30	4800	-0.26	-0.26	-0.26	0.31	0.31	0.31
			50	8000	-0.19	-0.19	-0.19	0.19	0.19	0.19
	16	0.26	3	960	-3.80'	-3.45'	-3.35'	3.98	3.87	3.73
			5	1600	-2.25'	-2.15'	-2.08'	2.33	2.30	2.24
			10	3200	-1.14'	-1.12'	-1.10'	1.17	1.17	1.15
			15	4800	-0.74	-0.73	-0.72	0.78	0.77	0.77
			20	6400	-0.55	-0.55	-0.54	0.57	0.57	0.57
			30	9600	-0.38	-0.38	-0.37	0.38	0.38	0.38
			50	16000	-0.19	-0.19	-0.19	0.23	0.23	0.23

Reference values obtained with the data generating mechanism. All SE of reference $< 5 * 10^{-5}$. R: Number of predictors. EPV_m: multinomial events per variable. ML: Maximum Likelihood. N: total sample size. SE are indicated as follows: omitted $< 0.05 \leq$ ' $< 0.10 \leq$ * ≤ 0.18.

Figure 6.3: Percent difference in Brier scores between reference and ML, lasso and ridge.



Zero (i.e. no difference with the data generating mechanism) has been included as reference. Left: stratified by number of predictors, frequency marginalized out. Right: stratified by frequency, number of predictors marginalized out. Dotted lines: within-sample Brier scores. Solid lines: out-of-sample Brier scores.

6.4.4 Sensitivity Analyses

Correlations between predictors

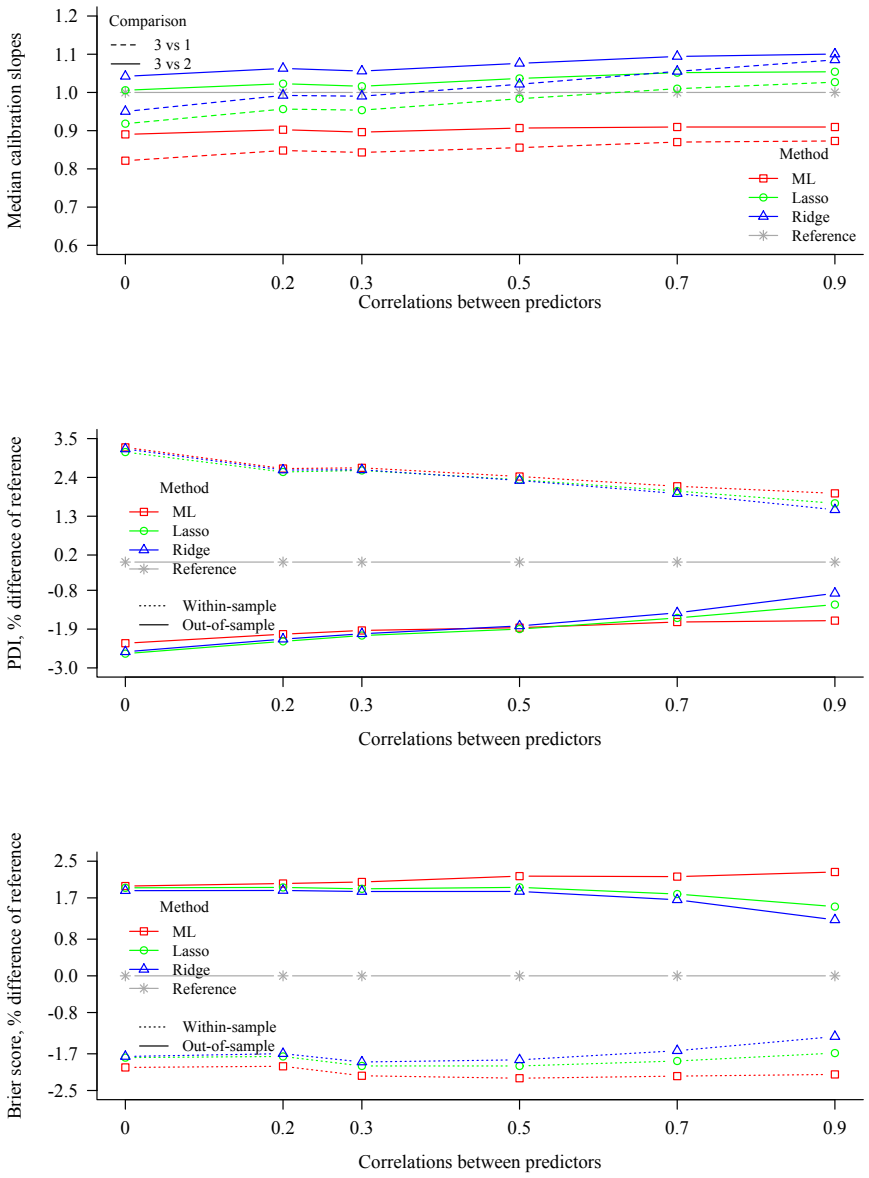
The results of the calibration slopes, PDI and Brier score are shown in Figure 6.4for different values of the correlations between the predictors. For ML, as the correlations between the predictors was increased a small improvement was observed in the calibration slopes and PDI, while the Brier score deteriorated. The calibration slopes increased as the correlations between the predictors increased, for both penalized methods. When the correlations between predictors were very high, both penalization methods yielded underfitted models. The PDI improved for both penalization methods as the correlations between the predictors increased, contrasting with ML, where little difference could be observed. For both penalization methods, the Brier score was better when the correlations between the predictors were very high, as compared to when the correlations were moderate or low. The Brier scores for both penalization methods were superior or equivalent to those for ML, for all values of the correlations between the predictors.

Type of predictors

The results of the calibration slopes, PDI and Brier score are shown for a scenario with continuous and with binary predictors in Figure 6.5. For ML, the calibration slopes were smaller when the predictors were binary, indicating more overfit. Also, the out-of-sample PDI was further from the reference, and the difference with the within-sample PDI was also larger (larger optimism), when the predictors were binary than when they were continuous. We observed barely any difference in the Brier scores between binary and continuous predictors.

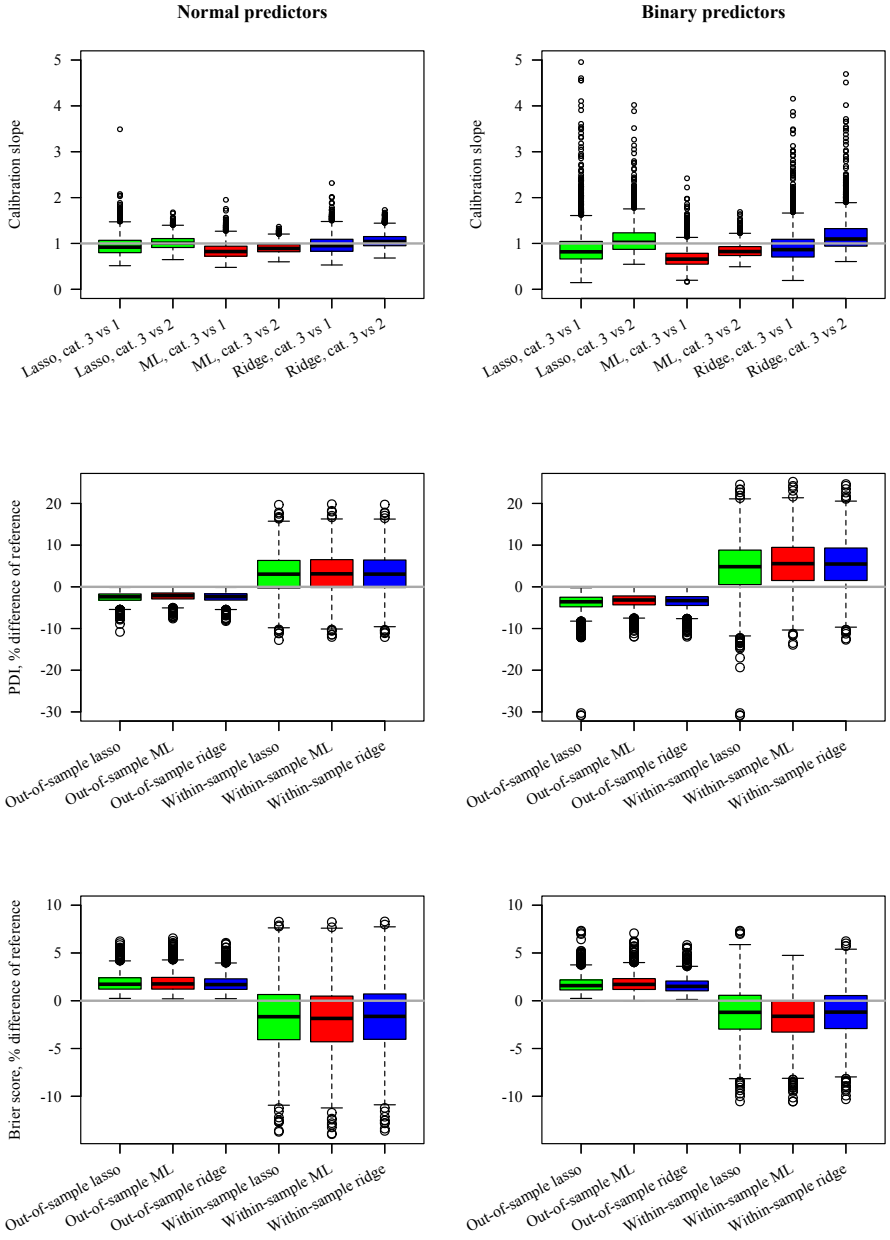
For the penalization methods, the calibration slopes were further away from 1 when the predictors were binary, indicating both more underfit and overfit, than when they were continuous (Figure 6.5). This contrasts with the calibrations slopes for ML, which consistently showed more overfit when the predictors were binary. Similar to ML, the difference between the PDI for the penalization methods and the reference was slightly larger when the predictors were binary. Finally, we observed little difference in the out-of-sample Brier scores for the penalization methods when the type of predictor was varied, similar to ML.

Figure 6.4: Predictive performance for various values of correlations between predictors, for ML, lasso and ridge.



EPV_m = 10, the number of predictors = 4, and the frequencies of the outcome categories are equal, giving a total sample size of 240. Top: median calibration slopes, where 1 is included as reference. Middle: Percent difference in PDI compared with reference. Bottom: Percent difference in Brier score compared with reference. For the PDI and Brier score, zero (i.e. no difference with the data generating mechanism) has been included as reference. Solid and dashed lines: out-of-sample. Dotted lines: within-sample.

Figure 6.5: Predictive performance for ML, lasso and ridge for normal and binary predictors.



$EPV_m = 10$, the number of predictors = 4, and the frequencies of the outcome categories are equal, giving a total sample size of 240. Top: Calibration slopes, where perfect calibration (1) has been included as reference. Middle: Percent difference in PDI compared with reference. Bottom: percent difference in Brier score compared with reference. For the PDI and Brier score, zero (i.e. no difference with the data generating mechanism) has been included as reference. Some extreme values are not shown.

Table 6.5: Prediction Models for Ovarian Tumors

			ML		Lasso (% shrinkage)		Ridge (% shrinkage)	
EPV_m	N	Predictor	Borderline	Invasive	Borderline	Invasive	Borderline	Invasive
3	614	Intercept	−3.96	−5.97	−3.74(5%)	−5.55(7%)	−3.89(2%)	−5.51(8%)
		Age	0.00	0.04	0.00(61%)	0.04(11%)	0.00(−18%)	0.04(9%)
		Solid diameter	0.04	0.10	0.04(3%)	0.09(5%)	0.04(6%)	0.09(8%)
		Papillations flow	1.27	0.52	1.23(3%)	0.47(10%)	1.31(−3%)	0.51(3%)
		Irregular	1.37	0.50	1.21(12%)	0.47(4%)	1.26(8%)	0.49(2%)
		Shadows	−17.95	−4.28	−3.77(79%)	−3.77(12%)	−3.07(83%)	−3.63(15%)
		Ascites	1.77	3.23	1.54(13%)	2.92(10%)	1.51(15%)	2.90(10%)
5	1024	Intercept	−3.87	−5.40	−3.74(3%)	−5.16(4%)	−3.84(1%)	−5.11(5%)
		Age	0.01	0.04	0.01(11%)	0.03(8%)	0.01(−1%)	0.03(6%)
		Solid diameter	0.03	0.09	0.03(2%)	0.09(3%)	0.03(2%)	0.08(6%)
		Papillations flow	1.74	0.88	1.71(2%)	0.83(5%)	1.74(0%)	0.85(4%)
		Irregular	1.02	0.47	0.91(11%)	0.46(2%)	0.96(6%)	0.46(1%)
		Shadows	−2.53	−3.24	−2.53(0%)	−2.96(9%)	−2.18(14%)	−2.91(10%)
		Ascites	1.72	3.20	1.55(10%)	2.99(7%)	1.51(12%)	2.95(8%)
10	2049	Intercept	−3.80	−5.40	−3.70(3%)	−5.22(3%)	−3.78(1%)	−5.15(5%)
		Age	0.00	0.03	0.00(14%)	0.03(6%)	0.00(−9%)	0.03(5%)
		Solid diameter	0.03	0.09	0.03(1%)	0.09(3%)	0.03(0%)	0.08(6%)
		Papillations flow	1.92	1.13	1.89(2%)	1.09(4%)	1.90(1%)	1.09(4%)
		Irregular	1.21	0.56	1.11(8%)	0.55(1%)	1.14(6%)	0.55(1%)
		Shadows	−2.15	−2.87	−2.14(0%)	−2.66(7%)	−1.93(10%)	−2.62(9%)
		Ascites	1.45	2.85	1.31(10%)	2.68(6%)	1.28(12%)	2.66(6%)

The reference category is benign tumors. The models estimated by lasso and ridge have been reparametrized into the reference-category model of equation 6.1. The shrinkage by lasso and ridge is calculated relative to Maximum Likelihood (ML). EPV_m: multinomial events per variable. N: total size of development sample. Age: age in years. Diameter: maximum diameter of solid component (continuous, but no increase > 50 mm). Papillations flow: presence of papillations with blood flow. Irregular: irregular cyst walls. Shadows: presence of acoustic shadows on the echo. Ascites: presence of ascites in the Pouch of Douglas.

6.5 Case study of ovarian cancer

We here present a case study applying penalized and unpenalized MLR to data from a clinical study with the objective to produce a clinical prediction model to predict whether an ovarian tumor is benign ($n = 3183$ or 66%), borderline malignant ($n = 284$ or 6%) or invasive ($n = 1381$ or 28%). The appropriateness of treatment strategies for ovarian tumors depends on the assessment of the tumor using noninvasive procedures, and choosing the most suitable treatment is important as invasive treatments may worsen the prognosis. [329] Candidate predictors were: age (years), presence of papillations with blood flow (yes/no), irregular cyst walls (yes/no), presence of acoustic shadows on the echo (yes/no), presence of ascites in the Pouch of Douglas (yes/no), and maximum diameter of solid component (continuous, but no increase > 50 mm).

For illustrative purposes, we partitioned the data set into disjoint development ($N = 2049, EPV_m = 10$) and validation sets ($N = 2799$). The relative frequencies of the outcome categories were kept constant between development and validation data. Further, we sampled from the development set to obtain two smaller development data sets, sized $N = 1024$ ($EPV_m = 5$) and $N = 616$ ($EPV_m = 3$). We used ML, lasso and ridge to estimate the multinomial logistic regression (MLR) models in the development data sets (Table 6.5). In the $EPV_m = 10$ and $EPV_m = 5$ development sets, the largest shrinkage by penalization we observed was 14%, compared to the model estimated by ML. We observed up to 83% shrinkage in the $EPV_m = 3$ sample.

The developed prediction models were tested in the validation set, thereby quantifying the out-of-sample performance (Table 6.6). We observed that the PDI and Brier scores of the penalized and unpenalized models improved as EPV_m and the total sample size increased, in accordance with the results of our simulations. For $EPV_m = 3$ the model estimated by ML showed overfit, as quantified by the multinomial calibration slopes, whereas the penalized models were close to perfectly calibrated. For $EPV_m \geq 5$ we observed minor miscalibration for all models. Finally, we observed negligible differences in values of the PDI and Brier score between the three models, for each size of the development data, also in accordance with the results of our simulations.

Table 6.6: Performance of Prediction Models for Ovarian Tumors

EPV_m	N	Estimator	slope 3 vs 1	slope 3 vs 2	PDI	Brier score
3	614	ML	0.85	0.71	0.762	0.0759
		Lasso	0.95	0.99	0.762	0.0756
		Ridge	0.97	1.02	0.763	0.0753
5	1024	ML	0.98	0.94	0.767	0.0745
		Lasso	1.03	0.99	0.768	0.0744
		Ridge	1.05	1.01	0.767	0.0743
10	2049	ML	1.01	0.91	0.769	0.0741
		Lasso	1.05	0.95	0.769	0.0740
		Ridge	1.07	0.97	0.768	0.0740

EPV_m: multinomial events per variable. N: total size of development sample. PDI: polytomous discrimination index. ML: Maximum Likelihood. Performance was calculated on an independent sample.

6.6 Discussion

We conducted an extensive simulation study to examine the predictive performance of MLR models that are developed in samples with a ratio of 3 to 50 observations in the smallest outcome category relative to the number of parameters estimated, excluding intercepts. This ratio, which we here call 'multinomial EPV' (EPV_m), is closely related to EPV as known from the binary logistic regression literature. [324, 304] In agreement with earlier studies focusing on binary models, [326, 306, 3] we found that sufficient size of the smallest multinomial category is a factor for the predictive performance of the MLR model. In this study, we have used the definition for EPV_m that most closely matches the EPV definition for binary outcomes. Further research could be focused on other possible EPV definitions. This study has implications for the development of diagnostic and prognostic multinomial prediction models, as it draws the basic outlines of what affects predictive performance in multinomial logistic prediction models in practice.

Our results show that MLR models estimated with ML (i.e. unpenalized) tend to be overfit even in samples with a relatively high number of EPV_m. Overall sample size and the method of analysis, i.e. whether or not shrinkage techniques are applied, are clearly also important factors. The extent of overfit (i.e. model miscalibration) was further affected by the relative sizes of the outcome categories. We observed that calibration was worst when all outcome categories were of equal size, EPV_m was small and the number of predictors was low. When EPV_m is kept constant, model calibration improves as at least one of the outcome categories grows in size, and as the number of predictors increases. In both scenarios, the total sample size also increases. Total sample size is therefore likely an underlying factor affecting model calibration.

Although MLR estimated with ridge and lasso tended to be slightly overfit or underfit (or a combination thereof when one linear predictor was overfit while the other was underfit), these penalized models generally showed better calibration than ML, which in many scenarios showed overfit. Penalization reduces overfit of the estimates by inducing a small bias in the coefficients, which reduces the variance of the estimated probabilities. [313, 330] As overall performance is composed of discrimination and calibration, [3] the improvement in calibration improves the overall performance. Our results indeed showed that the overall performance was slightly better for penalized than for unpenalized MLR, which is in agreement with earlier simulation studies on binary logistic regression. [330, 240]

As noted earlier, in some scenarios lasso and ridge MLR produced models for which one calibration slope was underfit while the other was overfit. This may be a consequence of the (default) parametrization of penalized MLR, which applied only one tuning parameter to two linear predictors. Possibly, $J-1$ tuning parameters are necessary for calibrating penalized models for J categories, such that each slope has its own tuning parameter. Further research is necessary to elucidate this phenomenon.

The conducted sensitivity analyses revealed that the discriminatory performance of unpenalized MLR improved slightly by increased correlations between predictors, though the reference PDI improved as well. Thus, the performance improved as the model strength of the data generating mechanism (the reference) improved. Fur-

ther, the model strength of the data generating mechanism was also affected by the number of predictors and the relative frequencies of the outcome categories. Here we also observe that the calibration and discrimination relative to the reference improved as the model strength of the data generating mechanism increased. Though, note that the Brier scores did not improve compared to the reference as the number of predictors increased.

As the correlations between the predictors increased, the predictive performance of both lasso and ridge improve considerably, though both became underfit when the correlations were very high. When lasso MLR is applied to highly correlated predictors, predictors may be selected randomly and the coefficients of the other predictors may be set to zero.[331] For lasso MLR the effective number of used degrees of freedom is decreased by shrinkage, which can be estimated unbiasedly by the number of predictors retained. [332] Thus, the number of events per effective degrees of freedom for the lasso increases as the correlations between the predictors increase, as the effective number of used degrees of freedom is reduced due to the correlations. This may explain why the predictive performance of lasso MLR improved considerably with increasing correlations.

For ridge MLR, correlations between predictors cause the estimated coefficients to be drawn towards each other by the squared penalty. [310] This stabilizes the estimates, reduces the number of effective degrees of freedom as the coefficients are shrunk [315] and improves the predictive performance. For unpenalized MLR, with predictors specified a priori, the number of effective degrees of freedom equals the number of estimated parameters, regardless of the correlations between the predictors. [315] Hence, for unpenalized MLR the ratio of events per effective degrees of freedom used did not change when the correlation changed, which may explain that little change in predictive performance occurred.

Our sensitivity analyses also show that predictive performance is worse with binary predictors than with continuous predictors, for all methods. This particularly seems to affect calibration. For binary predictors, it is more likely that situations arise where the predictors can (almost) perfectly predict the outcome in the development set, a phenomenon described as 'separation'. [333, 334] In such cases, the unpenalized MLR estimates may attain extreme values, hence the calibration slope of these models will be close to zero in the validation set.

Our simulation study also has some limitations. First, we limited our study to situations where all predictors had non-zero effects (i.e. no noise variables). Our results may therefore not generalize to situations with a large number of noise variables. In a recent simulation study, Pavlou et al. [330] found that penalization improves discrimination for binary logistic prediction models when noise variables are considered. Our results showed little difference in discriminatory performance between penalized and unpenalized MLR. Perhaps, if noise predictors or more weakly predictive variables are considered for MLR, penalized methods could also have better discrimination than unpenalized methods. In our simulation without noise predictors, ridge MLR tended to yield models with better calibration and overall performance than lasso MLR. Though, the relative predictive performance of lasso MLR compared to ridge MLR may improve when the number of noise variables increases, as has recently been shown for binary logistic regression. [242]

Table 6.7: Guidance and recommendations

- Predictive performance gradually improves as the number of multinomial EPV (EPV_m) increases, at least until 50 EPV_m.
- Higher EPV_m may be necessary when the event rates are equal, than when the smallest category is rare.
- Interpret (penalized and unpenalized) models with caution when estimated with $EPV_m < 10$.
- Use penalized methods for best predictive performance.
- Correct for optimism, as within-sample performance measures are overly optimistic.

Additionally, we only considered MLR for three outcome categories in our study, which is the simplest extension of the binary logistic model. When the number of outcome categories is increased and the number of EPV_m is kept constant, the total sample size increases. As our study showed that predictive performance tends to improve with increasing total sample size, we anticipate that a larger number of outcome categories will yield better overall predictive performance for the same number of multinomial EPV. Furthermore, future research on the interaction between the number of outcome categories and their distribution on predictive performance is warranted.

