EVIDENCE SYNTHESIS FOR PROGNOSIS AND PREDICTION: APPLICATION, METHODOLOGY AND USE OF INDIVIDUAL PARTICIPANT DATA

By

JOIE ENSOR

A thesis submitted to

The University of Birmingham

for the degree of

DOCTOR OF PHILOSOPHY

Institute of Applied Health Research

The University of Birmingham

March 2017

UNIVERSITY OF
BIRMINGHAM

University of Birmingham Research Archive

e-theses repository

This unpublished thesis/dissertation is copyright of the author and/or third parties. The intellectual property rights of the author or third parties in respect of this work are as defined by The Copyright Designs and Patents Act 1988 or as modified by any successor legislation.

Any use made of information contained in this thesis/dissertation must be in accordance with that legislation and must be properly acknowledged. Further distribution or reproduction in any format is prohibited without the permission of the copyright holder.

ABSTRACT

Prognosis research summarises, explains and predicts future outcomes in patients with a particular condition. This thesis investigates the application and development of evidence synthesis methods for prognosis research, with particular attention given to improving individualised predictions from prognostic models developed and/or validated using meta-analysis techniques.

A review of existing prognostic models for recurrence of venous thromboembolism highlighted several methodological and reporting issues. This motivated the development of a new model to address previous shortcomings, in particular by explicitly modelling and reporting the baseline hazard to enable individualised risk predictions over time. The new model was developed using individual participant data from several studies, using a novel internal-external cross-validation approach. This highlighted the potential for between-study heterogeneity in model performance, and motivated the investigation of recalibration methods to substantially improve consistency in model performance across populations.

Finally, a new multiple imputation method was developed to investigate the impact of missing threshold information in meta-analysis of prognostic test accuracy. Computer code was developed to implement the method, and applied examples indicated missing thresholds could have a potentially large impact on conclusions. A simulation study indicated that the new method generally improves on the current standard, in terms of bias, precision and coverage.

ACKNOWLEDGEMENTS

Firstly, I would like to say a heartfelt thank you to my main supervisor, Richard Riley. I won the lottery with you as my supervisor. You have always supported me, pushed me, and looked out for me, and I thank you so much for that. You are a great teacher and I have learned so much from you, and I look forward to learning much more and working together in the future. I cannot thank you enough for everything you have done, and for always having my back when the sandwiches have dodgy fillings.

I would also like to thank David Moore; thank you for your invaluable support without which I would never have been able to complete a HTA report, and thank you for teaching me so much about systematic reviewing, and the many other things you know so much about. Many thanks also go to the support of Jon Deeks; thank you for your expert knowledge and guidance through this work, you have an amazing ability to hone in on the crux of problems, which is an invaluable skill. Thank you for looking out for me when my time at Birmingham came to an end. And thank you for breaking my computer programs.

Thank you to my friends for supporting me in my quest to become a “Dr of Numbers”, as you say. Most importantly thank you to Kym and Dani, I am lucky to get to work with such a great team of like-minded people, thank you for all your help and support in this and other work. Also thank you for checking on me after conference dinners.

To my mum, dad, sister and grandparents, thank you for always supporting me even when it makes no sense to you. Mum and Dad, thank you for always believing in me and pushing me to do the best I can, thank you for always being there for me, listening and giving me your sound advice. I wouldn’t have made it here without you. Pen, thank you for always making me laugh, and grandma thank you for always keeping the cupboard stocked with Jaffa cakes.

Finally, I would like to thank Emma. Emma, without you I would never have believed that I could do a PhD, or that I would actually enjoy doing one. Thank you so much, I really wouldn’t be here without you. You are my rock, you have always been my biggest fan, and I couldn’t have completed this without your selfless support. Thank you for helping me get through this, for always humouring me even when I'm being stupid, for always being there to talk to when the simulations broken, and for your patience and understanding. You are amazing, and you inspire me every day.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AIC	Akaike Information Criteria
AUC	Area Under the (ROC) Curve
BIC	Bayes Information Criteria
CI	Confidence Interval
EPP	Events Per Predictor
FP	Flexible Parameteric (Model)
IECV	Internal-External Cross-Validation
IPD	Individual Participant Data
LP	Linear Predictor
MAR	Missing At Random
MCAR	Missing Completely At Random
MFP	Multivariable Fractional Polynomial
MI	Multiple Imputation
MICE	Multiple Imputation by Chained Equations
MIDC	Multiple Imputation using Discrete Combinations
ML	Maximum Likelihood
MNAR	Missing Not At Random
NI	No Imputation
OAC	Oral Anticoagulants
REML	Restricted Maximum Likelihood
RFS	Recurrence-Free Survival
ROC	Receiver Operating Characteristic (Curve)
RVTE	Recurrent Venous Thromboembolism (Database)
SI	Single Imputation
VTE	Venous Thromboembolism

CHAPTER 1: INTRODUCTION

1.1 Overview of the thesis

In medicine, prognosis is defined as the prediction of future outcomes in patients with a certain baseline health condition (1). Prognosis is of ever increasing importance to clinicians, researchers and funders as more and more patients are living with primary conditions and many of these with additional comorbidities (2, 3). Prognosis is crucial to inform clinicians of the risk of a patient's future health outcomes, which allows decisions to be made on appropriate treatment strategies. Similarly estimating a patient's prognosis provides useful information for the patient on the likely future course of their illness (1). Knowledge of prognosis enables selection or stratification of patients for clinical trials, typically using prognostic models, which tailor predictions to individuals based on their own set of prognostic factors (4-6). Prognostic models and single factors are also important for adjustment for, and understanding of case-mix variation in patient outcomes across centres or studies (7).

Prognosis research is therefore a crucial part of medical research, and the PROGRESS initiative suggests a framework of four key areas; overall prognosis (3), prognostic factor research (8), multivariable prognostic modelling (9), and stratified medicine research (10). The overall or average prognosis of a group of people describes the course of future outcomes in the context of current strategies for diagnosis and treatment of a given health condition (3). For example, the overall risk of recurrent venous thromboembolism (VTE) in patients with an unprovoked initial VTE is much greater than those with a provoked VTE over time (see Figure 1.1); there

may be differences in prognosis for individual patients, but on average unprovoked patients are at greater risk of recurrence. This motivates clinical research to address this risk, in developing better diagnostic strategies, different treatment regimens specific to unprovoked patients, or identifying those at highest risk for further care.

Prognostic modelling and prognostic factor research are popular topics because of their many uses, and these areas form the main interest of this thesis (9, 11). Prognostic factors may be useful as predictors of treatment response, particularly where the factor is potentially modifiable, or may be used to decide on further testing or procedures (8). Prognostic factors also form the basis of prognostic models, to allow prediction of an individual's risk of outcome given their values of the included factors (predictors) (9).