Our results are in agreement with other reports that the adequate sample size for a prediction model is not simply given by the number of EPV. [239, 335, 242] Instead, prediction model performance is related to both EPV and total sample size. Thus both should be considered when developing a prediction model. However, based on our findings, some general recommendations for MLR prediction model development can be given, which are summarized in Table 6.7. We believe that the penalization methods (lasso and ridge) are applicable for MLR even for large samples, albeit the added value of penalization in terms of predictive performance decreases with increasing EPV_m and total sample size. For samples with EPV_m 30 or lower we advise that the total sample size be taken into consideration. When the total sample size is large, reasonable predictive performance may be attained with 10 EPV_m. Conversely, when the total sample size is low, predictive performance can be poor if EPV_m is 10. Below 10 EPV_m a MLR model is at risk of being seriously miscalibrated. Penalization and optimism corrections for ≤ 10 EPV_m are highly recommended.

Acknowledgements

We thank Hajime Uno for providing code for the PDI.

Chapter 7

General discussion

Adapted from

Valentijn M.T. de Jong*, Thomas P.A. Debray*, Karel G.M. Moons, Richard D. Riley. Evidence synthesis in prognosis research. *Diagnostic and Prognostic Research.* 2019 Jul 11;3(1):13. DOI: 10.1186/s41512-019-0059-4

* Contributed equally

In this thesis, we have investigated methods for performing an individual participant data meta-analysis (IPD-MA) in prediction model research. We aimed to develop and evaluate methods that allow for enhanced development and validation of prediction models that are more reproducible and better transportable to other settings and populations. The main findings of this thesis are:

- Chapter 2 summarizes available methods for conducting an IPD-MA of randomized intervention studies with time-to-event outcomes. We focused on modeling frailty of trial participants across trials, modeling heterogeneity of intervention effects, choosing appropriate association measures, dealing with (trial differences in) censoring and follow-up times, and addressing time-varying intervention effects and effect modification (interactions). We discuss how to do this using either parametric or semi-parametric methods and how to implement these approaches in a one-stage or two-stage IPD-MA framework. These methods form the foundation of modeling to predict the survival time in new participants and to predict the intervention effect for individual participants.

- Chapter 3 illustrates the use of Stepwise Internal-External Cross-Validation (SIECV) to assess and reduce heterogeneity in a model's predictive performance. This method allows for the development of prediction models that are more robust and require less tailoring when applied to different settings and populations. We propose a predictor selection algorithm that optimizes the (weighted) average performance whilst minimizing its variability across the hold-out clusters (or studies). Our methodology may improve the generalizability of developed models to different settings and populations and reduce the need for tailoring the model to local circumstances.

- Chapter 4 describes propensity-score methods for standardizing IPD from multiple data sets for prediction model validation purposes. The performance of a developed prediction model may deteriorate in a model validation study due to differences in patient characteristics or regression coefficients between the development and validation samples. By weighting samples towards a specific target population, we can provide more precise estimates of reproducibility and enhance the interpretation of validation study results. We illustrate how samples that are poorly representative (e.g. due to the choice of eligibility criteria, as is commonly the case in RCT data) can be standardized with respect to a specific target population.

- Chapter 5 describes the impact of misclassification of predictors in an IPD-MA. The presence of misclassification may introduce bias in estimates of prediction model parameters, even when the error is entirely random. We developed Bayesian statistical methods for addressing such misclassification, where the extent and nature of measurement error may vary across studies and may depend on participant characteristics. With these methods one can facilitate unbiased estimation of unadjusted and adjusted predictor effects, as well as approximately unbiased estimates of between-study heterogeneity, as is shown in our simulation and motivating example on the diagnosis of dengue.

- Chapter 6 addresses the predictive performance and necessary sample sizes for Multinomial Logistic regression (MLR) prediction models. The use of these models has been advocated when three or more unordered outcomes need to be predicted. Unlike Binary Logistic Regression, the sample size necessary for developing an MLR prediction model had not yet been investigated. We highlight the importance of both the number of outcome events per candidate predictor and the total sample size when determining the necessary sample size in the multinomial prediction modeling context. We recommend the use of penalized MLR when prediction models are developed in small data sets, or in medium sized data sets with a small total sample size (i.e. when the sizes of the outcome categories are balanced).

We continue this final chapter with an overview and discussion on evidence synthesis methods in prognostic prediction research and finish with summary points on prediction research in general.

7.1 Evidence synthesis in prognosis research

Thorough and systematic appraisal of the existing evidence has become mainstream in medical research and practice [336, 337]. Over the past few decades, meta-analysis has become the *de facto* statistical method for summarizing the results from a systematic review and appraisal of existing data on a certain topic. In meta-analysis, estimates of interest (e.g. for a specific treatment effect [42] or diagnostic test-outcome association) are obtained from individual studies and then combined into a weighted average. Such quantitative data synthesis potentially increases statistical power to detect genuine associations or effects, to investigate sources of variation within and across studies, and to answer questions that were not posed by individual studies [338, 339].

Meta-analysis is commonly applied in the domain of randomized therapeutic intervention studies [42] and, more recently, in that of diagnostic test accuracy studies. In the current era of personalized or precision medicine, the use of prognostic information is considered increasingly important to predict outcomes of individuals (off or on treatment) in order to make tailored treatment decisions [5, 340, 341, 264, 342, 343]. It therefore seems timely to apply meta-analytic approaches that allow the quantitative synthesis of prognostic evidence [344].

Key barriers of quantitative synthesis of data from prognosis studies are, among others, the lack of high quality data often due to poor reporting, lack of uniformity in statistical analysis across studies, lack of agreement on relevant statistical measures, and lack of meta-analytical guidance for synthesis of prognosis study data. Recently much guidance has been written on how to define a review question [345], define the PICOTS (Patients, Index prognostic factor or model, Comparator factor or model, Outcomes, Timing of prognostication, Setting of prognostication), define the search strategy, design the data extraction list [346], and do risk of bias assessments [346, 347]. However, there is relatively little guidance on how to do the actual meta-analysis of results from prognosis studies.

In this paper, we discuss how the data or prognostic results from individual

studies, routine care sources (e.g. hospital records or registries), and biobanks can be combined quantitatively. Hereto, we describe statistical methods for the meta-analysis of aggregate data (AD), individual participant data (IPD), or a combination thereof. The aim of this gentle overview is to inform researchers of available methods for synthesis of data of prognostic factor and prognostic model studies, and to encourage their use when individual studies fail to provide generalizable evidence, as we wish to highlight recent advances in these fields.

7.2 Quantitative synthesis in prognostic factor research

Estimates of overall prognosis (e.g. population outcome risk) are rarely sufficient to inform treatment recommendations and individual patient management. For this reason, it is often helpful to distinguish groups of people with a different average prognosis [5, 340]. A common approach is to identify specific factors that, among people with a given starting point (such as diagnosis of disease), are associated with a subsequent endpoint [341]. This generally requires estimation of a factor-outcome association which can, for instance, be quantified using a hazard ratio or an odds ratio [341].

Several meta-analysis methods can be used to generate summary estimates of the association between a prognostic factor and a certain outcome. Although it is fairly straightforward to summarize crude (i.e. unadjusted) estimates of a particular factor-outcome association, this practice is generally discouraged because in practice hardly any prognostication is done based on a single factor only [4, 348]. For this reason, we here focus on meta-analysis methods to summarize the adjusted estimates of a certain prognostic factor and outcome. An overview of the presented methods is provided in Table 7.1.

7.2.1 Meta-analysis of prognostic factor estimates using aggregate data (AD)

A relatively simple situation arises when the prognostic factor of interest is unadjusted in all studies or has been adjusted for the same other prognostic factors (co-variates) in all studies. Traditional meta-analysis methods – as used in meta-analysis of intervention studies – can then be used to summarize the corresponding AD [349]. The most well known approach, also from other types of meta-analysis, is the so-called fixed effect meta-analysis approach, which can be formulated as follows [350, 351]:

$$\hat{\theta}_i \sim \mathcal{N}\left(\mu, \hat{s}_i^2\right) \tag{7.1}$$

where $\hat{\theta}_i$ is the estimated factor-outcome association (e.g. log hazard ratio) from the i^{th} study, with an estimated standard error $\hat{s}_i$. This approach yields a summary estimate of the prognostic effect (μ), which simply represents a weighted average of the $\hat{\theta}_i$s.

A common interpretation of fixed effect meta-analysis is that the *true* factor-outcome association is identical for all studies (i.e. $\theta_i = \mu$). In practice, however, true values for factor-outcome associations are likely to vary across studies due to differences in, e.g., study design, follow-up, variable definitions, adjustment factors, settings and health care standards. It may therefore be more reasonable to assume that the factor-outcome associations θ_i are unrelated, and to adopt a fixed effects meta-analysis [352]. In this approach, the weight for each study is proportional to both the number of study participants, and to how much information is contributed per subject. The meta-analysis then produces an average effect applicable to an amalgamation of the contributing study populations.

Finally, a third option is to adopt a so-called random effects meta-analysis approach, which assumes that the factor-outcome associations θ_i are different but related across studies. A major advantage of this approach is that the presence of between-study heterogeneity can directly be quantified [350, 351]:

$$\hat{\theta}_i \sim \mathcal{N}\left(\mu, \tau^2 + \hat{s}_i^2\right) \tag{7.2}$$

The random effects model includes an additional parameter τ representing the (unknown) between-study standard deviation. The overall summary result (μ) now represents the average (mean) prognostic effect of the factor across the studies.

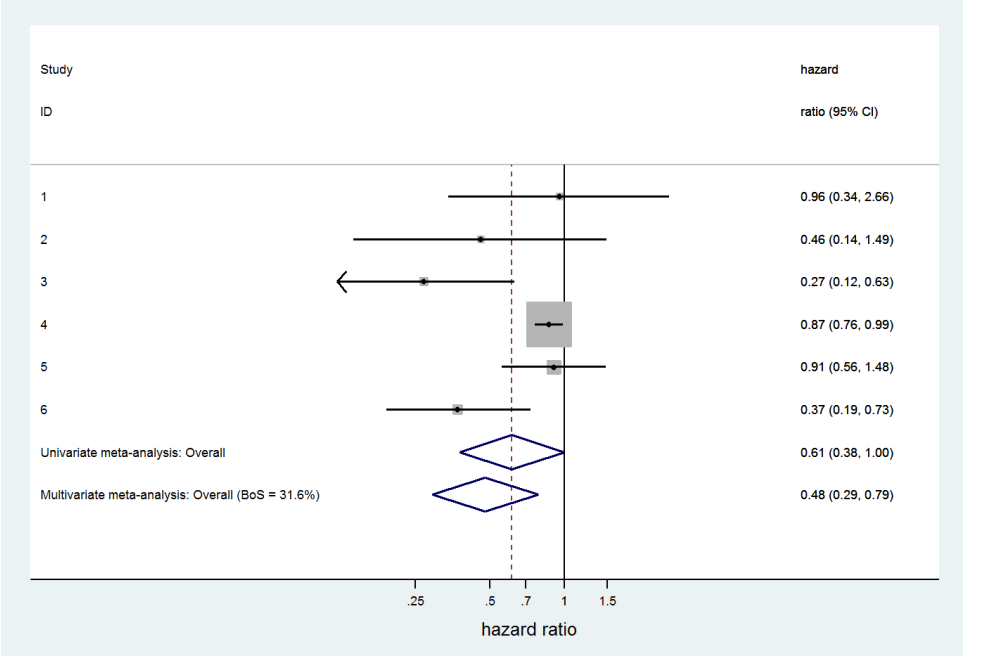
Several methods exist for estimating the weighted average μ and the between-study standard deviation τ [118, 119]. One approach is to estimate μ and τ simultaneously, e.g. by adopting (restricted) maximum likelihood estimation. Alternatively, it is possible to first estimate τ and then use the corresponding value to obtain an estimate for μ. When this strategy does not take the uncertainty of τ into account, confidence intervals for μ may become too narrow [109]. For this reason, it is generally recommended to adjust these intervals using the methods proposed by Hartung and Knapp [126], and Sidik and Jonkman [125].

As an example, investigated the prognostic effect of progesterone receptor status in cancer-specific survival in endometrial cancer [353]. Aggregate data from 6 studies were pooled using a random effects meta-analysis (Der Simonian and Laird method), yielding a summary hazard ratio of 0.62 and a corresponding 95% confidence interval (95% CI) ranging from 0.42 to 0.93. When adopting restricted maximum likelihood estimation, the summary estimate changed to 0.61 with a 95% CI from 0.38 to 1.00 (Figure 7.1). The wider CI is due to a larger estimate of τ when using restricted maximum likelihood estimation rather than DerSimonian and Laird.

Multivariate meta-analysis

Whereas traditional meta-analysis methods are applied to summarize multiple estimates of a single parameter, it is also possible to jointly summarize multiple estimates of two (or more) parameters using so-called bivariate (or multivariate) meta-analysis methods [179, 354, 351]. These methods are well known in the meta-analysis of diagnostic test accuracy, where one jointly estimates the sensitivity and specificity of the test under review [355]. Multivariate meta-analysis methods aim to account for the correlation between the different parameter estimates and can

Figure 7.1: Forest plot for prognostic effect of progesterone on cancer specific survival in endometrial cancer, with summary results for univariate and multivariate meta-analysis



The multivariate meta-analysis of cancer specific survival and progression-free survival used the approach of Riley et al. to handle missing within study correlations, through restricted maximum likelihood estimation [176]. Heterogeneity was similar in both univariate and multivariate meta-analyses ($I^2 = 70\%$).

therefore be used to deal with situations where two or more correlated parameters/statistics are to be synthesized per study. The (bivariate) random effects model for jointly summarizing the AD for two parameters of interest is given as follows:

$$\begin{pmatrix} \hat{\theta}_{1i} \\ \hat{\theta}_{2i} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \tau_1^2 & \rho\tau_1\tau_2 \\ \rho\tau_1\tau_2 & \tau_2^2 \end{pmatrix} + \begin{pmatrix} \hat{s}_{i1}^2 & \hat{r}_i\hat{s}_{i1}\hat{s}_{i2} \\ \hat{r}_i\hat{s}_{i1}\hat{s}_{i2} & \hat{s}_{i2}^2 \end{pmatrix} \right) \tag{7.3}$$

where $\hat{r}_i$ and ρ represent the (estimated) within-study and, respectively, the (unknown) between-study correlation coefficients. For example, $\hat{\theta}_1$ and $\hat{\theta}_2$ may be the prognostic effect on outcome 1 and outcome 2, respectively.

A common application of multivariate meta-analysis arises when researchers are interested in a prognostic factor's association with multiple outcomes [179]. For instance, in the endometrial cancer example, the unadjusted hazard ratio (HR) of progesterone was estimated for cancer specific survival (6 studies) and for progression-free survival (11 studies). The corresponding hazard ratios of the 17 studies were

then jointly pooled using a bivariate random effects meta-analysis [179]. As illustrated in Figure 7.1, this strategy yielded a different and more precise summary estimate of cancer-specific survival (unadjusted HR=0.48, 95% CI: 0.29 to 0.79) as compared to the univariate meta-analysis approach above (unadjusted HR=0.61, 95% CI: 0.38 to 1.00).

Multivariate meta-analysis can also be used to jointly summarize prognostic factor-outcome associations that have been adjusted for different sets of prognostic factors (covariates). Researchers then need to distinguish between estimates that are adjusted for all relevant covariates and estimates that are only adjusted for some (but not all) of the relevant covariates.

Unfortunately, the within-study correlations $\hat{r}_i$ are rarely reported, thereby complicating the multivariate meta-analysis approach. Riley previously demonstrated that simply ignoring these correlations can lead to meta-analysis results with inferior statistical properties [356]. Researchers may therefore assume a common within-study correlation (e.g. $\hat{r}_i = 0$ for all studies), recover its magnitude from reported summary statistics [357], or replace all within- and between-study correlations by an overall correlation parameter that is estimated from the AD at hand [176].

Other meta-analysis approaches

Several extensions for AD meta-analysis of prognostic factor studies have been proposed and can be used to explore sources of between-study heterogeneity [36, 351], to combine studies with different methods of measurement [358], or to combine studies that categorized continuous factors [358, 359, 360].

7.2.2 Meta-analysis using individual participant data (IPD)

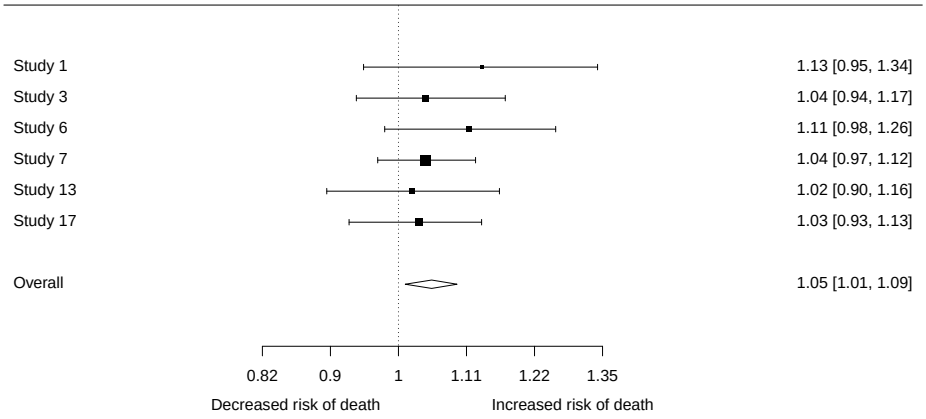
When IPD are available from multiple prognostic factor studies, various random effects meta-analysis models are possible that employ a one-stage or two-stage approach [361, 42, 82].

Two-stage meta-analysis

In the two-stage approach, each study is first summarized by its factor-outcome association estimate and standard error. These AD are then appropriately combined across studies into a summary effect using traditional meta-analysis methods. For instance, Trivella *et al.* performed a two-stage IPD-MA to investigate the role of angiogenesis as a prognostic factor in patients with non-small-cell lung carcinoma [362]. They estimated the log hazard ratio of microvessel-density counts for each participating study center, adjusted for age and cancer stage. These estimates were then pooled using random effects inverse-variance meta-analysis (Figure 7.2).

The two-stage IPD-MA approach can also be used to summarize the association of non-linear prognostic factors [363, 364]. In the first stage, the factor-outcome association of interest is modeled separately for each study with a certain functional form (e.g. cubic spline) and parameterization (e.g. location of knots). An overall function can then be obtained in the second stage by meta-analysing the study-specific function values for distinct factor values [363, 364].

Figure 7.2: Meta-analysis of multivariable predictor effects



Association between risk of death and increase of one microvessel count, as measured by the Chalkley method. Estimates represent multivariable hazard ratios, adjusted for age and cancer stage. [362]

For instance, Sauerbrei *et al.* combined IPD from nine population-based registries to study the prognostic effect of age in T1-2 breast cancer patients [363]. They estimated a Cox regression model separately in each registry and adjusted for 5 to 10 other prognostic factors such as the type of surgery and radiotherapy. Studywise selected fractional polynomials (FP) were used to model the adjusted effect of age. The resulting FP functions were then averaged pointwise, with weights for each registry depending on the variance of the the log relative hazard at distinct age values. Results indicated that the mortality risk is low for women between about 40 and 65 years, and increases outside this range.

Multivariate (two-stage) meta-analysis

Also for IPD meta-analysis, it is possible to simultaneously analyze multiple outcomes by adopting multivariate meta-analysis methods. This typically involves a two-stage approach where the IPD of each study is first reduced to AD (including estimates of the within-study correlation) and subsequently pooled across studies. Multivariate meta-analysis methods have, for instance, been proposed to summarize the association of (non-linear) continuous markers [365]. In the first stage, a common function (e.g. spline with a common location and number of knots for all studies) is estimated separately in each study. The resulting AD (e.g. multivariable regression coefficients) are then pooled across studies in the second stage. In contrast to univariate pooling of estimated effects on a grid of exposure values [363], a

major advantage of this approach is that it better accounts for correlations, thereby decreasing bias and improving precision.

One-stage meta-analysis

An alternative approach for IPD meta-analysis (IPD-MA) of prognostic factor studies is a one-stage approach which synthesises the IPD from all studies in a single step, whilst accounting for clustering of patients within studies [366, 97]. The estimation of a pooled factor-outcome association then involves the fitting of a mixed effect model, where each parameter (e.g. regression coefficient) can be specified as common, random or independent (fixed) across studies. One-stage methods appear particularly advantageous when few studies or few patients per study are available [361], or when studies involve time-to-event outcomes [48, 114].

For instance, Den Ruijter *et al.* performed a one-stage meta-analysis using IPD from 14 cohorts to estimate the association between (log-transformed) carotid intima-media thickness (CIMT) and the incidence of first-time myocardial infarction or stroke [367]. They first assessed between-study heterogeneity by estimating statistical interaction between cohort and CIMT measurements. Subsequently, a multivariable Cox proportional-hazards model was fitted with random effects for the baseline hazard and common effects for the regression coefficients.

When adopting a one-stage approach, it is generally recommended to account for potential ecological bias [36]. This bias may, for instance, arise when patient outcomes are associated with the mean value of the prognostic factor, rather than the individual covariate values. Ecological bias can be mitigated by separating the within-study and across-study associations, as described elsewhere [98].

7.2.3 Meta-analysis using IPD and AD

Although IPD meta-analyses are generally considered as the gold standard, IPD cannot always be obtained from all relevant studies. To avoid (data availability) bias, it is often helpful to supplement the available IPD with AD for those studies where IPD are not available [192]. This strategy can be implemented using the approaches described below, assuming suitable AD can be obtained from the non-IPD studies.

Two-stage meta-analysis

A simple approach is to generate AD from each available IPD set, and to jointly summarize the newly derived (from IPD studies) and previously published AD (from non-IPD studies) using aforementioned meta-analysis methods for AD [192]. When critical information from the non-IPD studies is missing (e.g. within-study correlations), the IPD studies can be used to derive the relevant statistics, thereby reducing the risk of bias in summary estimates [368, 233, 356, 358].

A specific situation arises when the non-IPD studies provide factor-outcome associations that are not adjusted for all relevant covariates. A two-stage bivariate meta-analysis can then be used to combine these partially adjusted estimates with the (fully and partially adjusted) factor-outcome associations from the IPD studies.

The adaptation method

As mentioned earlier, it is common that AD studies do not adjust for all relevant covariates, and only provide factor-outcome associations that are partially adjusted. An alternative method to combine fully adjusted associations with the partially adjusted ones is to use the difference in value between the corresponding regression coefficient(s) [369, 245]. This difference is first estimated in the IPD at hand, and then applied to the summary estimate of the partially adjusted factor-outcome association. The adaptation method has, for instance, been applied in a study investigating risk factors for Methicillin-resistant Staphylococcus aureus acute bacterial skin and skin structure infections [370]. The study authors conducted a literature review to retrieve unadjusted odds ratios for 7 potential risk factors. These odds ratios were then summarized for each risk factor using a random effects meta-analysis, and *adapted* into an adjusted odds ratio using the IPD at hand.

The adaptation method is strongly related, and in some situations even equivalent, to the aforementioned two-stage meta-analysis approach [371]. Although formal comparisons are lacking, it has been argued that the adaptation method may be less statistically and computationally efficient.

Hierarchical-related regression

This one-stage approach directly combines the available IPD and AD by specifying a distinct likelihood for each data source [98, 36]. This enables the IPD studies to contribute in all parameter estimates, whereas the AD studies are only used to estimate the study-level parameters and across-study relationships. For example, Riley and Steyerberg adopted hierarchical-related regression to investigate the relationship between age and the risk of 6-month mortality in patients with traumatic brain injury (TBI) [36]. They used a Bernoulli distribution to model the binary outcomes from 4 IPD studies, and a Binomial distribution for the observed event counts in 10 AD studies. To account for potential ecological bias, the within-study and across-study effects for participant age were separated when jointly analyzing the 14 studies. It was found that an individual's probability of death by 6 months increases as their individual age increases and also as the mean age in their study (or population) increases. A possible explanation for this is that studies with a higher mean age involved clinicians with less experience of treating TBI patients.

Table 7.1: Available methods for quantitative synthesis in prognostic factor research

	Available data	Estimate of interest	Possible methods for evidence synthesis
AD	Baseline characteristics	Linear FOA	Meta-regression [36]
	Similarly adjusted FOAs	Linear FOA	Univariate meta-analysis [350], Multivariate meta-analysis [351, 176, 372]
		Non-linear FOA	Univariate meta-analysis [360, 359], Multivariate meta-analysis [358]
	Not similarly adjusted FOAs	Linear FOA	Multivariate meta-analysis [371, 176, 372]
IPD		Linear FOA	One-stage meta-analysis [361, 36], Two-stage meta-analysis [361], Multivariate meta-analysis [371, 361], Graphical meta-analysis [373]
		Non-linear FOA	One-stage meta-analysis [363, 36], Two-stage meta-analysis [363, 360], Multivariate meta-analysis [365]
IPD + AD	Baseline characteristiscs	Linear FOA	Hierarchical-related regression [36]
		Non-linear FOA	Hierarchical-related regression [36]
	Similarly adjusted FOAs	Linear FOA	Two-stage meta-analysis, Hierarchical-related regression [98]
		Non-linear FOA	Two-stage meta-analysis [360], Hierarchical-related regression [36]
	Not similarly adjusted FOAs	Linear FOA	Multivariate meta-analysis [371, 368], Adaptation method [369, 245]

FOA = factor-outcome association.

7.2.4 Summary points

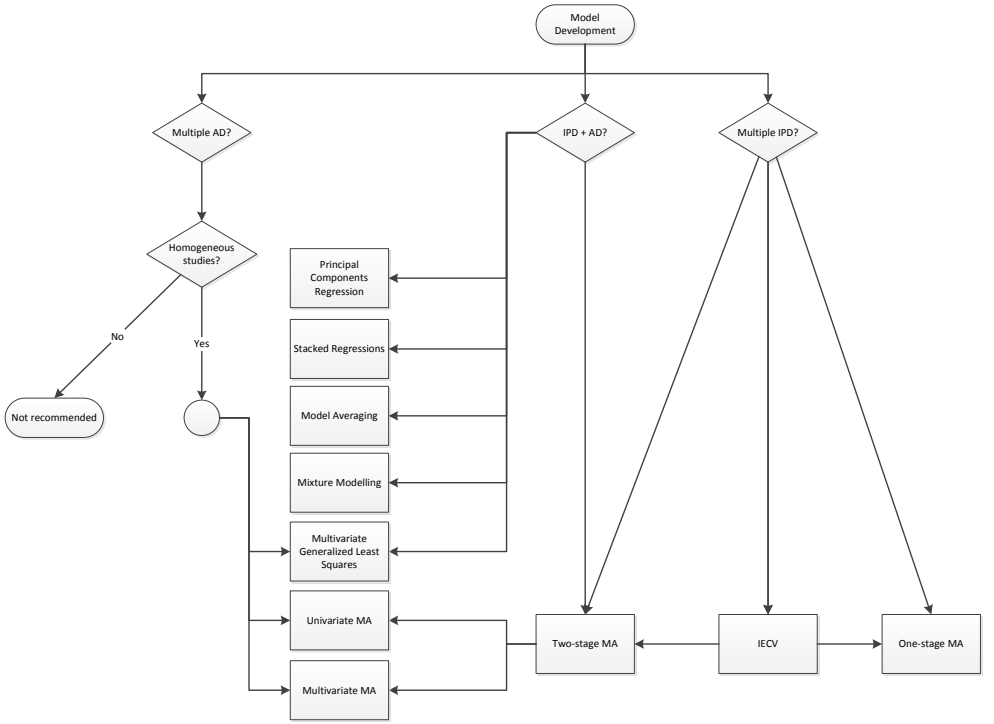
Evidence synthesis in prognostic factor research may help to identify factors that are associated with a certain clinical outcome, to explore their functional form and to quantify their incremental value over established prognostic factors [341]. When IPD are unavailable, traditional meta-analysis methods can be used to summarize published prognostic factor estimates in order to identify genuine prognostic factors [349]. Although IPD are not strictly required to assess the incremental value of a prognostic factor or to explore its functional form, this may often be unfeasible using published AD only [366]. For this reason, when IPD are available for a few studies, corresponding information can be used to restore unreported AD (e.g. missing within-study correlations) or to adapt unadjusted factor-outcome associations. Evidence synthesis in prognostic factor research is, however, most appealing when multiple sources of IPD are available, as this allows to derive desired prognostic factor results directly, and to analyze continuous factors more appropriately [341]. Meta-analysis of IPD is preferably initiated using a two-stage approach, as corresponding methods are relatively straightforward to implement and guard against ecological bias. One-stage meta-analysis methods may, however, be more appealing when few studies or few subjects per study are available, as they are more flexible, resistant against small sample bias, and avoid the need for estimating correlations between random effects [361].

7.3 Quantitative synthesis in prognostic model research

Prognostic model research aims to examine multiple prognostic factors in combination [5], in order to predict the absolute risk of future outcomes in single individuals. Corresponding studies may derive new prognostic models (so-called development studies), evaluate the performance of existing models in new individuals (so-called validation studies) and if necessary tailor their predictions, or examine the model's impact on health-related outcomes.

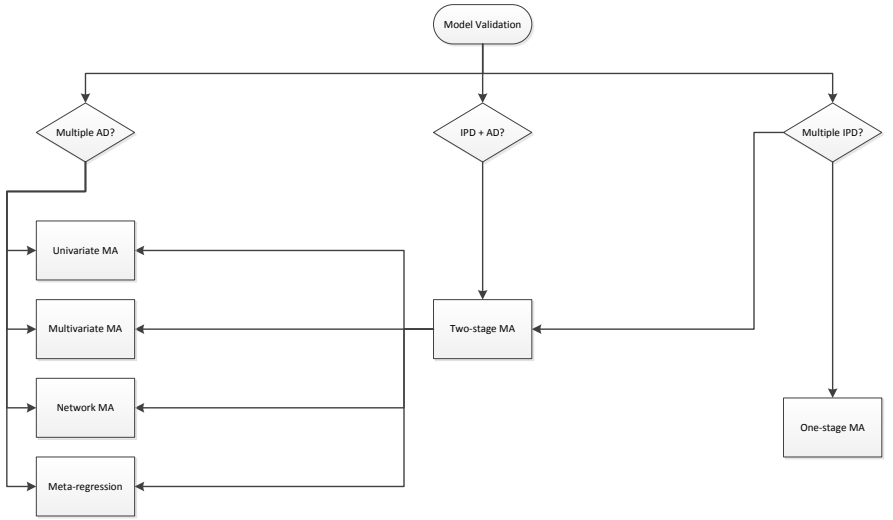
Currently, most prognostic models are developed based on relatively small studies. Hence, many of these models do not perform adequately when applied to other individuals [11, 264, 9, 22]. To investigate and improve the performance of prognostic models across different settings and populations, researchers may consider meta-analysis methods during their development and validation [25, 374, 5, 20, 26, 221]. Several strategies for this purpose are described below and summarized in Figures 7.3 and 7.4. As before, we distinguish between situations where the available data sources comprise of aggregate data, individual participant data, or a combination of both.

Figure 7.3: Available methods for quantitative synthesis during prognostic model development



Abbreviations: MA, meta-analysis; IECV, internal-external cross-validation; AD, aggregate data; IPD, individual participant data

Figure 7.4: Available methods for quantitative synthesis during prognostic model validation



Abbreviations: MA, meta-analysis; AD, aggregate data;
IPD, individual participant data

7.3.1 Meta-analysis using aggregate data (AD)

Validation of an existing prognostic model

A common source of AD are so-called external validation studies assessing the (discrimination and calibration) performance of a certain prognostic model when tested in other individuals than from which the model was developed. By summarizing these performance estimates, it becomes possible to identify whether the model's predictions are sufficiently accurate across different settings and populations. This typically requires the retrieval of multiple performance statistics (e.g. concordance statistic, calibration-in-the-large, calibration slope) and corresponding standard errors [21, 234]. The resulting estimates can then be pooled using traditional meta-analysis methods, provided that an appropriate scale [235] or link function [375, 234] is used. Although different study weights can be used [232, 352], it is generally recommended to allow for between-study heterogeneity as validation studies are likely to differ in their design and execution [234, 21, 235]. As is the case in meta-analysis of prognostic factor research, meta-regression can be used to explore potential sources of between-study heterogeneity.

For instance, van Doorn et al. reviewed 19 published validations of CHA2DS2-VASc, a prediction model for estimating stroke risk in patients with atrial fibrillation [376]. A random effects meta-analysis was applied to summarize estimates of model discrimination (logit c-statistic) and annual risk per score (square root risks). The summary c-statistic was 0.64 (95% CI 0.56 – 0.71), which increased to 0.71 (95% CI 0.62 – 0.79) for studies recruiting patients from a hospital care setting. Further, stroke risks were found to vary substantially within the different scores and were notably elevated in hospital patients as compared to patients from the general population.

Development of a new prognostic model

It is also possible to summarize AD from multiple but similar prognostic model development studies and to combine their regression coefficients into a new prediction model (for example, via a multivariate meta-analysis) [357, 372]. This strategy is, however, often complicated by the poor reporting of key model parameters (and their standard errors and within-study correlations), by inconsistent covariate adjustment across studies, and by the presence of between-study heterogeneity. For this reason, meta-analysis of previously developed prognostic models only seems reasonable when the corresponding studies are fairly homogeneous and when the required AD are reported in sufficient detail (see also Figure 7.3).

7.3.2 Meta-analysis using IPD

When IPD are available, it becomes possible to assess and optimize the prognostic model's performance across different settings and populations using a one-stage or a two-stage meta-analysis approach.

Validation of an existing prognostic model

In the two-stage approach, the model is first validated separately in each IPD, yielding study-specific estimates of model discrimination and calibration. These estimates are then pooled across studies in the second stage, using univariate [21, 377, 232] or multivariate [231] meta-analysis methods (Figure 7.4). For instance, Snell *et al.* adopted multivariate IPD meta-analysis to summarize the calibration slope and concordance statistic of a prognostic model for breast cancer incidence. The summary estimates were then used in combination with estimates of between-study heterogeneity to calculate the probability that model performance would be adequate (i.e. within certain ranges) in new populations [231].

Model validation can also be performed through a one-stage approach. For instance, the summary calibration slope can be derived by fitting a mixed effect model with study-specific intercept terms and a random effect for the prognostic index.

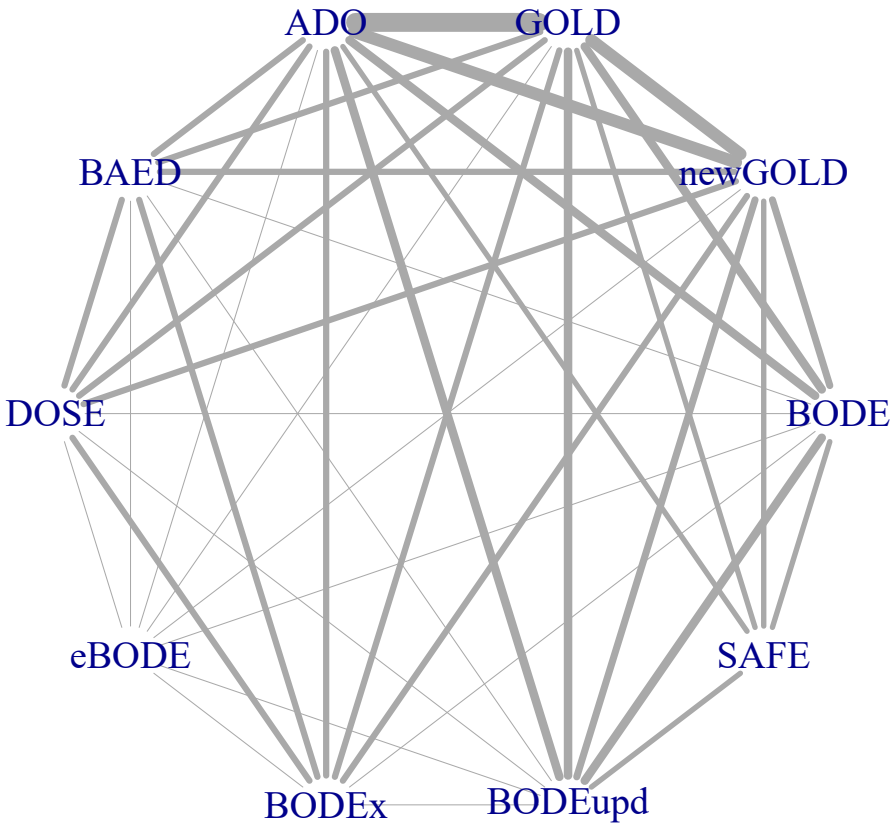
Finally, several extensions of one-stage and two-stage meta-analysis are possible. For instance, network meta-analysis (NMA) can be used to assess the (relative) performance of multiple prognostic models [378], which is particularly helpful when direct comparisons are not feasible for some studies. As an example, Haile *et al.* compared the performance of 10 prognostic models for calculating mortality risk in patients with chronic obstructive pulmonary disease [378]. Although IPD were available for 24 cohort studies ($N = 15\,762$), information on important variables was often missing such that some models could not be validated in one or more studies (Figure 7.5). A two-stage NMA was therefore adopted to summarize all available evidence on the models' comparative performance, and to allow the inclusion of studies where only few models could be validated.

Development of a new prognostic model

Meta-analysis of IPD is used increasingly often to develop new prognostic models, with improved generalizability across different settings and populations. Meta-analysis approaches are similar to prognostic factor research, and may involve a one-stage or a two-stage approach (see also Figure 7.3) [232]. In the two-stage approach, the parameters of the prognostic model (e.g. intercept term and regression coefficients) are estimated separately in each study and subsequently combined across studies using either a fixed or random effects meta-analysis. Conversely, in the one-stage approach, all IPD are simultaneously analysed by assuming a common, fixed, or random effect for each model parameter. Both approaches then yield a set of study-specific and/or "pooled" regression coefficients that can be used for making absolute risk predictions in a variety of populations. One-stage approaches are particularly helpful when studies are relatively small, or contain few events, as they use a more exact statistical approach and do not require continuity corrections when (partial) separation occurs [361]. Conversely, two-stage approaches are generally preferred when modeling interactions or non-linear terms, as they guard against over-parameterization and ecological bias [365].

As an example, Westeneng *et al.* recently performed a meta-analysis with IPD from 14 European cohorts to develop the ESCALC model for predicting survival

Figure 7.5: Validation of 10 prognostic models for 3-year mortality in patients with chronic obstructive pulmonary disease.



Depiction of network structure with lines weighted by the total number of participants available for each model comparison. [378] Abbreviations: GOLD, Global initiative for chronic Obstructive Lung Disease; BODE, Body mass index, airflow Obstruction, Dyspnoea and severe Exacerbations; BODE upd., BODE updated; ADO, Age, Dyspnoea, airflow Obstruction (we use the updated version of the ADO score in our analysis); e-BODE, severe acute exacerbation of COPD plus BODE; BODEx, Body mass index, airflow Obstruction, Dyspnoea, severe acute Exacerbation of COPD; DOSE, Dyspnoea, Obstruction, Smoking and Exacerbation frequency; SAFE, Saint George's Respiratory Questionnaire (SGRQ) score, Air-Flow limitation and Exercise capacity; B-AE-D, Body-mass index, Acute Exacerbations, Dyspnoea.

in patients with amyotrophic lateral sclerosis [379]. They fitted a Royston-Parmar survival model in the entire set of $N = 11\,475$ patients, and assumed a common baseline hazard and regression coefficients across cohorts. Because the resulting model showed some extent of mis-calibration upon validation, cohort-specific baseline hazard functions were reported to enable researchers to tailor model predictions to their population.

A particular advantage of IPD meta-analysis is that it enables the direct evaluation and optimization of a model's generalizability across different settings and populations through internal-external cross-validation (see chapter 3). [23, 26, 116, 221, 226] Briefly, this method iteratively omits one study from the meta-analysis to externally validate a model that is developed on the remaining studies. This process is repeated several times, leading to multiple estimates of model performance, which in turn can be summarized using aforementioned meta-analysis methods [235, 231]. If performance appears adequate across the available studies, the pooled data is used to develop a final model. Otherwise, it flags heterogeneous study populations where a developed model might not perform well and signals that additional predictors or more advanced modeling approaches (such as the inclusion of non-linear terms) or updating strategies (such as recalibration) might be needed.

Internal-external cross-validation has, for instance, been adopted during the development of ESCALC, a prognostic model for predicting survival in patients with amyotrophic lateral sclerosis. A one-stage approach was used to estimate a Royston-Parmar model using IPD from all but one study, after which its external validity was evaluated in the omitted study. The process was repeated for all studies, providing 14 estimates of discrimination and calibration performance. These estimates were then pooled using a random effects meta-analysis, yielding a summary c-statistic and calibration slope of, respectively, 0.78 (95% PI 0.74 to 0.82) and 1.01 (95% PI 0.83 to 1.18). These results suggest that the model is likely to perform well across different settings and populations.

7.3.3 Meta-analysis using IPD and AD

Validation of an existing prognostic model

Because IPD is commonly unavailable for one or more relevant validation studies, researchers may consider a two-stage meta-analysis to combine published estimates of prediction model performance with those derived from the IPD at hand. This approach has, however, not extensively been studied yet and further research is also warranted to explore alternative strategies such as hierarchical-related regression.

Development of a new prognostic model

For many disease areas, there is an abundance of competing models that predict similar outcomes in related populations. Hence, rather than developing a new prognostic model from scratch, it can be advantageous to combine the (AD of the) existing models with the available IPD [380, 381, 18, 382]. One approach is to summarize the models' regression coefficients together with the associations from the IPD [233, 368]. This is particularly useful if the data are reasonably homogeneous,

as synthesis then yields a prognostic model that is applicable to the "average" population. Conversely, when studies have different baseline risk or predictor-outcome associations, some tailoring will often be necessary to ensure that the new model remains sufficiently accurate in local settings. In these situations, the IPD can be used to adjust the existing models to specific populations by adopting Bayesian inference [233], model averaging [18], regression analysis [380, 18, 222, 383] or mixture models [222].

For example, model averaging was recently applied to combine the logistic EuroSCORE and EuroSCORE II models for predicting short-term mortality in patients undergoing coronary artery bypass graft surgery [382]. These models showed substantial mis-calibration in contemporary registry data and were therefore combined into a single model that was tailored to the contemporary population.

7.3.4 Summary points

Many prognostic model studies are based on relatively small samples, leading to overfitting, poor generalizability and over-optimism [384, 11]. Evidence synthesis allows to increase the effective sample size, and to study more diverse settings and populations [374, 26]. Although synthesis is ideally based on IPD, a systematic review and meta-analysis of published data can initially be performed to study the (discrimination and calibration) performance of a previously developed model. Estimates of between-study heterogeneity can then help to reveal the extent of necessary improvements (e.g., local tailoring), and to calculate the probability that the model(s) will be clinically useful in certain settings [227, 231]. In general, a good model will have satisfactory performance across different settings and populations. However, if prediction model performance is poor overall or prone to substantial between-study heterogeneity, retrieval of IPD may help to study causes of detrimental performance [21, 234, 24], and to establish whether distinct models are needed for different settings and populations [25].

When developing new or updating existing models, it is important to consider heterogeneity in baseline risk, predictor effects, the linear predictor and the absolute risk predictions [25]. Risk predictions should be reasonably similar across studies for a prediction model to be labeled 'generalizable', and therefore it is helpful to limit any heterogeneity in baseline risk and predictor effects whilst keeping the model's overall performance sufficiently high. In chapter 3 we described a statistical framework for developing and validating models on IPD from multiple studies. These methods can be applied to ascertain already during model development whether certain predictor effects are generalizable across populations and settings.

Finally, for newly developed prediction models from IPD-MA, it is helpful to provide any information that allows for tailored predictions. For instance, appropriate intercept terms can often be derived from the outcome incidence, particularly if predictor variables have been centered around their local means [116]. Similarly, predictor effects can sometimes be tailored using information about their particular measurement [385]. When it remains unclear which parameter values (e.g., intercept term) are most appropriate for predictions in new populations, researchers may use the pooled estimates or, preferably, integrate over the distribution of the random effects [386].

7.4 Concluding Remarks

In this paper, we have summarized and sign-posted various methods for meta-analysis of prognostic factor and prognostic model studies. Because these primary prognosis studies may address very different types of research questions and are often poorly reported, advanced meta-analysis methods are usually needed to provide (meaningful) summary estimates and understand sources of between-study heterogeneity. Regardless, researchers should not be daunted by their complexity, as we have shown that many of these methods have been implemented in traditional software packages and lead to an improved understanding of prognosis-related research questions.

For researchers embarking on a meta-analysis, the following issues should be taken into account. First, it is important to ensure that available data are of sufficient relevance and quality. It is recommended to conduct a systematic review of the literature and to harmonize available IPD sets. For instance, the methods we describe in chapter 4 can be applied to standardize from a non-target population or setting in model validation. Similarity of data sets can also be improved by standardizing related measurement scales [387], by adopting measurement error correction methods [287, 388, 389], or by treating bias arising from measurement error as a missing data problem [288, 387, 388]. For instance, the methods we described in chapter 5 can be applied to account for measurement error of binary predictor variables in an IPD-MA. Second, when data sets are affected by missing data, advanced imputation methods are needed to ensure valid inferences [253, 390, 391]. Finally, it is important to realize that not all meta-analysis methods have yet been rigorously assessed, and that further research is still needed to explore their potential areas of application.

Bibliography

[1] Adams ST, Leveson SH. Clinical prediction rules. BMJ. 2012 Jan;344:d8312. Available from: https://www.bmj.com/content/344/bmj.d8312.

[2] Plüddemann A, Wallace E, Bankhead C, Keogh C, Van der Windt D, Lasserson D, et al. Clinical prediction rules in practice: review of clinical guidelines and survey of GPs. Br J Gen Pract. 2014;64(621):e233–e242.