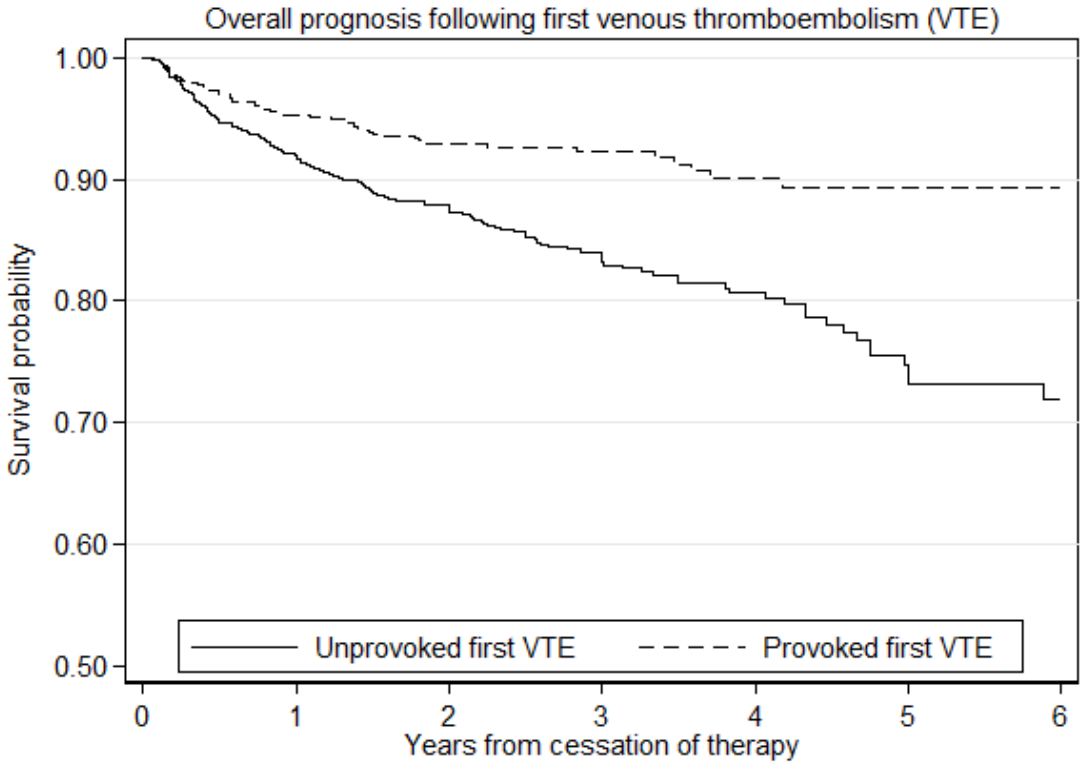


Figure 1.1 - Overall prognosis for recurrent VTE following initial provoked and unprovoked VTE

The increasing popularity of prognostic factor and model research in medicine has been facilitated by the growing trend of data sharing, and availability of large datasets such as patient electronic health records (EHR) (12-14). Large datasets of individual participant data (IPD) are becoming more easily accessible providing a wealth of information often from multiple studies, centres, clinical settings and geographical locations. Availability of IPD from multiple studies provides a unique opportunity for both prognostic factor and model research. The effect of prognostic factors can be assessed and combined ('meta-analysis') across multiple studies, increasing sample size to detect true effects, and allowing for patient case-mix variation to give better estimates of prognostic effect (15). Prognostic models can also be developed allowing for such variation by incorporating factors which explain the clustering of patients in studies or centres, and allowing for different implementation strategies in new populations (16-19).

Given this clustering of patients within studies, centres and countries, traditional evidence synthesis methods have a natural application for both prognostic factor and model research, which is a major theme of this thesis. IPD meta-analysis techniques have further potential uses for model development and validation (12, 20, 21), while others have led the way in the use of evidence synthesis methods for prognosis research using multiple clusters (12, 20, 22-24).

This research thesis focuses on the application and development of statistical methods for prognostic model and factor research. Specifically, this thesis uses evidence synthesis techniques in the development, validation, implementation and assessment of prognostic models and factors, with the aim of improving their performance, transportability and external validity.

In this introduction chapter, a framework for prognosis research is described and the key statistical concepts outlined, in regard to both single studies and meta-analysis of multiple studies. The rationale for the thesis is then given, and then the aims and outline of subsequent chapters is provided.

1.2 Prognosis research structure

Despite the importance of prognosis in medicine, historically prognosis research in the literature is of low quality and low impact (25). Over the last decade there has been a concerted effort to improve the quality, reporting and impact of prognosis research (1, 3, 25). The PROGnosis RESearch Strategy (PROGRESS) partnership, published a series of articles proposing a framework of four prognosis research themes (3, 8-10), one on each of the following areas;

(1) Fundamental (overall) prognosis research – The study of the natural course of health conditions in the context of current care (3).

(2) Prognostic factor research – The study of single factors and their association with patient outcomes (8).

(3) Prognostic model research - The development, validation, and impact of statistical models which combine several factors in order to predict individuals risk of outcome (9).

(4) Stratified medicine research - The use of prognostic information to help stratify treatment decisions for individuals (10).

These articles provide guidance on the undertaking of prognosis research in each of the above themes, and highlight particular challenges in the field and ways to improve the current standards of prognosis research. Further to these articles, special attention has been paid to the area of prognostic model research (1, 26-28). Prognosis is often a multifactorial problem and as such prognostic models naturally replicate the process many clinicians use to make predictions for their patients. Often single prognostic factors do not give accurate individual predictions for longer term outcomes, meaning prognostic models are required (8). Prognostic model research is also complicated, and there is no agreed strategy for development of a model, increasing the likelihood of poorly developed models (27).

This thesis focuses primarily on the examination of single prognostic factors and the development/validation of prognostic models; the aim is to utilise and develop statistical methods for improving the development, validation and implementation of factors and models when data from multiple studies are available. The following sections discuss prognostic factor and model research in more detail.

1.3 Prognostic factor research

A prognostic factor is any measure that is associated with a future outcome in a group of patients with a given health condition (8). These factors may be biological, environmental or psychological, and can range from simple (e.g. sex, age or weight), to complex factors measured in individuals (e.g. biomarkers, physiological or imaging variables). Prognostic factors are referred to by many names in the medical literature including; predictors, risk

factors, prognostic variables, and prognostic markers. In this thesis the terms 'prognostic factor' and 'predictor' will be most often used.

Prognostic factors have many potential uses in improving health outcomes including (8);

- Classifying disease at diagnosis (e.g. presence of elevated blood pressure combined with significant proteinuria used to define pre-eclampsia (29, 30)),
- Informing treatment strategies,
- Forming the building blocks for prognostic models (e.g. use of patients D-dimer levels to predict recurrence risk in VTE patients (15)),
- As potential predictors of treatment response,
- Monitoring disease progression (e.g. the RECIST criteria used to define response to treatment in cancer patients, which uses patient scans and measurements of tumour size (31)),
- As potential treatment effect modifiers (e.g. treatment response to tamoxifen is dependent on patients oestrogen receptor (ER) status (32) as shown in Figure 1.2)
- And, as potential confounders to be adjusted for in analyses

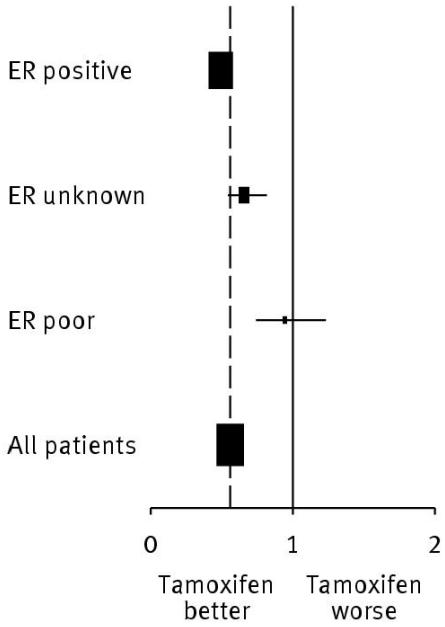


Figure 1.2 - Illustration of estrogen receptor (ER) status as a treatment effect modifier for tamoxifen in breast cancer (3).