[3] Steyerberg EW. Clinical Prediction Models: a Practical Approach to Development, Validation, and Updating. Springer Science & Business Media; 2008.

[4] Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? Bmj. 2009;338:b375.

[5] Riley RD, van der Windt D, Croft P, Moons KGM. Prognosis Research in Healthcare: Concepts, Methods, and Impact. Oxford University Press; 2019.

[6] Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. BMJ. 2017 May;357. Available from: https://www.bmj.com/content/357/bmj.j2099.

[7] Nashef SA, Roques F, Sharples LD, Nilsson J, Smith C, Goldstone AR, et al. Euroscore ii. European journal of cardio-thoracic surgery. 2012;41(4):734–745.

[8] Wai CT, Greenson JK, Fontana RJ, Kalbfleisch JD, Marrero JA, Conjeevaram HS, et al. A simple noninvasive index can predict both significant fibrosis and cirrhosis in patients with chronic hepatitis C. Hepatology. 2003;38(2):518–526.

[9] Moons KGM, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012 May;98(9):691–698.

[10] Moons KG, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio) marker. Heart. 2012;98(9):683–690.

[11] Siontis GCM, Tzoulaki I, Castaldi PJ, Ioannidis JPA. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. Journal of Clinical Epidemiology. 2015 Jan;68(1):25–34.

[12] Charlson ME, Ales KL, Simon R, MacKenzie CR. Why predictive indexes perform less well in validation studies: is it magic or methods? Archives of Internal Medicine. 1987;147(12):2155–2161.

[13] Vergouwe Y, Moons KGM, Steyerberg EW. External validity of risk models: use of benchmark values to disentangle a case-mix effect from incorrect coefficients. American Journal of Epidemiology. 2010;172(8):971–980.

[14] Austin PC, Steyerberg EW. Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. BMC medical research methodology. 2012;12(1):82.

[15] Debray TPA, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KGM. A new framework to enhance the interpretation of external validation studies of clinical prediction models. Journal of Clinical Epidemiology. 2015 Mar;68(3):279–289.

[16] Luijken K, Groenwold RHH, Calster BV, Steyerberg EW, Smeden Mv. Impact of predictor measurement heterogeneity across settings on the performance of prediction models: A measurement error perspective. Statistics in Medicine. 2019;38(18):3444–3459. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.8183.

[17] Toll DB, Janssen KJM, Vergouwe Y, Moons KGM. Validation, updating and impact of clinical prediction rules: A review. Journal of Clinical Epidemiology. 2008 Nov;61(11):1085–1094. Available from: http://www.sciencedirect.com/science/article/pii/S0895435608001650.

[18] Debray TPA, Koffijberg H, Nieboer D, Vergouwe Y, Steyerberg EW, Moons KGM. Meta-analysis and aggregation of multiple published prediction models. Statistics in Medicine. 2014 Jun;33(14):2341–2362. Available from: http://doi.wiley.com/10.1002/sim.6080.

[19] Damen JAAG, Hooft L, Schuit E, Debray TPA, Collins GS, Tzoulaki I, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ. 2016 May;353:i2416.

[20] Riley RD, Ensor J, Snell KIE, Debray TPA, Altman DG, Moons KGM, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ. 2016;353:i3140.

[21] Debray TPA, Damen JAAG, Snell KIE, Ensor J, Hooft L, Reitsma JB, et al. A guide to systematic review and meta-analysis of prediction model performance. BMJ. 2017 Jan;356:i6460. Available from: https://www.bmj.com/content/356/bmj.i6460.

[22] Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. BMJ. 2009 May;338(may28 1):b605–b605. Available from: http://www.bmj.com/cgi/doi/10.1136/bmj.b605.

[23] Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-external, and external validation. Journal of Clinical Epidemiology. 2016 Jan;69:245–247.

[24] Austin PC, van Klaveren D, Vergouwe Y, Nieboer D, Lee DS, Steyerberg EW. Geographic and temporal validity of prediction models: different approaches were useful to examine model performance. Journal of Clinical Epidemiology. 2016 Nov;79:76–85. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0895435616301408.

[25] Steyerberg EW, Nieboer D, Debray TPA, van Houwelingen HC. Assessment of heterogeneity in an individual participant data meta-analysis of prediction models: An overview and illustration. Statistics in Medicine. 2019;38(22):4290–4309.

[26] Debray TPA, Riley RD, Rovers MM, Reitsma JB, Moons KGM. Individual participant data (IPD) meta-analyses of diagnostic and prognostic modeling studies: guidance on their use. PLoS Medicine. 2015;12(10):e1001886.

[27] Bross I. Misclassification in 2 x 2 tables. Biometrics. 1954;10(4):478–486.

[28] Keys A, Kihlberg JK. Effect of misclassification on estimated relative prevalence of a characteristic: Part I. Two populations infallibly distinguished. Part II. Errors in two variables. American Journal of Public Health and the Nations Health. 1963;53(10):1656–1665.

[29] Weinberg CA, Umbach DM, Greenland S. When will nondifferential misclassification of an exposure preserve the direction of a trend? American Journal of Epidemiology. 1994;140(6):565–571.

[30] Copeland KT, Checkoway H, McMichael AJ, Holbrook RH. Bias due to misclassification in the estimation of relative risk. American journal of epidemiology. 1977;105(5):488–495.

[31] Dosemeci M, Wacholder S, Lubin JH. Does nondifferential misclassification of exposure always bias a true effect toward the null value? American journal of epidemiology. 1990;132(4):746–748.

[32] Birkett NJ. Effect of Nondifferential Misclassification on Estimates of Odds Ratios with Multiple Levels of Exposure. American Journal of Epidemiology. 1992 Aug;136(3):356–362. Available from: https://academic.oup.com/aje/article/136/3/356/96123.

[33] Gustafson P. Measurement error and misclassification in statistics and epidemiology: impacts and Bayesian adjustments. CRC Press; 2003.

[34] Carroll R, Ruppert D, Stefanski LA, Crainiceanu CM. Measurement error in nonlinear models. A modern Perspective: Chapmn & Hall/CRC. 2006;.

[35] Gustafson P, Greenland S. Misclassification. In: Ahrens W, Pigeot I, editors. Handbook of Epidemiology. New York, NY: Springer New York; 2014. p. 639–658. Available from: https://doi.org/10.1007/978-0-387-09834-0_58.

[36] Riley RD, Steyerberg EW. Meta-analysis of a binary outcome using individual participant data and aggregate data. Research Synthesis Methods. 2010 Jan;1(1):2–19. Available from: http://doi.wiley.com/10.1002/jrsm.4.

[37] Stewart LA, Parmar MKB. Meta-analysis of the literature or of individual patient data: is there a difference? The Lancet. 1993 Feb;341(8842):418–422. Available from: http://www.sciencedirect.com/science/article/pii/014067369393004K.

[38] Tierney JF, Vale C, Riley R, Tudur Smith C, Stewart L, Clarke M, et al. Individual participant data (IPD) meta-analyses of randomised controlled trials: guidance on their use. PLoS Medicine. 2015;12(7):e1001855.

[39] Lyman GH, Kuderer NM. The strengths and limitations of meta-analyses based on aggregate data. BMC Medical Research Methodology. 2005;5:14. Available from: http://dx.doi.org/10.1186/1471-2288-5-14.

[40] Tudur Smith C, Williamson PR, Marson AG. An overview of methods and empirical comparison of aggregate data and individual patient data results for investigating heterogeneity in meta-analysis of time-to-event outcomes. Journal of Evaluation in Clinical Practice. 2005 Oct;11(5):468–478. Available from: http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2753.2005.00559.x/abstract.

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

[42] Debray TPA, Moons KGM, van Valkenhoef G, Efthimiou O, Hummel N, Groenwold RHH, et al. Get real in individual participant data (IPD) meta-analysis: a review of the methodology. Research Synthesis Methods. 2015 Dec;6(4):293–309.

[43] Thomas D, Platt R, Benedetti A. A comparison of analytic approaches for individual patient data meta-analyses with binary outcomes. BMC Medical Research Methodology. 2017;17:28. Available from: http://dx.doi.org/10.1186/s12874-017-0307-7.

[44] Jackson D, Law M, Stijnen T, Viechtbauer W, White IR. A comparison of seven random-effects models for meta-analyses that estimate the summary odds ratio. Statistics in Medicine. 2018;37(7):1059–1085.

[45] Legha A, Riley RD, Ensor J, Snell KI, Morris TP, Burke DL. Individual participant data meta-analysis of continuous outcomes: A comparison of approaches for specifying and estimating one-stage models. Statistics in Medicine. 2018;.

[46] Riley RD, Kauser I, Bland M, Thijs L, Staessen JA, Wang J, et al. Meta-analysis of randomised trials with a continuous outcome according to baseline imbalance and availability of individual participant data. Statistics in Medicine. 2013;32(16):2747–2766.

[47] Tudur Smith C, Williamson PR, Marson AG. Investigating heterogeneity in an individual patient data meta-analysis of time to event outcomes. Statistics in Medicine. 2005 May;24(9):1307–1319. Available from: http://onlinelibrary.wiley.com/doi/10.1002/sim.2050/abstract.

[48] Bowden J, Tierney JF, Simmonds M, Copas AJ, Higgins JP. Individual patient data meta-analysis of time-to-event outcomes: one-stage versus two-stage approaches for estimating the hazard ratio under a random effects model. Research Synthesis Methods. 2011 Sep;2(3):150–162.

[49] Wienke A. Frailty models in survival analysis. Boca Raton, FL: CRC Press; 2011.

[50] Klein JP, van Houwelingen HC, Ibrahim JG, Scheike TH. Handbook of Survival Analysis. Chapman and Hall/CRC; 2013.

[51] Duchateau L, Janssen P. The Frailty Model. New York, NY: Springer-Verlag; 2008.

[52] Hougaard P. Analysis of Multivariate Survival Data. New York, NY: Springer-Verlag; 2012.

[53] Cox DR. Regression Models and Life-Tables. Journal of the Royal Statistical Society Series B (Methodological). 1972;34(2):187–220. Available from: http://www.jstor.org/stable/2985181.

[54] Vaupel JW, Manton KG, Stallard E. The Impact of Heterogeneity in Individual Frailty on the Dynamics of Mortality. Demography. 1979;16(3):439–454. Available from: http://www.jstor.org/stable/2061224.

[55] Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. Biometrika. 1984;71(3):431–444.

[56] Hougaard P. Fundamentals of Survival Data. Biometrics. 1999;55(1):13–22. Available from: http://www.jstor.org/stable/2533890.

[57] Lin NX, Logan S, Henley WE. Bias and Sensitivity Analysis When Estimating Treatment Effects from the Cox Model with Omitted Covariates. Biometrics. 2013 Dec;69(4):850–860. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4230475/.

[58] Harrell FE. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. 2nd ed. Cham, Switzerland: Springer; 2015. Available from: //www.springer.com/gp/book/9783319194240.

[59] Keiding N, Andersen PK, Klein JP. The Role of Frailty Models and Accelerated Failure Time Models in Describing Heterogeneity Due to Omitted Covariates. Statistics in Medicine. 1997 Jan;16(2):215–224. Available from: http://onlinelibrary.wiley.com/doi/10.1002/(SICI)1097-0258(19970130)16:2<215::AID-SIM481>3.0.CO;2-J/abstract.

[60] Gompertz B. On the Nature of the Function Expressive of the Law of Human Mortality, and on a New Mode of Determining the Value of Life Contingencies. Philosophical Transactions of the Royal Society of London. 1825;115:513–583. Available from: http://www.jstor.org/stable/107756.

[61] Royston P, Lambert PC. Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model. College Station, Texas, USA: Stata Press; 2011. Available from: stata.com/bookstore/flexible-parametric-survival-analysis-stata/.

[62] Bennett S. Log-Logistic Regression Models for Survival Data. Journal of the Royal Statistical Society Series C (Applied Statistics). 1983;32(2):165–171. Available from: http://www.jstor.org/stable/2347295.

[63] Bennett S. Analysis of survival data by the proportional odds model. Statistics in Medicine. 1983;2(2):273–277. Available from: http://onlinelibrary.wiley.com/doi/10.1002/sim.4780020223/full.

[64] Makeham WM. On the Law of Mortality and the Construction of Annuity Tables. The Assurance Magazine, and Journal of the Institute of Actuaries. 1860;8(6):301–310. Available from: http://www.jstor.org/stable/41134925.

[65] Hougaard P. Modelling Heterogeneity in Survival Data. Journal of Applied Probability. 1991;28(3):695–701. Available from: http://www.jstor.org/stable/3214503.

[66] Abbring JH, van den Berg GJ. The unobserved heterogeneity distribution in duration analysis. Biometrika. 2007 Mar;94(1):87–99. Available from: https://academic.oup.com/biomet/article/94/1/87/228674/The-unobserved-heterogeneity-distribution-in.

[67] Hernández AV, Eijkemans MJC, Steyerberg EW. Randomized Controlled Trials With Time-to-Event Outcomes: How Much Does Prespecified Covariate Adjustment Increase Power? Annals of Epidemiology. 2006 Jan;16(1):41–48. Available from: http://www.sciencedirect.com/science/article/pii/S1047279705003248.

[68] Hjort NL. On Inference in Parametric Survival Data Models. International Statistical Review. 1992;60(3):355–387. Available from: http://www.jstor.org/stable/1403683.

[69] 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 Aug;21(15):2175–2197.

[70] Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. Biometrika. 1994;81(3):515–526.

[71] Thompson S, Kaptoge S, White I, Wood A, Perry P, Danesh J, et al. Statistical methods for the time-to-event analysis of individual participant data from multiple epidemiological studies. International Journal of Epidemiology. 2010 Oct;39(5):1345–1359.

[72] Wei LJ. The accelerated failure time model: A useful alternative to the cox regression model in survival analysis. Statistics in Medicine. 1992 Jan;11(14-15):1871–1879. Available from: http://onlinelibrary.wiley.com/doi/10.1002/sim.4780111409/abstract.

[73] Kay R, Kinnersley N. On the Use of the Accelerated Failure Time Model as an Alternative to the Proportional Hazards Model in the Treatment of Time to Event Data: A Case Study in Influenza. Drug Information Journal. 2002 Jul;36(3):571–579. Available from: http://journals.sagepub.com/doi/abs/10.1177/009286150203600312.

[74] Lambert P, Collett D, Kimber A, Johnson R. Parametric accelerated failure time models with random effects and an application to kidney transplant survival. Statistics in Medicine. 2004 Oct;23(20):3177–3192.

[75] Freeman SC, Carpenter JR. Bayesian one-step IPD network meta-analysis of time-to-event data using Royston-Parmar models. Research Synthesis Methods. 2017;8(4):451–464.

[76] Lin DY. Cox regression analysis of multivariate failure time data: the marginal approach. Statistics in Medicine. 1994;13(21):2233–2247. Available from: http://onlinelibrary.wiley.com/doi/10.1002/sim.4780132105/full.

[77] Johnson B, Carlin BP, Hodges JS. Cross-Study Hierarchical Modeling of Stratified Clinical Trial Data. Journal of Biopharmaceutical Statistics. 1999 Jan;9(4):617–640. Available from: http://dx.doi.org/10.1081/BIP-100101199.

[78] Glidden DV, Vittinghoff E. Modelling clustered survival data from multicentre clinical trials. Statistics in Medicine. 2004 Feb;23(3):369–388.

[79] Gardiner JC, Luo Z, Roman LA. Fixed effects, random effects and GEE: what are the differences? Statistics in Medicine. 2009;28(2):221–239. Available from: http://onlinelibrary.wiley.com/doi/10.1002/sim.3478/full.

[80] Katsahian S, Latouche A, Mary JY, Chevret S, Porcher R. Practical methodology of meta-analysis of individual patient data using a survival outcome. Contemporary Clinical Trials. 2008 Mar;29(2):220–230. Available from: http://www.sciencedirect.com/science/article/pii/S1551714407001267.

[81] Aalen OO, Cook RJ, Røysland K. Does Cox analysis of a randomized survival study yield a causal treatment effect? Lifetime Data Analysis. 2015;21(4):579–593. Available from: DOI:10.1007/s10985-015-9335-y.

[82] Burke DL, Ensor J, Riley RD. Meta-analysis using individual participant data: one-stage and two-stage approaches, and why they may differ. Statistics in Medicine. 2017;36(5):855–875.

[83] Hua H, Burke DL, Crowther MJ, Ensor J, Tudur Smith C, Riley RD. One-stage individual participant data meta-analysis models: estimation of treatment-covariate interactions must avoid ecological bias by separating out within-trial and across-trial information. Statistics in Medicine. 2017;36(5):772–789.

[84] Berlin JA, Santanna J, Schmid CH, Szczech LA, Feldman HI. Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head. Statistics in Medicine. 2002 Feb;21(3):371–387. Available from: http://onlinelibrary.wiley.com/doi/10.1002/sim.1023/abstract.

[85] Riley RD, Higgins JPT, Deeks JJ. Interpretation of random effects meta-analyses. BMJ. 2011 Feb;342:d549. Available from: https://www.bmj.com/content/342/bmj.d549.

[86] Rondeau V, Michiels S, Liquet B, Pignon JP. Investigating trial and treatment heterogeneity in an individual patient data meta-analysis of survival data by means of the penalized maximum likelihood approach. Statistics in Medicine. 2008 May;27(11):1894–1910. Available from: http://onlinelibrary.wiley.com/doi/10.1002/sim.3161/abstract.

[87] Fine JP, Gray RJ. A Proportional Hazards Model for the Subdistribution of a Competing Risk. Journal of the American Statistical Association. 1999 Jun;94(446):496–509. Available from: http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1999.10474144.

[88] Beyersmann J, Allignol A, Schumacher M. Competing Risks and Multistate Models with R. Springer Science & Business Media; 2011.

[89] Wolkewitz M, Cooper BS, Palomar-Martinez M, Alvarez-Lerma F, Olaechea-Astigarraga P, Barnett AG, et al. Multilevel competing risk models to evaluate the risk of nosocomial infection. Critical Care. 2014 Apr;18(2):R64.

[90] Satagopan JM, Ben-Porat L, Berwick M, Robson M, Kutler D, Auerbach AD. A note on competing risks in survival data analysis. British Journal of Cancer. 2004 Oct;91(7):1229–1235. Available from: https://www.nature.com/articles/6602102.

[91] Lee KH, Dominici F, Schrag D, Haneuse S. Hierarchical models for semicompeting risks data with application to quality of end-of-life care for pancreatic cancer. Journal of the American Statistical Association. 2016;111(515):1075–1095.

[92] Giorgi R, Belot A, Gaudart J, Launoy G, French Network of Cancer Registries FRANCIM. The performance of multiple imputation for missing covariate data within the context of regression relative survival analysis. Statistics in Medicine. 2008 Dec;27(30):6310–6331. Available from: http://onlinelibrary.wiley.com/doi/10.1002/sim.3476/epdf.

[93] White IR, Royston P. Imputing missing covariate values for the Cox model. Statistics in Medicine. 2009 Jul;28(15):1982–1998.

[94] Resche-Rigon M, White IR, Bartlett JW, Peters SAE, Thompson SG, PROG-IMT Study Group. Multiple imputation for handling systematically missing confounders in meta-analysis of individual participant data. Statistics in Medicine. 2013 Dec;32(28):4890–4905.

[95] Falcaro M, Nur U, Rachet B, Carpenter JR. Estimating excess hazard ratios and net survival when covariate data are missing: strategies for multiple imputation. Epidemiology. 2015 May;26(3):421–428.

[96] Resche-Rigon M, White IR. Multiple imputation by chained equations for systematically and sporadically missing multilevel data. Statistical Methods in Medical Research. 2018 Jun;27(6):1634–1649.

[97] Abo-Zaid G, Guo B, Deeks JJ, Debray TPA, Steyerberg EW, Moons KGM, et al. Individual participant data meta-analyses should not ignore clustering. Journal of Clinical Epidemiology. 2013 Aug;66(8):865–873.e4. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0895435613000723.

[98] Riley RD, Lambert PC, Staessen JA, Wang J, Gueyffier F, Thijs L, et al. Meta-analysis of continuous outcomes combining individual patient data and aggregate data. Statistics in Medicine. 2008;27(11):1870–1893. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.3165.

[99] Firth D. Bias reduction of maximum likelihood estimates. Biometrika. 1993;80(1):27–38.

[100] Heinze G, Schemper M. A solution to the problem of monotone likelihood in Cox regression. Biometrics. 2001;57(1):114–119.

[101] Lambert PC, Sutton AJ, Abrams KR, Jones DR. A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. Journal of Clinical Epidemiology. 2002 Jan;55(1):86–94. Available from: http://www.sciencedirect.com/science/article/pii/S0895435601004140.

[102] Localio AR, Berlin JA, Ten Have TR, Kimmel SE. Adjustments for center in multicenter studies: an overview. Annals of Internal Medicine. 2001 Jul;135(2):112–123. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.128.1827&rep=rep1&type=pdf.

[103] Aalen OO. Effects of frailty in survival analysis. Statistical Methods in Medical Research. 1994 Oct;3(3):227–243. Available from: http://dx.doi.org/10.1177/096228029400300303.

[104] Whitehead A, Whitehead J. A general parametric approach to the meta-analysis of randomized clinical trials. Statistics in Medicine. 1991 Nov;10(11):1665–1677.

[105] Hartung J. An alternative method for meta-analysis. Biometrical Journal. 1999;41(8):901–916.

[106] Sidik K, Jonkman JN. Robust variance estimation for random effects meta-analysis. Computational Statistics & Data Analysis. 2006;50(12):3681–3701.

[107] Knapp G, Hartung J. Improved tests for a random effects meta-regression with a single covariate. Statistics in Medicine. 2003;22(17):2693–2710.

[108] Hartung J, Knapp G. On confidence intervals for the among-group variance in the one-way random effects model with unequal error variances. Journal of Statistical Planning and Inference. 2005;127(1-2):157–177.

[109] Jackson D, Law M, Rücker G, Schwarzer G. The Hartung-Knapp modification for random-effects meta-analysis: A useful refinement but are there any residual concerns? Statistics in Medicine. 2017;36(25):3923–3934.

[110] Higgins JPT, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2009;172(1):137–159.

[111] Su X, Zhou T, Yan X, Fan J, Yang S. Interaction trees with censored survival data. The International Journal of Biostatistics. 2008 Jan;4(1):Article 2.

[112] van Houwelingen HC, Eilers PHC. Non-proportional hazards models in survival analysis. In: Bethlehem JG, van der Heijden PGM, editors. COMPSTAT. Heidelberg: Physica;; 2000. p. 151–160. Available from: https://link.springer.com/chapter/10.1007/978-3-642-57678-2_14.

[113] Crowther MJ, Look MP, Riley RD. Multilevel mixed effects parametric survival models using adaptive Gauss-Hermite quadrature with application to recurrent events and individual participant data meta-analysis. Statistics in Medicine. 2014 Sep;33(22):3844–3858.

[114] Siannis F, Barrett JK, Farewell VT, Tierney JF. One-stage parametric meta-analysis of time-to-event outcomes. Statistics in Medicine. 2010 Dec;29(29):3030–3045.