1.4 Prognostic model research

Multivariable prediction models combine multiple individual predictors to either predict the risk of developing a future outcome (prognostic model), or the risk of an outcome being present or absent (diagnostic model) (33, 34). In this thesis the term 'prediction model' is used interchangeably with 'prognostic model', as the focus is always on future outcomes, but key differences are highlighted where necessary. As with prognostic factors, prediction models have many names in the literature including; prediction/risk score, prediction/risk tool, prediction index, prediction rule, or classification rule among others. Prediction models combine two or more individual predictors within a multivariable regression model, commonly a logistic or cox regression model, which are discussed in detail in the next section, following which further details on the phases of prediction model research are discussed. Examples of prognostic models include the QFracture algorithm to predict individuals 10-year probability of osteoporotic or hip fracture (35, 36), and the DASH score for predicting risk of recurrent VTE in unprovoked VTE patients after cessation of therapy (37). The DASH score combines scores for predictors including abnormal D-dimer levels (+2 score), age ≤ 50 years (+1 score), male sex (+1 score) and hormone use (-2 score), to calculate patient's cumulative recurrence risk at one, two and five years from cessation of therapy, with estimated 95% confidence intervals (see Table 1.1).

Table 1.1 – Predicted cumulative recurrence risks for groups of patients as defined by the DASH score (37).

DASH score	Cumulative recurrence (%, 95% Confidence interval)		
	1 Year	2 Years	5 Years
-2	2.4 (0.3–15.8)	5.2 (1.3–19.2)	5.2 (1.3–19.2)
-1	1.9 (0.3–5.9)	1.9 (0.6–5.9)	5.7 (1.5–20.5)
0	4.2 (2.3–7.7)	5.4 (3.1–9.3)	9.5 (3.8–22.3)
1	5.1 (3.4–7.5)	8.7 (6.3–12.0)	15.9 (10.1–24.3)
2	8.4 (6.2–11.5)	12.8 (9.9–16.4)	25.3 (17.6–35.7)
3	14.6 (11.3–18.8)	20.5 (16.4–25.5)	40.9 (31.2–52.4)
4	21.9 (13.6–34.1)	33.6 (23.3–46.8)	61.3 (44.3–78.5)

Prediction models are a natural step towards the stratified medicine approach which is the final theme described by the PROGRESS partnership (10). They allow clinicians and patients to jointly make informed decisions on treatment strategies based on the patients' prognosis (9). Importantly such models can be used to identify subgroups of patients with different levels of risk, in whom relative treatment effects may have greater absolute benefit or harm (4). Prediction models can also be used to increase power and reduce sample size requirements in clinical trials by prognostic targeting or predictor adjustment (5, 6), and in understanding differences in prognosis across centres or countries (7).

1.5 Statistical methods for prognosis and prediction models

The following introduces the key statistical methods for prognosis research, specifically those used in assessing the prognostic effect of single factors, and in developing clinical prediction models. In prognosis research the most common prediction models aim to predict either binary outcomes or time-to-event outcomes, for which typically either logistic or survival models are used, respectively.

1.5.1 Logistic regression

Logistic regression models are used in prognosis to model the relationship between one or more predictors and a binary outcome (e.g. mortality), usually a short-term outcome for which all patients have complete follow-up (34). They are also commonly used as diagnostic models for example modelling disease presence or absence. As a generalised linear model the logistic regression model (see Equation 1.1) uses a logit link to associate a binary outcome, Y with a combination of predictors $\boldsymbol{X} = (x_1, x_2, \dots)^T$, and regression coefficients $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots)^T$. The outcome relates to the probability of having the event, $p = P(Y = 1)$.

$$Logit(p) = \log\left(\frac{p}{1-p}\right) = \alpha + \boldsymbol{\beta X} = LP$$

Equation 1.1

The logit link function limits the back-transformed probability (p) from the model to lie between zero and one. The linear predictor (LP) is defined as the combination of the intercept term, α and the set of predictors with their associated coefficients $\boldsymbol{\beta X}$. Where interest lies in the effect of a single predictor or prognostic factor the vector $\boldsymbol{X}$ may contain only one predictor, x_1, or alternatively a set of other predictors as adjustment factors to ascertain the prognostic ability of the predictor beyond other potential confounders. Each coefficient within $\boldsymbol{\beta}$ relates to a log odds ratio, giving the change in log odds for a 1-unit increase in the associated x (conditional on any other factors in the model), typically estimated using maximum likelihood. This is sometimes known as the prognostic or predictive effect of a factor. The model can be easily extended to incorporate random effects, to allow for differences in predictor effects across clusters within the data for example.