[115] Andreano A, Rebora P, Valsecchi MG. Measures of single arm outcome in meta-analyses of rare events in the presence of competing risks. Biometrical Journal Biometrische Zeitschrift. 2015 Jul;57(4):649–660.

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

[117] DerSimonian R, Laird N. Meta-analysis in clinical trials. Controlled Clinical Trials. 1986;7(3):177–188. Available from: http://www.sciencedirect.com/science/article/pii/0197245686900462.

[118] Langan D, Higgins JPT, Simmonds M. Comparative performance of heterogeneity variance estimators in meta-analysis: a review of simulation studies: A Review of Simulation Studies. Research Synthesis Methods. 2017 Jun;8(2):181–198. Available from: http://doi.wiley.com/10.1002/jrsm.1198.

[119] Veroniki AA, Jackson D, Viechtbauer W, Bender R, Bowden J, Knapp G, et al. Methods to estimate the between-study variance and its uncertainty in meta-analysis. Research Synthesis Methods. 2016;7(1):55–79.

[120] Brockwell SE, Gordon IR. A comparison of statistical methods for meta-analysis. Statistics in Medicine. 2001;20(6):825–840.

[121] Sidik K, Jonkman JN. A comparison of heterogeneity variance estimators in combining results of studies. Statistics in Medicine. 2007;26(9):1964–1981.

[122] Langan D, Higgins JPT, Jackson D, Bowden J, Veroniki AA, Kontopantelis E, et al. A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. Research Synthesis Methods. 2018 Aug;.

[123] Ioannidis JPA, Patsopoulos NA, Evangelou E. Uncertainty in heterogeneity estimates in meta-analyses. BMJ. 2007;335(7626):914–916.

[124] Viechtbauer W. Confidence intervals for the amount of heterogeneity in meta-analysis. Statistics in Medicine. 2007;26(1):37–52.

[125] Sidik K, Jonkman JN. A simple confidence interval for meta-analysis. Statistics in Medicine. 2002;21(21):3153–3159. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.1262.

[126] Hartung J, Knapp G. A refined method for the meta-analysis of controlled clinical trials with binary outcome. Statistics in Medicine. 2001;20(24):3875–3889.

[127] IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Medical Research Methodology. 2014;14(1):25.

[128] Normand SLT. Meta-analysis: formulating, evaluating, combining, and reporting. Statistics in Medicine. 1999;18(3):321–359.

[129] Veroniki AA, Jackson D, Bender R, Kuss O, Langan D, Higgins JPT, et al. Methods to calculate uncertainty in the estimated overall effect size from a random-effects meta-analysis. Research Synthesis Methods. 2019 Mar;10(1):23–43. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/jrsm.1319.

[130] Fisher DJ. Two-stage individual participant data meta-analysis and generalized forest plots. Stata Journal. 2015;15(2):369–396. Available from: https://journals.sagepub.com/doi/pdf/10.1177/1536867X1501500203.

[131] Tudur Smith C, Williamson PR. A comparison of methods for fixed effects meta-analysis of individual patient data with time to event outcomes. Clinical Trials. 2007;4(6):621–630. Available from: http://ctj.sagepub.com/content/4/6/621.short.

[132] Michiels S, Baujat B, Mahé C, Sargent DJ, Pignon JP. Random effects survival models gave a better understanding of heterogeneity in individual patient data meta-analyses. Journal of Clinical Epidemiology. 2005 Mar;58(3):238–245. Available from: http://www.sciencedirect.com/science/article/pii/S089543560400294X.

[133] Therneau T M, Grambsch PM. Modeling Survival Data: Extending the Cox Model. Springer, New York; 2000.

[134] Sargent DJ. A general framework for random effects survival analysis in the Cox proportional hazards setting. Biometrics. 1998 Dec;54(4):1486–1497. Available from: http://www.jstor.org/stable/pdf/2533673.pdf.

[135] Vaida F, Xu R. Proportional hazards model with random effects. Statistics in Medicine. 2000 Dec;19(24):3309–3324. Available from: http://doi.wiley.com/10.1002/1097-0258%2820001230%2919%3A24%3C3309%3A%3AAID-SIM825%3E3.0.CO%3B2-9.

[136] Simmonds MC, Higgins JPT, Stewart LA. Random-effects meta-analysis of time-to-event data using the expectation-maximisation algorithm and shrinkage estimators. Research Synthesis Methods. 2013 Jun;4(2):144–155.

[137] Morris C, Christiansen C. Fitting Weibull duration models with random effects. Lifetime Data Analysis. 1995;1(4):347–359. Available from: http://link.springer.com/10.1007/BF00985449.

[138] Rondeau V, Filleul L, Joly P. Nested frailty models using maximum penalized likelihood estimation. Statistics in Medicine. 2006 Dec;25(23):4036–4052.

[139] Donohue MC, Overholser R, Xu R, Vaida F. Conditional Akaike information under generalized linear and proportional hazards mixed models. Biometrika. 2011 Sep;98(3):685–700. Available from: https://academic.oup.com/biomet/article/98/3/685/235983/Conditional-Akaike-information-under-generalized.

[140] Munda M, Legrand C. Adjusting for centre heterogeneity in multicentre clinical trials with a time-to-event outcome. Pharmaceutical Statistics. 2014 Apr;13(2):145–152.

[141] Lanke J. How to Describe the Impact of the Family-Specific Frailty (Appendix of "Quantifying the Family Frailty Effect in Infant and Child Mortality by Using Median Hazard Ratio (MHR)"). Historical Methods: A Journal of Quantitative and Interdisciplinary History. 2010;Available from: dx.doi.org/10.1080/01615440903270299.

[142] Bengtsson T, Dribe M. Quantifying the Family Frailty Effect in Infant and Child Mortality by Using Median Hazard Ratio (MHR). Historical Methods: A Journal of Quantitative and Interdisciplinary History. 2010 Jan;43(1):15–27. Available from: http://dx.doi.org/10.1080/01615440903270299.

[143] Austin PC, Wagner P, Merlo J. The median hazard ratio: a useful measure of variance and general contextual effects in multilevel survival analysis. Statistics in Medicine. 2017;36(6):928–938.

[144] Cai J, Prentice RL. Estimating Equations for Hazard Ratio Parameters Based on Correlated Failure Time Data. Biometrika. 1995;82(1):151–164. Available from: http://www.jstor.org/stable/2337635.

[145] Spiekerman CF, Lin DY. Marginal regression models for multivariate failure time data. Journal of the American Statistical Association. 1998;93(443):1164–1175. Available from: http://www.tandfonline.com/doi/abs/10.1080/01621459.1998.10473777.

[146] McGilchrist CA. REML estimation for survival models with frailty. Biometrics. 1993;p. 221–225.

[147] Ripatti S, Palmgren J. Estimation of multivariate frailty models using penalized partial likelihood. Biometrics. 2000;56(4):1016–1022.

[148] Therneau TM, Grambsch PM, Pankratz VS. Penalized survival models and frailty. Journal of Computational and Graphical Statistics. 2003;12(1):156–175.

[149] 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 Mar;12:34.

[150] Clayton DG. A Monte Carlo method for Bayesian inference in frailty models. Biometrics. 1991;p. 467–485. Available from: http://www.jstor.org/stable/2532139.

[151] Legrand C, Ducrocq V, Janssen P, Sylvester R, Duchateau L. A Bayesian approach to jointly estimate centre and treatment by centre heterogeneity in a proportional hazards model. Statistics in Medicine. 2005 Dec;24(24):3789–3804.

[152] Hobbs BP, Sargent DJ, Carlin BP. Commensurate Priors for Incorporating Historical Information in Clinical Trials Using General and Generalized Linear Models. Bayesian Analysis. 2012 Aug;7(3):639–674.

[153] Bennett MM, Crowe BJ, Price KL, Stamey JD, Seaman JW. Comparison of Bayesian and frequentist meta-analytical approaches for analyzing time to event data. Journal of Biopharmaceutical Statistics. 2013;23(1):129–145.

[154] Sobel M, Madigan D, Wang W. Causal Inference for Meta-Analysis and Multi-Level Data Structures, with Application to Randomized Studies of Vioxx. Psychometrika. 2016 Jul;.

[155] Brilleman SL, Crowther MJ, Moreno-Betancur M, Buros Novik J, Dunyak J, Al-Huniti N, et al. Joint longitudinal and time-to-event models for multilevel hierarchical data. Statistical Methods in Medical Research. 2019;28(12):3502–3515.

[156] Rotolo F, Paoletti X, Michiels S. surrosurv: An R package for the evaluation of failure time surrogate endpoints in individual patient data meta-analyses of randomized clinical trials. Computer Methods and Programs in Biomedicine. 2018 Mar;155:189–198.

[157] Boutitie F, Gueyffier F, Pocock SJ, Boissel JP. Assessing treatment–time interaction in clinical trials with time to event data: a meta-analysis of hypertension trials. Statistics in Medicine. 1998 Dec;17(24):2883–2903. Available from: http://onlinelibrary.wiley.com/doi/10.1002/(SICI)1097-0258(19981230)17:24<2883::AID-SIM900>3.0.CO;2-L/abstract.

[158] Cox C, Chu H, Schneider MF, Muñoz A. Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. Statistics in Medicine. 2007 Oct;26(23):4352–4374. Available from: http://onlinelibrary.wiley.com/doi/10.1002/sim.2836/abstract.

[159] Jung TH, Peduzzi P, Allore H, Kyriakides TC, Esserman D. A joint model for recurrent events and a semi-competing risk in the presence of multi-level clustering. Statistical Methods in Medical Research. 2018 Jul;0(0):1–15.

[160] Saramago P, Chuang LH, Soares MO. Network meta-analysis of (individual patient) time to event data alongside (aggregate) count data. BMC Medical Research Methodology. 2014 Sep;14:105.

[161] Goldstein H, Browne W, Rasbash J. Multilevel modelling of medical data. Statistics in Medicine. 2002;21(21):3291–3315.

[162] Kent DM, Steyerberg E, van Klaveren D. Personalized evidence based medicine: predictive approaches to heterogeneous treatment effects. BMJ. 2018;363:k4245.

[163] Hess KR. Assessing time-by-covariate interactions in proportional hazards regression models using cubic spline functions. Statistics in Medicine. 1994;13(10):1045–1062.

[164] Giorgi R, Abrahamowicz M, Quantin C, Bolard P, Esteve J, Gouvernet J, et al. A relative survival regression model using B-spline functions to model non-proportional hazards. Statistics in Medicine. 2003;22(17):2767–2784.

[165] Thomas L, Reyes EM. Tutorial: survival estimation for Cox regression models with time-varying coefficients using SAS and R. Journal of Statistical Software. 2014;61(c1):1–23.

[166] White IR, Kaptoge S, Royston P, Sauerbrei W, Collaboration ERF. Meta-analysis of non-linear exposure-outcome relationships using individual participant data: A comparison of two methods. Statistics in Medicine. 2019;38(3):326–338.

[167] Simmonds MC, Tierney J, Bowden J, Higgins JP. Meta-analysis of time-to-event data: a comparison of two-stage methods. Research Synthesis Methods. 2011 Sep;2(3):139–149.

[168] Royston P, Parmar MKB. The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. Statistics in Medicine. 2011 Aug;30(19):2409–2421. Available from: http://onlinelibrary.wiley.com/doi/10.1002/sim.4274/abstract.

[169] Wei Y, Royston P, Tierney JF, Parmar MKB. Meta-analysis of time-to-event outcomes from randomized trials using restricted mean survival time: application to individual participant data. Statistics in Medicine. 2015 Sep;34(21):2881–2898.

[170] Lueza B, Rotolo F, Bonastre J, Pignon JP, Michiels S. Bias and precision of methods for estimating the difference in restricted mean survival time from an individual patient data meta-analysis. BMC Medical Research Methodology. 2016;16:37. Available from: http://dx.doi.org/10.1186/s12874-016-0137-z.

[171] Barrett JK, Farewell VT, Siannis F, Tierney J, Higgins JPT. Two-stage meta-analysis of survival data from individual participants using percentile ratios. Statistics in Medicine. 2012 Dec;31(30):4296–4308.

[172] Noordzij M, Leffondré K, van Stralen KJ, Zoccali C, Dekker FW, Jager KJ. When do we need competing risks methods for survival analysis in nephrology? Nephrology Dialysis Transplantation. 2013 Nov;28(11):2670–2677. Available from: https://academic.oup.com/ndt/article/28/11/2670/1823847.

[173] Andersen PK, Geskus RB, de Witte T, Putter H. Competing risks in epidemiology: possibilities and pitfalls. International Journal of Epidemiology. 2012 Jun;41(3):861–870. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3396320/.

[174] Geskus RB. Data Analysis with Competing Risks and Intermediate States. Chapman and Hall/CRC; 2015.

[175] Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. Statistics in Medicine. 2007;26(11):2389–2430.

[176] Riley RD, Thompson JR, Abrams KR. An alternative model for bivariate random-effects meta-analysis when the within-study correlations are unknown. Biostatistics. 2008 Jan;9(1):172–186. Available from: http://biostatistics.oxfordjournals.org/content/9/1/172.

[177] Riley RD, Price MJ, Jackson D, Wardle M, Gueyffier F, Wang J, et al. Multivariate meta-analysis using individual participant data. Research Synthesis Methods. 2015 Jun;6(2):157–174.

[178] Efthimiou O, Debray TP, Valkenhoef G, Trelle S, Panayidou K, Moons KGM, et al. GetReal in network meta-analysis: a review of the methodology. Research Synthesis Methods. 2016;7(3):236–263.

[179] Riley RD, Jackson D, Salanti G, Burke DL, Price M, Kirkham J, et al. Multivariate and network meta-analysis of multiple outcomes and multiple treatments: rationale, concepts, and examples. BMJ. 2017 Sep;358:j3932. Available from: https://www.bmj.com/content/358/bmj.j3932.

[180] Veroniki AA, Straus SE, Soobiah C, Elliott MJ, Tricco AC. A scoping review of indirect comparison methods and applications using individual patient data. BMC Medical Research Methodology. 2016 Apr;16:47.

[181] Freeman SC, Fisher D, Tierney JF, Carpenter JR. A framework for identifying treatment-covariate interactions in individual participant data network meta-analysis. Research Synthesis Methods. 2018;9(3):393–407.

[182] ICH. General Considerations for Clinical Trials; 1997. Available from: http://www.ich.org/products/guidelines/efficacy/efficacy-single/article/general-considerations-for-clinical-trials.htmlAccessed25February2019.

[183] Shi Q, Sargent DJ. Meta-analysis for the evaluation of surrogate endpoints in cancer clinical trials. International Journal of Clinical Oncology. 2009 Apr;14(2):102–111. Available from: http://link.springer.com/article/10.1007/s10147-009-0885-4.

[184] Renfro LA, Shi Q, Xue Y, Li J, Shang H, Sargent DJ. Center-Within-Trial Versus Trial-Level Evaluation of Surrogate Endpoints. Computational Statistics & Data Analysis. 2014 Oct;78:1–20.

[185] Buyse M, Piedbois P. On the Relationship Between Response to Treatment and Survival Time. Statistics in Medicine. 1996 Dec;15(24):2797–2812. Available from: http://onlinelibrary.wiley.com/doi/10.1002/(SICI)1097-0258(19961230)15:24<2797::AID-SIM290>3.0.CO;2-V/abstract.

[186] Schluchter MD, Konstan MW, Davis PB. Jointly modelling the relationship between survival and pulmonary function in cystic fibrosis patients. Statistics in Medicine. 2002 May;21(9):1271–1287. Available from: http://onlinelibrary.wiley.com/doi/10.1002/sim.1104/abstract.

[187] Luo S, Su X, DeSantis SM, Huang X, Yi M, Hunt KK. Joint model for a diagnostic test without a gold standard in the presence of a dependent terminal event. Statistics in Medicine. 2014 Jul;33(15):2554–2566.

[188] Shi Q, Renfro LA, Bot BM, Burzykowski T, Buyse M, Sargent DJ. Comparative assessment of trial-level surrogacy measures for candidate time-to-event surrogate endpoints in clinical trials. Computational Statistics & Data Analysis. 2011 Sep;55(9):2748–2757. Available from: http://www.sciencedirect.com/science/article/pii/S0167947311001058.

[189] Renfro LA, Shi Q, Sargent DJ, Carlin BP. Bayesian adjusted R2 for the meta-analytic evaluation of surrogate time-to-event endpoints in clinical trials. Statistics in Medicine. 2012 Apr;31(8):743–761.

[190] Buyse M, Molenberghs G, Paoletti X, Oba K, Alonso A, van der Elst W, et al. Statistical evaluation of surrogate endpoints with examples from cancer clinical trials. Biometrical Journal. 2016 Jan;58(1):104–132.

[191] Poppe KK, Doughty RN, Yu CM, Quintana M, Møller JE, Klein AL, et al. Understanding differences in results from literature-based and individual patient meta-analyses: An example from meta-analyses of observational data. International Journal of Cardiology. 2011 Apr;148(2):209–213. Available from: http://www.sciencedirect.com/science/article/pii/S0167527309015794.

[192] 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 May;60(5):431.e1–431.e12. Available from: http://www.sciencedirect.com/science/article/pii/S0895435606004033.

[193] Jackson D, White IR, Seaman S, Evans H, Baisley K, Carpenter J. Relaxing the independent censoring assumption in the Cox proportional hazards model using multiple imputation. Statistics in Medicine. 2014 Nov;33(27):4681–4694.

[194] Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. Journal of Statistical Software. 2011;45(3). Available from: http://doc.utwente.nl/78938/.

[195] Jolani S, Debray TPA, Koffijberg H, Buuren S, Moons KGM. Imputation of systematically missing predictors in an individual participant data meta-analysis: a generalized approach using MICE. Statistics in Medicine. 2015;34(11):1841–1863.

[196] Audigier V, Resche-Rigon M. micemd: Multiple Imputation by Chained Equations with Multilevel Data; 2019. R package version 1.6.0. Available from: https://CRAN.R-project.org/package=micemd.

[197] Kline D, Andridge R, Kaizar E. Comparing multiple imputation methods for systematically missing subject-level data. Research Synthesis Methods. 2017;8(2):136–148.

[198] Grund S, Lüdtke O, Robitzsch A. Multiple imputation of missing covariate values in multilevel models with random slopes: A cautionary note. Behavior Research Methods. 2016;48(2):640–649.

[199] Marson AG, Williamson PR, Clough H, Hutton JL, Chadwick DW, On Behalf Of The Epilepsy Monotherapy Trial Group. Carbamazepine versus Valproate Monotherapy for Epilepsy: A Meta-analysis. Epilepsia. 2002 May;43(5):505–513. Available from: http://onlinelibrary.wiley.com/doi/10.1046/j.1528-1157.2002.20801.x/abstract.

[200] Therneau TM. coxme: Mixed Effects Cox Models; 2018. R package version 2.2-7. Available from: https://CRAN.R-project.org/package=coxme.

[201] R Core Team. R: A Language and Environment for Statistical Computing; 2018. Available from: https://www.R-project.org/.AccessedSeptember,13,2019.

[202] Jones E, Sweeting MJ, Sharp SJ, Thompson SG, EPIC-InterAct Consortium. A method making fewer assumptions gave the most reliable estimates of exposure-outcome associations in stratified case-cohort studies. Journal of Clinical Epidemiology. 2015 Dec;68(12):1397–1405.

[203] Marson AG, Williamson PR, Hutton JL, Clough HE, Chadwick DW. Carbamazepine versus valproate monotherapy for epilepsy. Cochrane Database of Systematic Reviews. 2000;3.

[204] Lin DY, Zeng D. On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. Biometrika. 2010 Jun;97(2):321–332.

[205] Commenges D, Andersen PK. Score test of homogeneity for survival data. Lifetime Data Analysis. 1995;1(2):145–156; discussion 157–159. Available from: https://link.springer.com/content/pdf/10.1007/BF00985764.pdf.

[206] Gray RJ. Tests for Variation Over Groups in Survival Data. Journal of the American Statistical Association. 1995;90(429):198–203. Available from: http://www.jstor.org/stable/pdf/2291143.pdf.

[207] Claeskens G, Nguti R, Janssen P. One-sided tests in shared frailty models. Test. 2008;17(1):69–82. Available from: http://www.springerlink.com/index/8356T3296V01NR5T.pdf.

[208] Economou P, Stehlík M. On Small Samples Testing for Frailty Through Homogeneity Test. Communications in Statistics - Simulation and Computation. 2015 Jan;44(1):40–65. Available from: http://dx.doi.org/10.1080/03610918.2013.763982.

[209] Biard L, Porcher R, Resche-Rigon M. Permutation tests for centre effect on survival endpoints with application in an acute myeloid leukaemia multicentre study. Statistics in Medicine. 2014 Jul;33(17):3047–3057.

[210] Hox JJ. Multilevel Analysis: Techniques and Applications. 2nd ed. New York, NY: Routledge, Taylor & Francis; 2010.

[211] Goldstein H. Multilevel Statistical Models. 4th ed. Hoboken, N.J.: Wiley; 2011. Available from: http://onlinelibrary.wiley.com/book/10.1002/9780470973394.

[212] Volinsky CT, Raftery AE. Bayesian information criterion for censored survival models. Biometrics. 2000 Mar;56(1):256–262.

[213] Vaida F, Blanchard S. Conditional Akaike Information for Mixed-Effects Models. Biometrika. 2005;92(2):351–370. Available from: http://www.jstor.org/stable/20441193.

[214] Greven S, Kneib T. On the behaviour of marginal and conditional AIC in linear mixed models. Biometrika. 2010 Dec;97(4):773–789. Available from: https://academic.oup.com/biomet/article/97/4/773/241321/On-the-behaviour-of-marginal-and-conditional-AIC.

[215] Campbell H, Dean CB. The consequences of proportional hazards based model selection. Statistics in Medicine. 2014;33(6):1042–1056. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.6021.

[216] Califf RM, Harrell Jr FE. Individual risk prediction using data beyond the medical clinic. Canadian Medical Association Journal. 2018;190(32):E947.

[217] Greving JP, Wermer MJH, Brown Jr RD, Morita A, Juvela S, Yonekura M, et al. Development of the PHASES score for prediction of risk of rupture of intracranial aneurysms: a pooled analysis of six prospective cohort studies. The Lancet Neurology. 2014;13(1):59–66.

[218] Aerts M, Minalu G, Bösner S, Buntinx F, Burnand B, Haasenritter J, et al. Pooled individual patient data from five countries were used to derive a clinical prediction rule for coronary artery disease in primary care. Journal of Clinical Epidemiology. 2017;81:120–128.

[219] Hilkens NA, Algra A, Diener HC, Reitsma JB, Bath PM, Csiba L, et al. Predicting major bleeding in patients with noncardioembolic stroke on antiplatelets: S2TOP-BLEED. Neurology. 2017;89(9):936–943.

[220] Roques F, Nashef SAM, Michel P, Gauducheau E, De Vincentiis C, Baudet E, et al. Risk factors and outcome in European cardiac surgery: analysis of the EuroSCORE multinational database of 19030 patients. European Journal of Cardio-thoracic Surgery. 1999;15(6):816–823.