Using the logistic model for prognosis

For prognosis, predicted probabilities for any individual, *i*, can be obtained by back-transforming the linear predictor from Equation 1.1 using the inverse function, also called the logistic function (38), as below;

$$p_i = logit^{-1}(LP_i) = logistic(LP_i) = \frac{\exp(LP_i)}{1 + \exp(LP_i)} = \frac{1}{1 + \exp(-LP_i)}$$

Equation 1.2

The top panel of Figure 1.3 shows the predicted values from Equation 1.1, which can be seen to lie outside of the range [0, 1] on the logit scale. In contrast the bottom panel in Figure 1.3 shows the probability of an event as calculated by transforming Equation 1.1 using the logistic function given in Equation 1.2. The typical sigmoid shape of the logistic function is evident and ensures predictions lie between zero and one. The HERDOO2 model is an example of a published prognostic model developed in a logistic framework, for predicting risk of recurrent VTE (see Figure 1.4) (39).

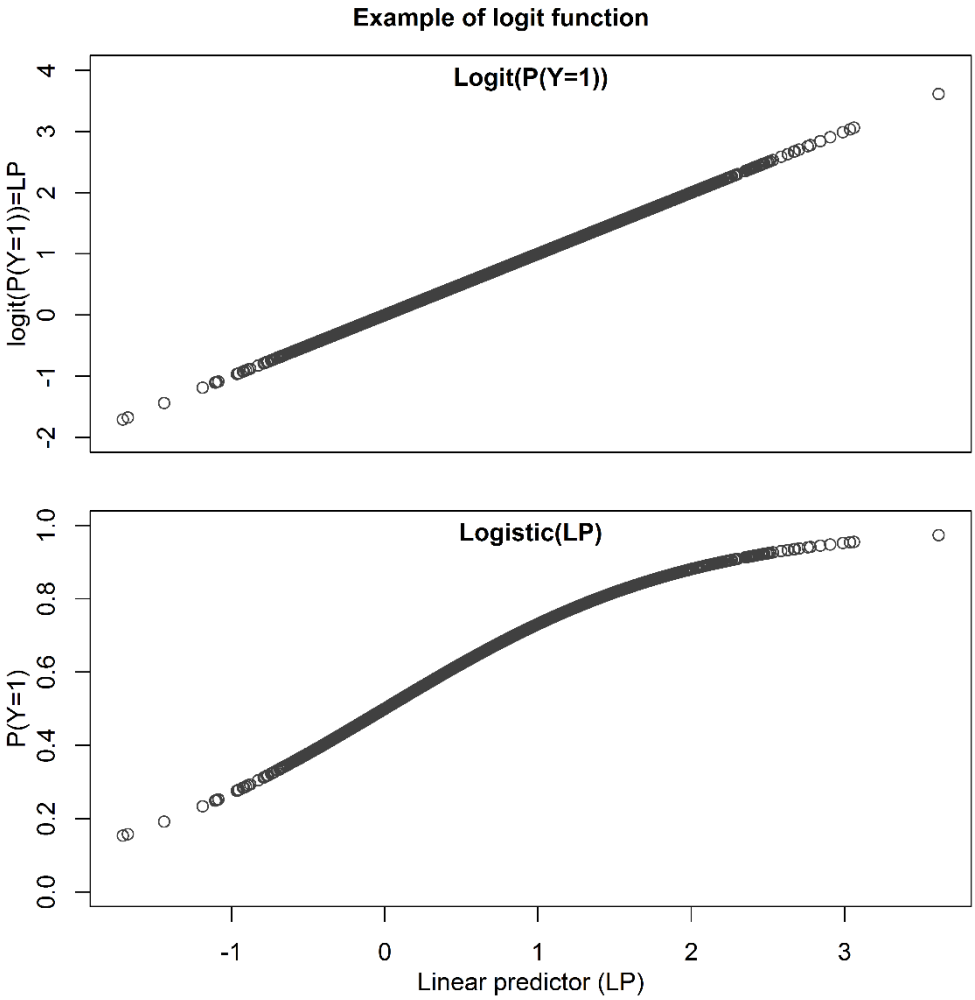


Figure 1.3 - Example of the probability of event from the logistic model (logit(p) = LP (as in Equation 1.1)) against the linear predictor (top panel) and the probability of an event transformed using the logistic function (p = 1/1+exp(-LP)) against the linear predictor (bottom panel).