[221] Ahmed I, Debray TPA, Moons KGM, Riley RD. Developing and validating risk prediction models in an individual participant data meta-analysis. BMC Medical Research Methodology. 2014;14(1):3.

[222] van Houwelingen HC. Validation, calibration, revision and combination of prognostic survival models. Statistics in Medicine. 2000;19(24):3401–3415.

[223] Janssen KJM, Moons KGM, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. Journal of Clinical Epidemiology. 2008 Jan;61(1):76–86. Available from: http://www.sciencedirect.com/science/article/pii/S0895435607002132.

[224] Vergouwe Y, Nieboer D, Oostenbrink R, Debray TP, Murray GD, Kattan MW, et al. A closed testing procedure to select an appropriate method for updating prediction models. Statistics in Medicine. 2017;36(28):4529–4539.

[225] Siontis GCM, Tzoulaki I, Ioannidis JPA. Predicting Death: An Empirical Evaluation of Predictive Tools for Mortality. Archives of Internal Medicine. 2011 Oct;171(19):1721–1726.

[226] 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.

[227] Wynants L, Riley RD, Timmerman D, van Calster B. Random-effects meta-analysis of the clinical utility of tests and prediction models. Statistics in Medicine. 2018 May;37(12):2034–2052. Available from: http://doi.wiley.com/10.1002/sim.7653.

[228] Debray TPA, de Jong VMT, Moons KGM, Riley RD. Evidence synthesis in prognosis research. Diagnostic and Prognostic Research. 2019 Jul;3(1):13. Available from: https://doi.org/10.1186/s41512-019-0059-4.

[229] R Core Team. R: A Language and Environment for Statistical Computing; 2019. Available from: https://www.R-project.org/.AccessedOctober,17,2019.

[230] Bouwmeester W, Moons KGM, Kappen TH, van Klei WA, Twisk JWR, Eijkemans MJC, et al. Internal validation of risk models in clustered data: a comparison of bootstrap schemes. American Journal of Epidemiology. 2013 Jun;177(11):1209–1217.

[231] Snell KIE, Hua H, Debray TPA, Ensor J, Look MP, Moons KGM, et al. Multivariate meta-analysis of individual participant data helped externally validate the performance and implementation of a prediction model. Journal of Clinical Epidemiology. 2016 Jan;69:40–50.

[232] Pennells L, Kaptoge S, White IR, Thompson SG, Wood AM, Emerging Risk Factors Collaboration. Assessing risk prediction models using individual participant data from multiple studies. American Journal of Epidemiology. 2014 Mar;179(5):621–632.

[233] Debray TPA, Koffijberg H, Vergouwe Y, Moons KGM, Steyerberg E. Aggregating published prediction models with individual participant data: a comparison of different approaches. Statistics in Medicine. 2012 Oct;31(23):2697–2712. Available from: http://doi.wiley.com/10.1002/sim.5412.

[234] Debray TPA, Damen JAAG, Riley RD, Snell K, Reitsma JB, Hooft L, et al. A framework for meta-analysis of prediction model studies with binary and time-to-event out-

comes. Statistical Methods in Medical Research. 2019 Sep;28(9):2768–2786. Available from: http://journals.sagepub.com/doi/10.1177/0962280218785504.

[235] Snell KIE, Ensor J, Debray TPA, Moons KGM, Riley RD. Meta-analysis of prediction model performance across multiple studies: Which scale helps ensure between-study normality for the C-statistic and calibration measures? Statistical Methods in Medical Research. 2018;27(11):3505–3522.

[236] Qaseem A, Snow V, Barry P, Hornbake ER, Rodnick JE, Tobolic T, et al. Current diagnosis of venous thromboembolism in primary care: a clinical practice guideline from the American Academy of Family Physicians and the American College of Physicians. Annals of Internal Medicine. 2007;146(6):454–458.

[237] Oudega R, Moons KGM, Hoes AW. Ruling out deep venous thrombosis in primary care. Thrombosis and Haemostasis. 2005;94(01):200–205.

[238] Geersing GJ, Zuithoff NPA, Kearon C, Anderson DR, Cate-Hoek AJt, Elf JL, et al. Exclusion of deep vein thrombosis using the Wells rule in clinically important subgroups: individual patient data meta-analysis. BMJ. 2014 Mar;348:g1340. Available from: https://www.bmj.com/content/348/bmj.g1340.

[239] van Smeden M, de Groot JAH, Moons KGM, Collins GS, Altman DG, Eijkemans MJC, et al. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. BMC Medical Research Methodology. 2016;16(1):16:163.

[240] Puhr R, Heinze G, Nold M, Lusa L, Geroldinger A. Firth's logistic regression with rare events: accurate effect estimates and predictions? Statistics in Medicine. 2017;36(14):2302–2317.

[241] Riley RD, Snell KI, Ensor J, Burke DL, Harrell Jr FE, Moons KG, et al. Minimum sample size for developing a multivariable prediction model: PART II-binary and time-to-event outcomes. Statistics in Medicine. 2019 Mar;38(7):1276–1296.

[242] van Smeden M, Moons KGM, de Groot JAH, Collins GS, Altman DG, Eijkemans MJC, et al. Sample size for binary logistic prediction models: beyond events per variable criteria. Statistical Methods in Medical Research. 2018;28(8):2455–2474.

[243] de Jong VMT, Eijkemans MJC, van Calster B, Timmerman D, Moons KGM, Steyerberg EW, et al. Sample size considerations and predictive performance of multinomial logistic prediction models. Statistics in Medicine. 2019;38(9):1601–1619.

[244] Cox DR. Two Further Applications of a Model for Binary Regression. Biometrika. 1958;45(3/4):562–565. Available from: https://www.jstor.org/stable/2333203.

[245] Steyerberg EW, Eijkemans MJC, Harrell Jr FE, Habbema JDF. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Statistics in Medicine. 2000;19(8):1059–1079.

[246] van Houwelingen J, le Cessie S. Predictive value of statistical models. Statistics in Medicine. 1990;9(11):1303–1325.

[247] David HA. Gini's mean difference rediscovered. Biometrika. 1968;55(3):573–575.

[248] Gerstenberger C, Vogel D. On the efficiency of Gini's mean difference. Statistical Methods & Applications. 2015 Nov;24(4):569–596.

[249] Quartagno M, Carpenter J. jomo: A package for Multilevel Joint Modelling Multiple Imputation; 2019. R package version 2.6-7. Available from: https://CRAN.R-project.org/package=jomo.

[250] Granger CB, Alexander JH, McMurray JJV, Lopes RD, Hylek EM, Hanna M, et al. Apixaban versus Warfarin in Patients with Atrial Fibrillation. New England Journal of Medicine. 2011 Sep;365(11):981–992. Available from: https://doi.org/10.1056/NEJMoa1107039.

[251] Wolf PA, Mitchell JB, Baker CS, Kannel WB, D'Agostino RB. Impact of Atrial Fibrillation on Mortality, Stroke, and Medical Costs. Archives of Internal Medicine. 1998 Feb;158(3):229–234. Available from: https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/191305.

[252] Hylek EM, Go AS, Chang Y, Jensvold NG, Henault LE, Selby JV, et al. Effect of Intensity of Oral Anticoagulation on Stroke Severity and Mortality in Atrial Fibrillation. New England

Journal of Medicine. 2003 Sep;349(11):1019–1026. Available from: https://doi.org/10.1056/NEJMoa022913.

[253] Audigier V, White IR, Jolani S, Debray TPA, Quartagno M, Carpenter J, et al. Multiple Imputation for Multilevel Data with Continuous and Binary Variables. Statistical Science. 2018 May;33(2):160–183. Available from: https://projecteuclid.org/euclid.ss/1525313140.

[254] Global Research on Acute Conditions Team (GREAT) Network. Managing acute heart failure in the ED - case studies from the acute heart failure academy; 2013. Available from: https://www.greatnetwork.org/.

[255] Grund S, Robitzsch A, Luedtke O. mitml: Tools for Multiple Imputation in Multilevel Modeling; 2019. R package version 0.3-7. Available from: https://CRAN.R-project.org/package=mitml.

[256] Geersing GJ, Janssen KJM, Oudega R, Bax L, Hoes AW, Reitsma JB, et al. Excluding venous thromboembolism using point of care D-dimer tests in outpatients: a diagnostic meta-analysis. BMJ. 2009 Aug;339:b2990.

[257] Geersing GJ, Toll DB, Janssen KJM, Oudega R, Blikman MJC, Wijland R, et al. Diagnostic Accuracy and User-Friendliness of 5 Point-of-Care D-Dimer Tests for the Exclusion of Deep Vein Thrombosis. Clinical Chemistry. 2010 Nov;56(11):1758–1766.

[258] Vergouwe Y, Royston P, Moons KGM, Altman DG. Development and validation of a prediction model with missing predictor data: a practical approach. Journal of Clinical Epidemiology. 2010 Feb;63(2):205–214.

[259] Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychological Methods. 2001;6(4):330–351.

[260] Pajouheshnia R, van Smeden M, Peelen LM, Groenwold RHH. How variation in predictor measurement affects the discriminative ability and transportability of a prediction model. Journal of clinical epidemiology. 2019;105:136–141.

[261] Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in medicine. 1996;15(4):361–387.

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

[263] Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Annals of internal medicine. 1999;130(6):515–524.

[264] Steyerberg EW, Moons KGM, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. PLOS Medicine. 2013 Feb;10(2):e1001381.

[265] Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. LWW; 2000.

[266] Sato T, Matsuyama Y. Marginal structural models as a tool for standardization. Epidemiology. 2003;14(6):680–686.

[267] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41–55.

[268] Shmueli G. To Explain or to Predict? Statistical Science. 2010 Aug;25(3):289–310. Available from: http://projecteuclid.org/euclid.ss/1294167961.

[269] Hernán MA. The C-word: scientific euphemisms do not improve causal inference from observational data. American journal of public health. 2018;108(5):616–619.

[270] Nowacki AS, Wells BJ, Yu C, Kattan MW. Adding propensity scores to pure prediction models fails to improve predictive performance. PeerJ. 2013;1:e123.

[271] Groenwold RHH, Moons KGM, Pajouheshnia R, Altman DG, Collins GS, Debray TPA, et al. Explicit inclusion of treatment in prognostic modeling was recommended in observational and randomized settings. Journal of Clinical Epidemiology. 2016 Oct;78:90–100. Available from: http://www.sciencedirect.com/science/article/pii/S0895435616300300.

[272] Pajouheshnia R, Peelen LM, Moons KGM, Reitsma JB, Groenwold RHH. Accounting for treatment use when validating a prognostic model: a simulation study. BMC Medical Research Methodology. 2017 Jul;17(1):103. Available from: https://doi.org/10.1186/s12874-017-0375-8.

[273] Pajouheshnia R, Groenwold RH, Peelen LM, Reitsma JB, Moons KG. When and how to use data from randomised trials to develop or validate prognostic models. bmj. 2019;365:l2154.

[274] Wu S, Flach P. A scored AUC Metric for Classifier Evaluation and Selection. Proceedings of the ICML 2005 workshop on ROC Analysis in Machine Learning. 2005;.

[275] Li J, Fine JP. Weighted area under the receiver operating characteristic curve and its application to gene selection. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2010;59(4):673–692.

[276] Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Statistics in Medicine. 2000;19(9):1141–1164. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-0258%2820000515%2919%3A9%3C1141%3A%3AAID-SIM479%3E3.0.CO%3B2-F.

[277] Stuart EA, Cole SR, Bradshaw CP, Leaf PJ. The use of propensity scores to assess the generalizability of results from randomized trials. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2011;174(2):369–386.

[278] Vo T, Porcher R, Chaimani A, Vansteelandt S. A novel approach for identifying and addressing case-mix heterogeneity in individual participant data meta-analysis. Research Synthesis Methods. 2019 Nov;p. jrsm.1382. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/jrsm.1382.

[279] Stewart LA, Tierney JF. To IPD or not to IPD? Advantages and disadvantages of systematic reviews using individual patient data. Evaluation & the health professions. 2002;25(1):76–97.

[280] van Smeden M, Lash TL, Groenwold RH, van Smeden M. Five myths about measurement error in epidemiologic research. 2019;Available from: https://www.researchgate.net/profile/Maarten_Van_Smeden/publication/333747171_Five_myths_about_measurement_error_in_epidemiologic_research/links/5d01f6a24585157d15a6b78f/Five-myths-about-measurement-error-in-epidemiologic-research.pdf.

[281] Brakenhoff TB, Mitroiu M, Keogh RH, Moons KGM, Groenwold RHH, van Smeden M. Measurement error is often neglected in medical literature: a systematic review. Journal of clinical epidemiology. 2018;98:89–97.

[282] Falley BN, Stamey JD, Beaujean AA. Bayesian estimation of logistic regression with misclassified covariates and response. Journal of Applied Statistics. 2018 Jul;45(10):1756–1769. Available from: https://doi.org/10.1080/02664763.2017.1391182.

[283] Nelson T, Song JJ, Chin YM, Stamey JD. Bayesian Correction for Misclassification in Multilevel Count Data Models. Computational and mathematical methods in medicine. 2018;2018.

[284] Lian Q, Hodges JS, MacLehose R, Chu H. A Bayesian approach for correcting exposure misclassification in meta-analysis. Statistics in Medicine. 2019;38(1):115–130. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.7969.

[285] Jaenisch T, Tam DTH, Kieu NTT, Van Ngoc T, Nam NT, Van Kinh N, et al. Clinical evaluation of dengue and identification of risk factors for severe disease: protocol for a multicentre study in 8 countries. BMC infectious diseases. 2016;16(1):120.

[286] Anders KL, Nguyet NM, Van Vinh Chau N, Hung NT, Thuy TT, Lien LB, et al. Epidemiological Factors Associated with Dengue Shock Syndrome and Mortality in Hospitalized Dengue Patients in Ho Chi Minh City, Vietnam. The American Journal of Tropical Medicine and Hygiene. 2011 Jan;84(1):127–134. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3005500/.

[287] Keogh RH, White IR. A toolkit for measurement error correction, with a focus on nutritional epidemiology. Statistics in medicine. 2014;33(12):2137–2155.

[288] Cole SR, Chu H, Greenland S. Multiple-imputation for measurement-error correction. International Journal of Epidemiology. 2006 Aug;35(4):1074–1081. Available from: https://academic.oup.com/ije/article/35/4/1074/686404.

[289] Richardson S, Gilks WR. A Bayesian approach to measurement error problems in epidemiology using conditional independence models. American Journal of Epidemiology. 1993;138(6):430–442.

[290] Williams DR, Rast P, Bürkner PC. Bayesian meta-analysis with weakly informative prior distributions. 2018;Available from: https://osf.io/7tbrm/download?format=pdf.

[291] Gelman A. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). Bayesian analysis. 2006;1(3):515–534.

[292] Richardson S, Gilks WR. Conditional independence models for epidemiological studies with covariate measurement error. Statistics in Medicine. 1993;12(18):1703–1722.

[293] Meng XL. Multiple-imputation inferences with uncongenial sources of input. Statistical Science. 1994;p. 538–558.

[294] Goldstein ND, Welles SL, Burstyn I. To Be or Not to Be: Bayesian Correction for Misclassification of Self-reported Sexual Behaviors Among Men Who Have Sex with Men. Epidemiology. 2015 Sep;26(5):637–644. Available from: http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00001648-201509000-00003.

[295] Spiegelman D, Rosner B, Logan R. Estimation and inference for logistic regression with covariate misclassification and measurement error in main study/validation study designs. Journal of the American Statistical Association. 2000;95(449):51–61.

[296] Collins G, Reitsma J, Altman D, Moons K. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMC medicine. 2015;13(1):1.

[297] Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): explanation and elaboration. Annals of internal medicine. 2015;162(1):W1–W73.

[298] Biesheuvel CJ, Vergouwe Y, Steyerberg EW, Grobbee DE, Moons KGM. Polytomous logistic regression analysis could be applied more often in diagnostic research. Journal of clinical epidemiology. 2008;61(2):125–134.

[299] Moons KGM, Grobbee DE. Diagnostic studies as multivariable, prediction research. Journal of epidemiology and community health. 2002;56(5):337–338.

[300] Schuit E, Kwee A, Westerhuis MEMH, van Dessel HJHM, Graziosi GCM, van Lith JMM, et al. A clinical prediction model to assess the risk of operative delivery. BJOG: An International Journal of Obstetrics & Gynaecology. 2012;119(8):915–923.

[301] Barnes DE, Mehta KM, Boscardin WJ, Fortinsky RH, Palmer RM, Kirby KA, et al. Prediction of recovery, dependence or death in elders who become disabled during hospitalization. Journal of general internal medicine. 2013;28(2):261–268.

[302] van Calster B, Valentin L, van Holsbeke C, Testa A, Bourne T, van Huffel S, et al. Polytomous diagnosis of ovarian tumors as benign, borderline, primary invasive or metastatic: development and validation of standard and kernel-based risk prediction models. BMC medical research methodology. 2010;10(1):1.

[303] Roukema J, van Loenhout RB, Steyerberg EW, Moons KGM, Bleeker SE, Moll H. Polytomous regression did not outperform dichotomous logistic regression in diagnosing serious bacterial infections in febrile children. Journal of clinical epidemiology. 2008;61(2):135–141.

[304] Steyerberg EW, Eijkemans MJC, Harrell Jr FE, Habbema JDF. Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. Medical Decision Making. 2001;21(1):45–56.

[305] Ambler G, Brady AR, Royston P. Simplifying a prognostic model: a simulation study based on clinical data. Statistics in medicine. 2002;21(24):3803–3822.

[306] Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. Journal of clinical epidemiology. 1996;49(12):1373–1379.

[307] Courvoisier DS, Combescure C, Agoritsas T, Gayet-Ageron A, Perneger TV. Performance of Logistic Regression modeling: beyond the number of events per variable, the role of data structure. Journal of Clinical Epidemiology. 2011;64(1):993–1000.

[308] Austin PC, Steyerberg EW. Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. Statistical methods in medical research. 2017;26(2):796–808.

[309] Smith GCS, Seaman SR, Wood AM, Royston P, White IR. Correcting for optimistic prediction in small data sets. American journal of epidemiology. 2014;180(3):318–324.

[310] Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Methodological). 1996;p. 267–288.

[311] le Cessie S, van Houwelingen JC. Ridge estimators in logistic regression. Applied statistics. 1992;p. 191–201.

[312] Agresti A. Categorical Data Analysis. 2nd ed. New Jersey: John Willey and Sons; 2002.

[313] Hoerl AE, Kennard RW. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics. 1970;12(1):55–67.

[314] Hoerl AE, Kennard RW. Ridge regression: applications to nonorthogonal problems. Technometrics. 1970;12(1):69–82.

[315] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. 2nd ed. Springer Series in Statistics. New York: Springer-Verlag; 2009. Available from: `//www.springer.com/gp/book/9780387848570`.

[316] Friedman JH, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of statistical software. 2010;33(1):1. Available from: `https://www.jstatsoft.org/article/view/v033i01`.

[317] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982 Apr;143(1):29–36. Available from: `http://pubs.rsna.org/doi/abs/10.1148/radiology.143.1.7063747`.

[318] Harrell Jr FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. Jama. 1982;247(18):2543–2546.

[319] van Calster B, van Belle V, Vergouwe Y, Timmerman D, van Huffel S, Steyerberg EW. Extending the c-statistic to nominal polytomous outcomes: the Polytomous Discrimination Index. Statistics in medicine. 2012;31(23):2610–2626.

[320] van Hoorde K, Vergouwe Y, Timmerman D, van Huffel S, Steyerberg EW, van Calster B. Assessing calibration of multinomial risk prediction models. Statistics in medicine. 2014;33(15):2585–2596.

[321] Miller ME, Hui SL, Tierney WM. Validation techniques for logistic regression models. Statistics in medicine. 1991;10(8):1213–1226.

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

[323] Nagelkerke NJD. A note on a general definition of the coefficient of determination. Biometrika. 1991;78(3):691–692.

[324] Harrell FE, Lee KL, Mark DB. Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in medicine. 1996;15:361–387.

[325] Pencina MJ, D'agostino RB, Pencina KM, Janssens ACJW, Greenland P. Interpreting incremental value of markers added to risk prediction models. American journal of epidemiology. 2012;176(6):473–481.

[326] Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. American journal of epidemiology. 2007;165(6):710–718.

[327] Croissant Y. mlogit: Multinomial Logit Models; 2019. Available from: `https://CRAN.R-project.org/package=mlogit.AccessedSeptember,13,2019`.

[328] Henningsen A, Toomet O. maxLik: A package for maximum likelihood estimation in R. Computational Statistics. 2011;26(3):443–458.

[329] Timmerman D, Testa AC, Bourne T, Ferrazzi E, Ameye L, Konstantinovic ML, et al. Logistic regression model to distinguish between the benign and malignant adnexal mass before surgery: a multicenter study by the International Ovarian Tumor Analysis Group. Journal of Clinical Oncology. 2005;23(34):8794–8801.

[330] Pavlou M, Ambler G, Seaman S, De Iorio M, Omar RZ. Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events. Statistics in medicine. 2016;35(7):1159–1177.

[331] Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2005;67(2):301–320.

[332] Zou H, Hastie T, Tibshirani R. On the "degrees of freedom" of the lasso. The Annals of Statistics. 2007;35(5):2173–2192.

[333] Albert A, Anderson JA. On the Existence of Maximum Likelihood Estimates in Logistic Regression Models. Biometrika. 1984 Apr;71(1):1–10.

[334] Santner TJ, Duffy DE. A note on A. Albert and JA Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models. Biometrika. 1986;73(3):755–758.

[335] Ogundimu EO, Altman DG, Collins GS. Adequate sample size for developing prediction models is not simply related to events per variable. Journal of clinical epidemiology. 2016;76:175–182.

[336] Sutton AJ, Cooper NJ, Jones DR. Evidence synthesis as the key to more coherent and efficient research. BMC medical research methodology. 2009;9(1):29.

[337] Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ : British Medical Journal. 1996 Jan;312(7023):71–72. Available from: `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2349778/`.

[338] Lau J, Ioannidis JPA, Schmid CH. Summing up evidence: one answer is not always enough. The lancet. 1998;351(9096):123–127.

[339] Egger M, Smith GD. Meta-analysis: potentials and promise. Bmj. 1997;315(7119):1371–1374.

[340] Hemingway H, Croft P, Perel P, Hayden JA, Abrams K, Timmis A, et al. Prognosis research strategy (PROGRESS) 1: a framework for researching clinical outcomes. Bmj. 2013;346:e5595.

[341] Riley RD, Hayden JA, Steyerberg EW, Moons KGM, Abrams K, Kyzas PA, et al. Prognosis Research Strategy (PROGRESS) 2: prognostic factor research. PLoS medicine. 2013;10(2):e1001380.

[342] Hingorani AD, van der Windt DA, Riley RD, Abrams K, Moons KG, Steyerberg EW, et al. Prognosis research strategy (PROGRESS) 4: stratified medicine research. Bmj. 2013;346:e5793.

[343] Rothwell PM. Can overall results of clinical trials be applied to all patients? The Lancet. 1995;345(8965):1616–1619.

[344] Damen JAAG, Hooft L. The increasing need for systematic reviews of prognosis studies: strategies to facilitate review production and improve quality of primary research. Diagnostic and Prognostic Research. 2019;3(1):2.

[345] Altman DG. Systematic reviews of evaluations of prognostic variables. Bmj. 2001;323(7306):224–228.

[346] Moons KGM, de Groot JAH, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS medicine. 2014;11(10):e1001744.

[347] Hayden JA, van der Windt DA, Cartwright JL, Côté P, Bombardier C. Assessing bias in studies of prognostic factors. Annals of internal medicine. 2013;158(4):280–286.

[348] Sauerbrei W, Holländer N, Riley RD, Altman DG. Evidence-based assessment and application of prognostic markers: the long way from single studies to meta-analysis. Communications in statistics-theory and methods. 2006;35(7):1333–1342.

[349] Riley RD, Moons KGM, Snell KIE, Ensor J, Hooft L, Altman DG, et al. A guide to systematic review and meta-analysis of prognostic factor studies. bmj. 2019;364:k4597.

[350] Borenstein M, Hedges LV, Higgins JP, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. Research Synthesis Methods. 2010;1(2):97–111.

[351] van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. Statistics in medicine. 2002;21(4):589–624.

[352] Rice K, Higgins JPT, Lumley T. A re-evaluation of fixed effect (s) meta-analysis. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2018;181(1):205–227.

[353] Zhang Y, Zhao D, Gong C, Zhang F, He J, Zhang W, et al. Prognostic role of hormone receptors in endometrial cancer: a systematic review and meta-analysis. World journal of surgical oncology. 2015;13(1):208.

[354] Jackson D, Riley R, White IR. Multivariate meta-analysis: potential and promise. Statistics in medicine. 2011;30(20):2481–2498.

[355] Reitsma JB, Glas AS, Rutjes AWS, Scholten RJPM, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. Journal of clinical epidemiology. 2005;58(10):982–990.

[356] Riley RD. Multivariate meta-analysis: the effect of ignoring within-study correlation. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2009;172(4):789–811.

[357] Yoneoka D, Henmi M. Synthesis of linear regression coefficients by recovering the within-study covariance matrix from summary statistics. Research Synthesis Methods. 2017;8(2):212–219.

[358] Riley RD, Elia EG, Malin G, Hemming K, Price MP. Multivariate meta-analysis of prognostic factor studies with multiple cut-points and/or methods of measurement. Statistics in Medicine. 2015 Jul;34(17):2481–2496. Available from: http://doi.wiley.com/10.1002/sim.6493.

[359] Shi JQ, Copas JB. Meta-analysis for trend estimation. Statistics in Medicine. 2004 Jan;23(1):3–19. Available from: http://doi.wiley.com/10.1002/sim.1595.

[360] Berlin JA, Longnecker MP, Greenland S. Meta-analysis of epidemiologic dose-response data. Epidemiology. 1993;p. 218–228.

[361] Debray TPA, Moons KGM, Abo-Zaid GMA, Koffijberg H, Riley RD. Individual Participant Data Meta-Analysis for a Binary Outcome: One-Stage or Two-Stage? PLoS ONE. 2013 Apr;8(4):e60650. Available from: https://dx.plos.org/10.1371/journal.pone.0060650.

[362] Trivella M, Pezzella F, Pastorino U, Harris AL, Altman DG. Microvessel density as a prognostic factor in non-small-cell lung carcinoma: a meta-analysis of individual patient data. The Lancet Oncology. 2007 Jun;8(6):488–499. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1470204507701456.

[363] Sauerbrei W, Royston P. A new strategy for meta-analysis of continuous covariates in observational studies. Statistics in Medicine. 2011 Dec;30(28):3341–3360.

[364] Royston P, Sauerbrei W. A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials. Statistics in Medicine. 2004 Aug;23(16):2509–2525. Available from: http://doi.wiley.com/10.1002/sim.1815.

[365] Gasparrini A, Armstrong B, Kenward MG. Multivariate meta-analysis for non-linear and other multi-parameter associations. Statistics in Medicine. 2012 Dec;31(29):3821–3839. Available from: http://doi.wiley.com/10.1002/sim.5471.

[366] Abo-Zaid G, Sauerbrei W, Riley RD. Individual participant data meta-analysis of prognostic factor studies: state of the art? BMC Medical Research Methodology. 2012 Dec;12(1):56. Available from: http://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-12-56.

[367] den Ruijter HM, Peters SAE, Anderson TJ, Britton AR, Dekker JM, Eijkemans MJC, et al. Common Carotid Intima-Media Thickness Measurements in Cardiovascular Risk Prediction: A Meta-analysis. JAMA. 2012 Aug;308(8):796. Available from: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2012.9630.

[368] Yoneoka D, Henmi M, Sawada N, Inoue M. Synthesis of clinical prediction models under different sets of covariates with one individual patient data. BMC Medical Research Methodology. 2015 Dec;15(1):101. Available from: http://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-015-0087-x.

[369] Debray TP, Koffijberg H, Lu D, Vergouwe Y, Steyerberg EW, Moons KGM. Incorporating published univariable associations in diagnostic and prognostic modeling. BMC Medical Research Methodology. 2012 Dec;12(1):121. Available from: http://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-12-121.

[370] Claeys KC, Zasowski EJ, Lagnf AM, Levine DP, Davis SL, Rybak MJ. Novel application of published risk factors for methicillin-resistant S. aureus in acute bacterial skin and skin structure infections. International Journal of Antimicrobial Agents. 2018 Jan;51(1):43–46. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0924857917302261.

[371] The Fibrinogen Studies Collaboration. Systematically missing confounders in individual participant data meta-analysis of observational cohort studies. Statistics in Medicine. 2009 Apr;28(8):1218–1237. Available from: http://doi.wiley.com/10.1002/sim.3540.

[372] Becker BJ, Wu MJ. The synthesis of regression slopes in meta-analysis. Statistical science. 2007;22(3):414–429.

[373] Kovačić J, Varnai VM. A graphical model approach to systematically missing data in meta-analysis of observational studies. Statistics in Medicine. 2016 Oct;35(24):4443–4458. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.7010.

[374] Wynants L, Kent DM, Timmerman D, Lundquist CM, Van Calster B. Untapped potential of multicenter studies: a review of cardiovascular risk prediction models revealed inappropriate analyses and wide variation in reporting. Diagnostic and Prognostic Research. 2019 Dec;3(1):6. Available from: https://diagnprognres.biomedcentral.com/articles/10.1186/s41512-019-0046-9.

[375] Stijnen T, Hamza TH, Özdemir P. Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data. Statistics in Medicine. 2010 Dec;29(29):3046–3067. Available from: http://doi.wiley.com/10.1002/sim.4040.

[376] van Doorn S, Debray TPA, Kaasenbrood F, Hoes AW, Rutten FH, Moons KGM, et al. Predictive performance of the CHA2DS2-VASc rule in atrial fibrillation: a systematic review and meta-analysis. Journal of Thrombosis and Haemostasis. 2017 Jun;15(6):1065–1077. Available from: http://doi.wiley.com/10.1111/jth.13690.

[377] van Klaveren D, Steyerberg EW, Perel P, Vergouwe Y. Assessing discriminative ability of risk models in clustered data. BMC Medical Research Methodology. 2014 Dec;14(1):5. Available from: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-14-5.

[378] Haile SR, Guerra B, Soriano JB, Puhan MA. Multiple Score Comparison: a network meta-analysis approach to comparison and external validation of prognostic scores. BMC medical research methodology. 2017;17(1):172.

[379] Westeneng HJ, Debray TP, Visser AE, van Eijk RP, Rooney JP, Calvo A, et al. Prognosis for patients with amyotrophic lateral sclerosis: development and validation of a personalised prediction model. The Lancet Neurology. 2018;17(5):423–433.

[380] Martin GP, Mamas MA, Peek N, Buchan I, Sperrin M. A multiple-model generalisation of updating clinical prediction models. Statistics in Medicine. 2018 Apr;37(8):1343–1358. Available from: http://doi.wiley.com/10.1002/sim.7586.

[381] Martin GP, Mamas MA, Peek N, Buchan I, Sperrin M. Clinical prediction in defined populations: a simulation study investigating when and how to aggregate existing models. BMC Medical Research Methodology. 2017 Dec;17(1):1. Available from: http://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-016-0277-1.

[382] Su TL, Jaki T, Hickey GL, Buchan I, Sperrin M. A review of statistical updating methods for clinical prediction models. Statistical methods in medical research. 2018;27(1):185–197.

[383] Merz CJ, Pazzani MJ. A principal components approach to combining regression estimates. Machine learning. 1999;36(1-2):9–32.

[384] Collins GS, de Groot JAH, Dutton S, Omar O, Shanyinde M, Tajar A, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Medical Research Methodology. 2014 Dec;14(1):40. Available from: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-14-40.

[385] Whittle R, Peat G, Belcher J, Collins GS, Riley RD. Measurement error and timing of predictor values for multivariable risk prediction models are poorly reported. Journal of Clinical Epidemiology. 2018 Oct;102:38–49. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0895435618300374.

[386] Pavlou M, Ambler G, Seaman S, Omar RZ. A note on obtaining correct marginal predictions from a random intercepts model for binary outcomes. BMC Medical Research Methodology. 2015 Dec;15(1):59. Available from: http://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-015-0046-6.

[387] Griffith LE, van den Heuvel E, Fortier I, Sohel N, Hofer SM, Payette H, et al. Statistical approaches to harmonize data on cognitive measures in systematic reviews are rarely reported. Journal of Clinical Epidemiology. 2015 Feb;68(2):154–162. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0895435614003497.

[388] Bartlett JW, Keogh RH. Bayesian correction for covariate measurement error: A frequentist evaluation and comparison with regression calibration. Statistical Methods in Medical Research. 2018 Jun;27(6):1695–1708. Available from: http://journals.sagepub.com/doi/10.1177/0962280216667764.

[389] Hossain S, Gustafson P. Bayesian adjustment for covariate measurement errors: A flexible parametric approach. Statistics in Medicine. 2009 May;28(11):1580–1600. Available from: http://doi.wiley.com/10.1002/sim.3552.

[390] Grund S, Lüdtke O, Robitzsch A. Multiple Imputation of Missing Data for Multilevel Models: Simulations and Recommendations. Organizational Research Methods. 2018 Jan;21(1):111–149. Available from: http://journals.sagepub.com/doi/10.1177/1094428117703686.

[391] Kunkel D, Kaizar EE. A comparison of existing methods for multiple imputation in individual participant data meta-analysis. Statistics in Medicine. 2017 Sep;36(22):3507–3532. Available from: http://doi.wiley.com/10.1002/sim.7388.

Appendices

Summary

Prediction models are developed, validated and used for the prediction of a patient's current (diagnostic prediction models) or future (prognostic prediction models) health status, and may thereby aid in medical decision making and to inform patients on their health. Risk predictions can be used to make decisions regarding the need for additional diagnostic tests, initiating life-style changes or other preventive strategies, identifying the most effective treatment for an individual and for benchmarking the quality of medical centers. Prediction models should be developed and validated in large samples from multiple populations and settings. This requires research groups to join efforts by sharing their individual participant data (IPD) and subsequently applying adequate statistical methods to synthesize the data across studies or research centers.

Many randomized trials evaluate an intervention effect on time-to-event outcomes. IPD from such trials can be obtained and combined in a so-called IPD meta-analysis (IPD-MA), to summarize the overall intervention effect. In **chapter 2** we present a narrative literature review to provide an overview of methods for conducting an IPD-MA of randomized intervention studies with a time-to-event outcome. We focused on identifying good methodological practice for modeling frailty of trial participants across trials, modeling heterogeneity of intervention effects, choosing appropriate association measures, dealing with (trial differences in) censoring and follow-up times, and addressing time-varying intervention effects and effect modification (interactions). We discuss how to achieve this using parametric and semi-parametric methods, and describe how to implement these in a one-stage or two-stage IPD-MA framework. We recommend exploring heterogeneity of the effect(s) through interaction and non-linear effects. Random effects should be applied to account for residual heterogeneity of the intervention effect. We provide further recommendations, many of which specific to IPD-MA of time-to-event data from randomized trials examining an intervention effect. We illustrate several key methods in a real IPD-MA, where IPD of 1225 participants from 5 randomized clinical trials were combined to compare the effects of Carbamazepine and Valproate on the incidence of epileptic seizures.

Prediction models often yield inaccurate predictions for new individuals. Although large data sets from individual participant data meta-analysis or electronic healthcare records may alleviate this, prevailing strategies for prediction model development generally do not account for heterogeneity between settings and populations. This limits the generalizability of developed models (even from large, combined, clustered data sets) and necessitates local revisions. In **chapter 3** we develop methodology for producing prediction models that are more robust and require less tailoring when applied to different settings and populations. We adopt Internal-External Cross-Validation to assess and reduce heterogeneity in a model's predictive performance during its development. We propose a predictor selection algorithm that optimizes the (weighted) average performance whilst minimizing its variability across the hold-out clusters (or studies). Predictors are added iteratively until the estimated generalizability is optimized. We illustrate this methodology by

developing a new model for predicting the risk of atrial fibrillation and updating an existing one for diagnosing deep vein thrombosis. We used individual participant data from 20 cohorts (N = 10873) and 11 diagnostic studies (N = 10014), respectively. Meta-analysis of calibration and discrimination in each hold-out cluster shows that trade-offs between average performance and heterogeneity occurred. Our methodology allows for the assessment of heterogeneity of prediction model performance during model development in multiple or clustered data sets, thereby informing researchers on predictor selection to minimize heterogeneity. This may improve the generalizability to different settings and populations, and reduce the need for tailoring the model.

Many prediction models perform worse when applied to new individuals, which may be due to a (lack of) representativeness of the validation sample for the prediction model. If the validation sample does not fully represent the model's intended target population, estimates of model performance in the validation set are misleading. In **chapter 4** we consider the use of propensity score weighting methods to standardize predictive performance measures estimated in multiple validation samples that are obtained from different but related populations and settings, by weighting with respect to the covariate distribution of the target population and setting. We show how standardized measures for a model's discrimination and calibration can be derived.

We illustrate our methods in a motivating example on the validation of eight different diagnostic prediction models for the detection of deep vein thrombosis (DVT) that may aid in the diagnosis of patients suspected of DVT in 12 external validation data sets. We applied random effects meta-analysis to analyze the estimates of prediction models' performance across these 12 external validation data sets. The between-study heterogeneity estimates of the random effects meta-analysis indicate that differences between discriminatory performance in the individual validation studies can partially be attributed to differences in case-mix, rather than the use of invalid model coefficients. Further, the meta-analysis showed that the between-study heterogeneity for the calibration slopes was increased by standardization for all models. This demonstrates that there were differences in case-mix between the development and validation samples, and that the case-mix differences partially masked the differences in optimal coefficients between these samples. When standardization filtered out these differences in case-mix, heterogeneity in the calibration slopes became more apparent. Propensity score-based standardization may help to facilitate the interpretation of (heterogeneity in) prediction model performance across multiple external validation studies and to guide model updating strategies or to accept that the validation sample does not reflect the target population of a developed model.

A common problem in the retrieval and analysis of multiple data sources, such as IPD-MA, is the presence of measurement error. Misclassification of binary predictors arises when these study variables are not accurately measured. The presence of misclassification may introduce bias in estimates of parameters (including predictor effects), even when the error is entirely random. Although several methods

for addressing misclassification during the development of a prediction model have been proposed, these do not account for the heterogeneity that is often present in IPD-MA. In **chapter 5** we develop a Bayesian framework for addressing predictor misclassification in an IPD-MA, where the extent and nature of measurement error may vary across studies and may be dependent on participant-level characteristics. This facilitates unbiased estimation of adjusted and unadjusted predictor-outcome associations, as well as unbiased estimates of between-study heterogeneity.

We illustrate our methodology in a motivating example of the diagnosis of dengue using two predictor variables. In this example, the gold standard measurement for one predictor variable is unavailable for half of the studies. Instead, these studies only measured a surrogate that is prone to misclassification. Our methods reduced the error in the estimates for the predictor-outcome association. In general, our methods yielded estimates with less error than an analysis that was naive with regard to measurement error and an analysis based on gold standard measurements alone. Estimates for heterogeneity of the predictor-outcome association were similar across all investigated methods. Further, our simulations show that our framework can appropriately accounted for misclassification that is dependent on study- and participant-level information. By implementing a proposed misclassification model that models participant-level effects and heterogeneity between studies in the outcome and gold standard and surrogate measurement of the predictor, we obtained valid estimates of the predictor-outcome association, with less RMSE, greater power and similar coverage compared to an analysis that was restricted to observations for which gold standard measurements were available. Heterogeneity estimates were adequate for all studied models. Our proposed framework can be used to address the presence of misclassification of a predictor variable in an IPD-MA. This framework requires that some studies supply IPD for the surrogate predictor and the gold standard predictor and misclassification is exchangeable across studies conditional on the observed covariates (and outcome).

Multinomial Logistic Regression (MLR) has been advocated for developing clinical prediction models that distinguish between three or more unordered outcomes. In **chapter 6** we present a full-factorial simulation study to examine the predictive performance of MLR models in relation to the relative size of outcome categories, number of predictors and the number of events per variable. It is shown that MLR estimated by maximum likelihood yields overfitted prediction models in small to medium sized data. In most cases, the calibration and overall predictive performance of the multinomial prediction model is improved by using penalized MLR. Our simulation study also highlights the importance of events per variable in the multinomial context as well as the total sample size. As expected, our study demonstrates the need for optimism correction of the predictive performance measures when developing the multinomial logistic prediction model. We recommend the use of penalized MLR when prediction models are developed in small data sets, or in medium sized data sets with a small total sample size (i.e. when the sizes of the outcome categories are balanced). Finally, we present a motivating example in which we illustrate the development and validation of penalized and unpenalized multinomial prediction models for predicting malignancy of ovarian cancer.

Finally, in **chapter 7** we provide an overview of meta-analysis methods for prognosis research. Over the past few years, evidence synthesis has become essential to investigate and improve the generalizability of medical research findings. This strategy often involves a meta-analysis to formally summarize quantities of interest, such as relative treatment effect estimates. The use of meta-analysis methods is, however, less straightforward in prognosis research because substantial variation exists in research objectives, analysis methods and in the level of reported evidence. We present a gentle overview of statistical methods that can be used to summarize data of prognostic factor and prognostic model studies. We discuss how aggregate data, individual participant data or a combination thereof can be combined through meta-analysis methods. Recent examples are provided throughout to illustrate the various methods. We finish with providing general recommendations for performing an IPD-MA in prediction modeling research.

Samenvatting

Voorspellingsmodellen worden gebruikt om de huidige (diagnostische voorspellingsmodellen) of toekomstige (prognostische voorspellingsmodellen) gezondheidsstatus van een patiënt te berekenen. Ze kunnen daardoor helpen bij het nemen van medische beslissingen en het informeren van patiënten over hun gezondheid. Op basis van voorspellingen van risico's kunnen beslissingen worden genomen over de noodzaak van het afnemen van extra diagnostische tests, het veranderen van de levensstijl van de patiënt, het kiezen voor andere preventieve strategieën, het identificeren van de meest effectieve behandeling voor een patiënt en voor het beoordelen van de kwaliteit van medische centra. Voorspellingsmodellen dienen ontwikkeld en getest te worden met behulp van data die voldoende representatief is. Dit vereist vaak het combineren van individuele patiëntdata (IPD) uit meerdere populaties en studies en het bundelen van de krachten van verschillende onderzoeksgroepen. Voor het analyseren van gecombineerde datasets zijn geavanceerde statistische methoden nodig. Tot op heden zijn zulke methoden voor diagnostische en prognostische toepassingen nog onvoldoende onderzocht. In dit proefschrift heb ik onderzocht welke statistische methoden beschikbaar zijn voor het analyseren van gecombineerde datasets bij wetenschappelijke vraagstukken die zich richten op interventie, diagnose en prognose. Daarnaast heb ik hiervoor nieuwe statistische methoden ontwikkeld voor diagnostisch en prognostisch onderzoek. Ook in interventieonderzoek is heterogeniteit aanwezig in gecombineerde en geclusterde datasets. Op dat gebied zijn al vele methoden beschikbaar die daarmee rekening houden. Daarom begint dit proefschrift met een overzicht daarvan.

IPD-meta-analyse in interventieonderzoek

In vele gerandomiseerde experimentele onderzoeken wordt het longitudinale effect (het effect door de tijd heen) van een interventie onderzocht. Wanneer data uit meerdere studies beschikbaar zijn en samengevoegd worden in een zogenaamde IPD-meta-analyse (IPD-MA), zijn geavanceerde methoden nodig om om te gaan met verschillen tussen studies. In **hoofdstuk 2** presenteren we een literatuuronderzoek waar we een overzicht van methoden weergeven voor het uitvoeren van een IPD-MA van gerandomiseerde interventieonderzoeken die als uitkomst een tijd tot een gebeurtenis hebben. We hebben ons hierbij gericht op een aantal punten: het identificeren van correcte werkwijzen voor het modeleren van de verschillen in *frailty* (ongemeten onderliggende gezondheid) van onderzoeksdeelnemers tussen onderzoeken, het modelleren van heterogeniteit van effecten van interventies, het kiezen van geschikte associatiematen, het omgaan met (verschillen tussen onderzoeken in) *censoring* (de gebeurtenis is voor een onderzoeksdeelnemer tijdens het onderzoek niet voorgekomen, maar mogelijk wel daarna) en opvolgingstijden en tot slot het modeleren van tijdsafhankelijke effecten van interventies en effectmodificatie (interacties). We bediscussiëren hoe dit bereikt kan worden met parametrische en semi-parametrische methoden en beschrijven hoe dit geïmplementeerd kan worden in een één-staps- en twee-staps-IPD-MA-raamwerk. We raden aan om heterogeniteit van het effect van een interventie uit te zoeken door middel van interactietermen

en niet-lineaire effecten. *Random effects* (effecten die een gegeven verdeling volgen) moeten gebruikt worden om rekening te houden met residuele heterogeniteit van het effect van de interventie. We geven verdere aanbevelingen die veelal specifiek zijn voor IPD-MA van gerandomiseerd experimenteel onderzoek waarin het effect van een interventie op de tijd tot een gebeurtenis wordt onderzocht. We illustreren enkele belangrijke methoden in een klinische IPD-MA. In deze IPD-MA zijn IPD van 1.225 deelnemers van vijf gerandomiseerde experimentele klinische onderzoeken samengevoegd om de effecten van Carbamazepine en Valproaat op de incidentie van epileptische aanvallen te vergelijken.
Na een overzicht en discussie van statistische methoden voor het analyseren van geclusterde en gecombineerde datasets in interventieonderzoek te hebben gegeven, behandelen we voorspellingsmethoden hiervoor in het kader van diagnostisch en prognostisch onderzoek.