Examples of published prediction models

The following give real examples of logistic and Cox regression models developed for the prediction of risk of recurrent VTE post cessation of therapy, in patients with an initial unprovoked VTE.

Logistic prediction model example

The HERDOO2 model is a logistic regression model for which the LP was defined as follows, where the regression coefficients represent log odds ratios as in Equation 1.1;

LP = (–3.9717 x intercept) + (1.2977 x BMI ≥ 30 kg/m2) + (0.6473 x post-thrombotic signs) + (0.9155 x D-dimer ≥ 250 μg/L) + (0.8084 x age ≥ 65 years)

Cox prediction model example

The Vienna model is presented as a nomogram which is based on a Cox regression model, meaning that the regression coefficients in the LP below represent log hazard ratios, as in Equation 1.3.

LP = (0.64 x Male) + (0.96 x PE) + (0.73 x Proximal DVT) + (0.24 x D-dimer (per doubling))

Figure 1.4 - Examples of published prediction models using logistic and Cox regression model structure (39, 40).

1.5.2 Cox regression

The Cox proportional hazards model is the most commonly used model in the literature for time-to-event outcomes (such as the Vienna prediction model (40), see Figure 1.4), allowing for patients with different lengths of follow-up and censoring (41). Right censoring is common in prognostic studies where patients may be lost to follow-up, or withdraw from the study, or

where the study ends before the patient has an event. In such cases the patient is censored at the last time they were known to have not had the event. It models the hazard function over time, $h(t; \boldsymbol{X})$ with $\boldsymbol{X} = (x_1, x_2, \dots)^T$, representing the vector of included predictors (42). As for the logistic model, a single predictor can be investigated in isolation by including only one predictor in the vector $\boldsymbol{X}$.

The baseline hazard, $h_0(t)$ represents the hazard rate at time *t*, when all predictors equal zero. The Cox model makes no assumptions about the form of the baseline hazard, allowing various complex hazard shapes to be captured by the model. It does however, in its simplest form, make the assumption that hazard rates are proportional between patient subgroups over time; the proportional hazards assumption. The vector of coefficients $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots)^T$ are estimated on the log scale (see Equation 1.3) by maximising the partial likelihood, which is independent of $h_0(t)$.

$$h(t) = h_0(t)\exp(\boldsymbol{\beta X})$$
$$\ln(h(t)) = \ln(h_0(t)) + \boldsymbol{\beta X}$$

Equation 1.3

Each β coefficient in the model relates to a log hazard ratio, giving the change in log hazard for a 1-unit increase in the associated x (conditional on any other factors in the model). The cumulative hazard function, $H(t)$ represents the total hazard accumulated up to time t, and can be directly calculated by summing the hazard function given in Equation 1.3 as follows;

$$H(t) = \int_0^t h(u)\, du$$

Equation 1.4

Using the Cox model for prognosis

As previously discussed, for prognosis interest usually lies in prediction of absolute risks for individuals, which can be calculated by first transforming the cumulative hazard function to obtain the baseline survival function $S_0(t)$;

$$S_0(t) = \exp(-H_0(t))$$

Equation 1.5

And then using the baseline survival function to obtain a predicted survival probability, $S_i(t)$ at time *t* for individual *i* by transforming Equation 1.3 as follows;