IPD-meta-analyse in voorspellingsonderzoek

Voorspellingsmodellen leveren vaak onnauwkeurige voorspellingen op voor nieuwe individuen. Hoewel het gebruik van grote datasets zoals IPD-MA of elektronische patiëntendossiers de nauwkeurigheid van voorspellingen kan verbeteren, wordt er tot zover weinig of zelfs geen rekening gehouden met heterogeniteit tussen populaties, datasets en/of studies bij het ontwikkelen en testen van voorspellingsmodellen. Dit beperkt de generaliseerbaarheid van reeds ontwikkelde modellen (zelfs van grote, samengevoegde, geclusterde datasets) en zorgt ervoor dat lokale aanpassingen vaak noodzakelijk zijn. In **hoofdstuk 3** ontwikkelen we een methode voor het ontwikkelen van voorspellingsmodellen die robuuster zijn en waarvoor minder lokale aanpassingen nodig zijn. We passen Interne-Externe Kruisvalidatie toe om de heterogeniteit van het voorspellend vermogen van een model te beoordelen en te verminderen tijdens de ontwikkeling van het model. We stellen een algoritme voor dat een selectie van voorspellende variabelen maakt en vervolgens combineert in een voorspellingsmodel. Hierbij wordt getracht het voorspellend vermogen in de geïncludeerde populaties te optimaliseren en de variabiliteit daarvan tussen deze populaties te minimaliseren. Voorspellende variabelen worden iteratief toegevoegd totdat de geschatte generaliseerbaarheid is geoptimaliseerd.

We illustreren deze methode door een nieuw model te ontwikkelen dat het risico op boezemfibrilleren voorspelt. Ook updaten we een model voor de diagnose van diepe veneuze trombose (DVT). We hebben hiervoor IPD van respectievelijk twintig cohortonderzoeken (N = 10.873) en elf diagnostische onderzoeken (N = 10.014) gebruikt. Uit meta-analyse van de schattingen van de kalibratie van voorspellingen en het vermogen om te discrimineren tussen zieke en niet zieke onderzoeksdeelnemers in iedere dataset die niet voor ontwikkeling werd gebruikt blijkt dat er een wisselwerking is opgetreden tussen het gemiddelde voorspellende vermogen en de heterogeniteit daarvan. Onze methodologie neemt de heterogeniteit van het voorspellend vermogen van modellen in aanmerking tijdens de ontwikkeling daarvan in meerdere of geclusterde datasets. Onderzoekers kunnen onze methodologie gebruiken om variabelen te selecteren die de heterogeniteit van het voorspellend vermogen verminderen. Dit kan de generaliseerbaarheid naar andere omgevingen en populaties

verbeteren, wat de noodzaak voor het aanpassen van het model vermindert.

Het voorspellend vermogen van veel voorspellingsmodellen verslechtert wanneer deze worden toegepast op nieuwe individuen. Dit kan veroorzaakt worden door (een gebrek aan) de representativiteit van de steekproef die gebruik wordt in een onderzoek om een voorspellingsmodel te valideren. Als de teststeekproef (*validation sample*) de beoogde populatie niet volledig representeert, dan zullen schattingen van het voorspellend vermogen misleidend zijn. In **hoofdstuk 4** stellen wij voor om weegmethoden d.m.v. *propensity scores* (geneigdheidscores) te gebruiken om schattingen van voorspellend vermogen in andere maar gerelateerde populaties en omgevingen te standaardiseren. Dit kan door de contributie van individuen te wegen naar hun gelijkenis tot een bepaalde doelpopulatie of -omgeving. We laten zien hoe gestandaardiseerde maten voor discriminatie en kalibratie kunnen worden afgeleid.

We illustreren onze methoden in een toegepast voorbeeld van het testen van het voorspellend vermogen van acht diagnostische voorspellingsmodellen voor de detectie van diepe veneuze trombose (DVT), die van behulp kunnen zijn bij de diagnose van patiënten bij wie DVT wordt vermoed. We hebben *random effects* meta-analyse toegepast op de schattingen van het voorspellingsvermogen van de acht modellen in twaalf externe testdatasets. Dit leverde schattingen van de heterogeniteit van het voorspellingsvermogen tussen datasets. Deze waarden laten zien dat verschillen in discriminerend vermogen in de individuele testonderzoeken deels verklaard kunnen worden door verschillen in de verdelingen van patiëntkarakteristieken in de verschillende onderzoeken. Dit laat zien dat de verschillen in voorspellend vermogen niet volledig te verklaren zijn door het gebruik van verkeerd geschatte modelcoëfficiënten (d.w.z. de weging van voorspellende variabelen). Verder laat de meta-analyse zien dat voor alle methoden de heterogeniteit van de kalibratie van de coëfficiënten werd vergroot door standaardisatie. Dit geeft aan dat er verschillen waren in de patiëntkarakteristieken van de ontwikkelings- en testpopulaties, maar ook dat deze verschillen deels de verschillen in optimale coëfficiënten gemaskeerd hebben. Na het filteren van deze verschillen door middel van standaardisatie, werd het duidelijk dat er heterogeniteit aanwezig was bij de hellingscoëfficiënt van de kalibratie. Het toepassen van standaardisatie door middel van *propensity scores* kan de interpretatie van (de heterogeniteit van) het voorspellingsvermogen in (meerdere) externe testonderzoeken gemakkelijker maken en kan gebruikt worden om het updaten van voorspellingsmodellen te leiden. Dit kan soms tot de conclusie leiden dat de teststeekproef de doelpopulatie niet voldoende representeert.

Een veelvoorkomend probleem in het verzamelen en analyseren van data uit meerdere bronnen, zoals in een IPD-MA, is de aanwezigheid van meetfouten. Wanneer meetfouten optreden bij een binaire variabele spreekt men van misclassificatie. De aanwezigheid van misclassificatie kan leiden tot onjuiste onderzoeksresultaten en onbetrouwbare voorspellingsmodellen, zelfs wanneer de meetfout volledig willekeurig is. Hoewel er verschillende methodes beschikbaar zijn om om te gaan met misclassificatie, houden deze geen rekening met de veelvoorkomende heterogeniteit tussen studies van een IPD-MA. In **hoofdstuk 5** ontwikkelen we een Bayesiaans statistisch raamwerk dat rekening houdt met misclassificatie in voorspellende

variabelen in een IPD-MA waar de aard en de omvang van de meetfout tussen onderzoeken kan variëren en afhankelijk kan zijn van eigenschappen van individuele onderzoeksdeelnemers. Dit maakt het mogelijk om zuivere schattingen van gecorrigeerde (als covariaten gemeten zijn) en ongecorrigeerde associaties tussen voorspeller en uitkomst te verkrijgen, alsmede zuivere schattingen van heterogeniteit tussen onderzoeken.

We illustreren onze methodologie in een toegepast voorbeeld van de diagnose van dengue (knokkelkoorts) met gebruik van twee (mogelijk) voorspellende variabelen. In dit voorbeeld is de gouden standaard (d.w.z. een foutloze meetmethode) voor één voorspellende variabele in de helft van de onderzoeken van de IPD-MA niet toegepast. In die onderzoeken is enkel een surrogaatmeting met mogelijke fouten gedaan. Onze methoden corrigeerden voor deze mogelijke meetfouten en verminderden de fout in de schattingen van de associatie tussen de voorspeller en dengue. Over het geheel genomen leverden onze methoden schattingen met kleinere fouten dan een analyse die geen rekening hield met de mogelijke misclassificatie en ook in vergelijking met een analyse waarin enkel de gouden standaard gebruikt werd. De schattingen van de heterogeniteit van de associatie tussen de voorspeller en dengue waren vergelijkbaar voor alle geanalyseerde methoden.

Daarnaast lieten onze simulaties zien dat er binnen ons raamwerk adequaat omgegaan kan worden met misclassificatie die afhankelijk is van eigenschappen van onderzoeken en individuen. Het geïmplementeerde misclassificatiemodel dat effecten en misclassificatie van de voorspellende variabele modelleert op het niveau van individuen en onderzoeken, leverde zuivere schattingen van de voorspeller-uitkomstassociatie op. De gemiddelde gekwadrateerde fout was kleiner, het onderscheidend vermogen (*power*) groter en de dekkingswaarschijnlijkheid (*coverage probability*) was vergelijkbaar met een analyse die beperkt was tot observaties waarvan de gouden standaard gemeten was. De schattingen van de heterogeniteit waren adequaat voor alle in de simulatie bestudeerde modellen. Het door ons voorgestelde raamwerk kan worden gebruikt om om te gaan met de mogelijke aanwezigheid van misclassificatie van een voorspellende variabele in een IPD-MA. Ons raamwerk vereist dat ten minste enkele onderzoeken IPD leveren voor metingen van de voorspeller d.m.v. beide de surrogaat en de gouden standaard en dat misclassificatie uitwisselbaar is tussen onderzoeken, gegeven de geobserveerde covariaten (en mogelijk de uitkomst).

Het gebruik van Multinomiale Logistische Regressie (MLR) wordt aangeraden bij de ontwikkeling van klinische voorspellingsmodellen waarbij onderscheid gemaakt moet worden tussen drie of meer ongeordende uitkomsten. In **hoofdstuk 6** presenteren we een simulatieonderzoek, waarin we het voorspellend vermogen van MLR-modellen in relatie tot de relatieve grootte van de uitkomstcategorieën, het aantal voorspellende variabelen en het aantal gebeurtenissen per variabele hebben onderzocht. Hieruit blijkt dat het schatten van een MLR-model met de methode *maximum likelihood* modellen oplevert die *overfitted* zijn, wanneer de dataset klein tot middelgroot is. In de meeste gevallen worden de kalibratie en het voorspellend vermogen over het algemeen verbeterd door het gebruik van de methode *penalized maximum likelihood*. Ons simulatieonderzoek laat zien dat in de multinomiale con-

text naast het aantal gebeurtenissen per variabele ook de totale steekproefgrootte van belang is voor het voorspellend vermogen. Zoals verwacht toont ons onderzoek de noodzaak van correctie voor *optimism* van het geschatte voorspellend vermogen wanneer een MLR-model ontwikkeld wordt. We raden het gebruik van de methode *penalized maximum likelihood* aan wanneer een model ontwikkeld wordt in een kleine dataset of in een middelgrote data set met een kleine totale steekproefgrootte (bijvoorbeeld wanneer de uitkomstcategorieën even vaak voorkomen). We presenteren ook een toegepast voorbeeld, waar we de ontwikkeling en validatie van *penalized* en *unpenalized* MLR-modellen voor het voorspellen van kwaadaardige eierstokkanker illustreren.

Discussie

Ten slotte geven we in **hoofdstuk 7** een overzicht van methoden voor meta-analyse in prognostisch onderzoek. De laatste jaren is het reviewen van onderzoeksresultaten essentieel geworden voor de generaliseerbaarheid van medisch wetenschappelijk onderzoek. Vaak worden de kwantitieve bevindingen uit gereviewde artikelen rekenkundig samengevat door middel van een meta-analyse. Een voorbeeld hiervan is het kwantitatief samenvatten van de effecten van twee medische behandelingen. Het gebruik van meta-analyse is echter minder voor de hand liggend in prognostisch onderzoek vanwege substantiële variatie in onderzoeksdoelstellingen, in analysemethoden en in het niveau van gerapporteerde onderzoeksresultaten in dat onderzoeksgebied. We geven een overzicht van statistische methoden die gebruikt kunnen worden om data van onderzoeken van prognostische factoren en prognostische modellen samen te vatten. We bespreken hoe gerapporteerde statistische gegevens, IPD of een combinatie daarvan samengevat kunnen worden met methoden voor meta-analyse. Aan de hand van recente voorbeelden illustreren we het gebruik van verschillende methoden. We sluiten af met enkele aanbevelingen voor het uitvoeren van een IPD-MA in onderzoek naar voorspellingsmodellen in het algemeen.

List of publications and conference presentations

Publications

- de Jong VMT, Eijkemans MJC, van Calster B, Timmerman D, Moons KGM, Steyerberg EW, et al. Sample size considerations and predictive performance of multinomial logistic prediction models. *Statistics in Medicine*. 2019;38(9): 1601–19.
- de Jong VMT*, Debray TPA*, Moons KGM, Riley RD. Evidence synthesis in prognosis research. *Diagnostic and Prognostic Research*. 2019 Jul 11;3(1):13. * *Contributed equally*
- de Jong VMT, Moons KGM, Riley RD, Smith CT, Marson AG, Eijkemans MJC, et al. Individual participant data meta-analysis of intervention studies with time-to-event outcomes: A review of the methodology and an applied example. *Research Synthesis Methods*. 2019.

Oral presentations

- Predictive performance of multinomial logistic prediction models - a simulation study, *International Society for Clinical Biostatistics (ISCB)*, 21-25 August 2016, Birmingham, United Kingdom
- Individual patient data meta-analysis of time-to-event data: A review of the methodology, *International Society for Clinical Biostatistics (ISCB)*, 9-13 July 2017, Vigo, Spain
- Developing and updating prediction models in large clustered data sets, *Methods for Evaluation of medical prediction Models, Tests And Biomarkers (MEMTAB)*, 2-3 July 2018, Utrecht, the Netherlands
- Developing and updating prediction models in large clustered data sets, *International Society for Clinical Biostatistics and Australian Statisical Conference (ISCB ASC)*, 26-30 August 2018, Melbourne, Australia
- Evidence synthesis in prognosis research, *International Society for Clinical Biostatistics (ISCB)*, 14-18 July 2019, Leuven, Belgium

Poster presentations

- Predictive performance of multinomial logistic prediction models, *Methods for Evaluation of medical prediction Models, Tests And Biomarkers (MEMTAB)*, 2-3 July 2018, Utrecht, the Netherlands
- Individual-participant-data meta-analysis of intervention studies with time-to-event outcomes: A review of methodology and an example, *International Society for Clinical Biostatistics (ISCB)*, 14-18 July 2019, Leuven, Belgium

Lists of software and supporting information

Software

I have implemented various newly developed methods, evaluations and applications that are presented in this thesis in software that is publicly available online in software repositories. Note that although these repositories retain the software as used for or presented in this thesis, they may contain more recent updates and/or additions. Here I give an overview thereof:

Chapter	What	Where
2	Epilepsy motivating example	GitHub
3	R package: SIECV methods	CRAN
		R-Forge
	DVT motivating example	GitHub
	AF motivating example	GitHub
4	R package: methods for validation	GitHub
6	Simulation and data	GitHub
	Ovarian cancer case study	GitHub
	R package: Performance measures	GitHub

Github: `https://github.com/VMTdeJong/`
CRAN & R-Forge: implemented in R package `metamisc`.

Supporting information

To save some trees, I omitted the five supporting information documents for chapters 2 and 6 from this thesis. Since these are part of Open Access publications, anyone can view them through the links below. To avoid dead links, I link to the DOI of the manuscripts, as the supporting information documents do not have their own DOI.

Chapter	Where
2	10.1002/jrsm.1384
5	page 109
6	10.1002/sim.8063

Dankwoord

Prof. dr. Moons, beste Carl, Je optimisme is erg motiverend en inspirerend. Je weet er altijd een positieve draai aan te geven als een hoofdstuk nog wat extra werk nodig heeft. Jouw inzicht heeft me geholpen om de vertaalslag naar de epidemiologie te maken. Vanaf de eerste dag van dit promotietraject maakte je duidelijk dat mijn leertraject voorop stond. We begonnen dus geen projecten waar ik alle kennis al voor had; het doel was steeds om nieuwe kennis en vaardigheden op te doen. Dit heeft dan ook geleid tot een zeer gevarieerd proefschrift.

Prof. dr. ir. Eijkemans, beste René, ik heb veel plezier gehad van onze discussies. Op elk wiskundig probleem weet je een oplossing. Je vertelde me niet wat ik moest doen, maar stelde de juiste vragen, waardoor ik enorme vrijheid heb gehad om dit proefschrift naar eigen inzicht in te vullen.

Dr. Debray, beste Thomas, ik heb enorm genoten van onze discussies. Ik heb heel veel van je geleerd over o.a. meta-analyse en missing data, en af en toe hadden we nog de nodige discussie over machine learning. Je hebt je ingezet om nieuwe mogelijkheden voor mij te creëren. Dat heeft er toe geleid dat we verder samen kunnen werken onder de vlag van ReCoDID. Daarvoor en voor al je hulp tijdens mijn promotie ben ik je eeuwig dankbaar. Het was en is een waar genoegen om met je samen te werken.

Beste leden van de beoordelingscommissie en de promotiecommissie, prof. dr. Scholten, prof. dr. Houwing-Duistermaat, prof. dr. Hoijtink, prof. dr. Bots, prof. dr. Nielen, dr. van Calster, dr. Oberski, ik dank u voor de bereidheid mijn proefschrift te lezen en beoordelen.

To all the coauthors who have contributed to this thesis, I would like say thank you: Richard, Catrin, Jeroen, Long, Paul, Harlan, Thomas, Ben & Ewout. Your worthy contributions are highly appreciated. Richard, your thoughtful comments have significantly improved not one, not two, but three chapters of this thesis.

Beste Maarten, jij hebt me geïnspireerd om onderzoek naar voorspellingsmodellen te gaan doen, en daar ben ik je nog steeds dankbaar voor. Én dat heeft tot een hoofdstuk van dit proefschrift geleid! Ik kijk er naar uit om weer met je samen te werken wanneer je weer naar Utrecht komt.

Colleagues from ReCoDID, Paul, Harlan, Lauren, Thomas, Kerstin, Frank, Heather, Till, I have enjoyed working with you, even if it has (mostly) been from far away. Paul and Harlan, I cannot stress enough how much your knowledge of misclassification, measurement error and Bayesian statistics have contributed to the respective chapter.

Beste kamergenoten, dr. Jenniskens, Kevin, Saskia, Carline, Nicole, Giske, Chris, Pauline, Anne-Karien, Romin, Anna-Maria, Suzanne, Anouk, Marjolein, Lenja, bedankt voor de leuke tijd samen. Ik heb genoten van alle wandelingen

en lunches. Kevin, en Saskia, eerst in het Stratenum, daarna in het van Geuns, met jullie was het altijd gezellig werken. Buiten het werk hebben we veel lol gehad bij o.a. het klimmen, tafelvoetbal en de estafette. Kevin, het was een eer om je paranimf te zijn, en bedankt dat je mijn paranimf wilde zijn! Ik kijk er naar uit om weer samen met je te kunnen werken "aan de overkant". Anouk, Marjolein, Lenja, toen ik vertelde dat ik mijn kamer moest verlaten hebben jullie mij direct geadopteerd en vervolgens mentaal gesteund bij het verrichten van de laatste loodjes van dit proefschrift.

Marian, ik weet dat je stiekem al heel goed bent in het Nederlands, dus schrijf ik ook voor jou dit dankwoord in het Nederlands. Met jouw humor en altijd aanwezige glimlach wist je me altijd weer op te vrolijken wanneer ik een PhD-dipje had. Hoewel onze proefschriften over totaal andere onderwerpen gaan, heb je me ook inhoudelijk kunnen helpen en op nieuwe ideeën kunnen brengen. Ten slotte bedankt dat je mijn paranimf wilde zijn!

Ik wil graag de hele epi-methoden- en biostatistiekteams en de vele anderen die zich aansloten bij de methoden- en predictievergaderingen bedanken voor de waardevolle discussies tijdens die vergaderingen. Ik heb veel van jullie presentaties geleerd, en het presenteren van mijn eigen onderzoek heeft enorm geholpen om enkele stukken van dit proefschrift te verduidelijken. Cas, Rebecca, Caroline en Paul, ik vond het leuk om met jullie les te geven en ik heb er veel van geleerd over statistiek en het overbrengen van kennis.

Medepromovendi, ik heb genoten van de vele borrels en uitjes en natuurlijk de promovenski's. Het was fijn om met jullie in hetzelfde schuitje te zitten. Josan en Anouk, het was een plezier om met jullie de JOB voor te zitten.

39-ers, we hebben veel lol gehad samen: vele feestjes thuis of in de doos, samen eten, koppen koffie tijdens het samen studeren (of juist niet). Joost, het samen programmeren in R & Python, de discussies over statistische methoden en juist de theorie daarachter hebben me extra gemotiveerd om onderzoek hiernaar te gaan doen.

IBBejaarden en medeadoptiebejaarden, ik heb veel lol gehad tijdens de spelletjesavonden, feestjes en natuurlijk de frankantie. Bedankt dat jullie me in de groep geadopteerd hebben.

Sebastiaan en Rens, de spelletjesavonden met uiteenlopende discussies over economie, politiek, statistiek, programmeren en vele andere zaken zijn een goede afleiding van dit proefschrift geweest. Sebastiaan, ik vind het leuk dat hoewel we totaal verschillende werk- en studiepaden hebben gevolgd, we toch beide in de datawereld terecht zijn gekomen en hier nu diepgaande discussies over kunnen hebben.

Lieve familie de Jong, van Woudenbergh en Verstegen, het is altijd fijn jullie weer te zien. we hebben zware tijden gehad en jullie steun heeft daarbij veel geholpen.

De Saas, lieve Saskia, ik ben tijdens dit promotietraject vaak in de avond of in het weekend druk geweest, vooral tijdens de laatste maanden. Dank je wel voor je liefde, steun en begrip. Ik kan niet in woorden uitdrukken hoeveel dit geholpen heeft.

Ineke, lieve mama, jij hebt me al die tijd gesteund. Ook al heb je me meermaals moeten vragen waar ik nou toch mee bezig was, had je er altijd vertrouwen in dat het goed ging komen met dit proefschrift.

En natuurlijk Marcel, lieve papa. Ik heb je natuurlijk voor het einde bewaard, omdat ik er nog steeds grote moeite mee heb dit op papier te zetten. Je hebt sinds ik klein was mijn nieuwsgierigheid aangemoedigd. Hoewel mijn eerste studiejaren niet zo vlot gingen, had je er het volste vertrouwen in dat het goed zou komen met mijn studie. Ik vind het heel jammer dat je mijn promotietraject en proefschrift niet hebt mogen meemaken.

Curriculum vitae

Valentijn M.T. de Jong was born on the 19th of June, 1990, in Montfoort, the Netherlands. He obtained his bachelor's degree in Natural and Social Sciences with a specialization in Brain and Cognitive Sciences from the University of Amsterdam in 2014. He then started the MSc programme in Methodology and Statistics for the Behavioural, Biomedical and Social Sciences at Utrecht University (UU). In 2015 he taught (bio-)statistics at the UU and the University Medical Center Utrecht (UMCU), as a student assistant. In 2016 he obtained his master's degree (cum laude), after completing an internship at the Department of Biostatistics and Research Support of the Julius Center for Health Sciences and Primary Care at the UMCU.

In 2016 he started working on the present PhD thesis at the UMCU, under supervision of dr. T.P.A. (Thomas) Debray, Prof. dr. K.G.M. (Carl) Moons and Prof. dr. ir. M.J.C. (René) Eijkemans. During his PhD he taught various (bio-)statistics and methodology courses at the bachelor and master level. He chaired the Jonge Onderzoekers Bijeenkomst (JOB, Junior Researchers Meeting) from 2018 till 2019. He also started the post-graduate master programme in Epidemiology and is due to graduate in 2020.

Valentijn currently works as a post-doctoral researcher at the Julius Center for Health Sciences and Primary Care, UMCU, where he works on methodological aspects relating to individual participant data meta-analysis.