$$S_i(t) = S_0(t)^{\exp(\boldsymbol{\beta X}_i)}$$

Equation 1.6

1.5.3 Flexible parametric models

Flexible parametric (FP) models go beyond the Cox model and extend standard parametric survival models, such as the Weibull or exponential model, by modelling the baseline hazard more accurately (43-45). Standard parametric models assume distributional shapes for the baseline hazard, but are restricted and often unable to capture realistic hazard functions which may rise and fall over time. For example the Weibull model assumes a monotonic shape, either rising over time or falling over time. Parameterisation of the baseline hazard is important for prognosis; firstly in order to obtain individualised absolute risk predictions over time and secondly, for out-of-sample prediction enabling external validation. The following sections briefly describe the framework for flexible parametric models which will be used in later chapters of the thesis, beginning with restricted cubic splines.

Restricted cubic splines

FP models utilise restricted cubic splines to flexibly model the baseline hazard on the log-cumulative hazard scale. By fitting cubic splines between cut points over time, known as knots, FP models can better capture fluctuations in the baseline hazard. Restricted cubic splines are used over cubic splines to force the function to be linear before the first knot and after the last knot, known as the boundary knots (defined as the minimum and maximum event times), so as to ensure a more biologically plausible function in the tails of the distribution where there is more likely to be sparse data (46). To fit a restricted cubic spline for variable x we create new variables, z, in the linear predictor, known as basis functions. We define *n* interior knots, k_1, …, k_n, and the boundary knots, k_{min} and k_{max}, then the spline function can be written in terms of parameters γ and the new variables z_1, …, z_{n+1} as below;

$$spline(x) = \gamma_0 + \gamma_0 z_1 + \cdots + \gamma_{n+1} z_{n+1}$$

Equation 1.7

The basis functions are calculated by;

$$z_1 = x$$
$$z_i = (x - k_i)_+^3 - \lambda_i (x - k_{min})_+^3 - (1 - \lambda_i)(x - k_{max})_+^3$$

Equation 1.8

Where λ_i may be calculated using the following formula for i = 2, …, n+1;

$$\lambda_i = \frac{k_{max} - k_i}{k_{max} - k_{min}}$$

Equation 1.9

FP model specification

FP models under the proportional hazards assumption are a generalisation of the Weibull model (46, 47). Royston and Parmar propose to extend the Weibull model as follows by first defining the Weibull log cumulative hazard;

$$\ln H(t) = \ln \lambda + \gamma_1 \ln t = \gamma_0 + \gamma_1 \ln t$$

Equation 1.10

Showing that the hazard can be described in terms of a constant and a linear function of log time, and therefore we can easily make the baseline component of the equation more flexible to improve on the Weibull model (where the baseline can only be a monotonic function over time). We can generalise Equation 1.10 with restricted cubic splines as follows,

$$\ln H(t) = spline(\ln t) = \gamma_0 + \gamma_1 \ln t + \gamma_2 z_1(\ln t) + \gamma_3 z_2(\ln t) + \cdots$$

Equation 1.11

Where we combine the constant and linear terms from Equation 1.10, with the basis functions from the restricted cubic spline as in Equation 1.8. There is a basis function for each knot in the model and a corresponding regression coefficient. It follows naturally that, Equation 1.11 collapses to a Weibull model in the case with zero knots. Equation 1.11 can be further extended to incorporate a set of predictors in a linear predictor, $\boldsymbol{\beta X}$ so that,

$$\ln H(t) = \ln H_0(t) + \boldsymbol{\beta X} = spline(\ln t) + \boldsymbol{\beta X}$$

Equation 1.12

Where $\boldsymbol{X} = (x_1 \dots x_j)^T$ and $\boldsymbol{\beta} = (\beta_1 \dots \beta_j)^T$ define vectors of predictors and their corresponding coefficients (log hazard ratios). Non-proportional hazards can also be assumed for the included predictors, $\boldsymbol{X}$ using restricted cubic splines, but this is not considered here.

The models are estimated using maximum likelihood, with suitable starting values for the coefficients, $\boldsymbol{\beta}$, derived from a Cox model with covariates $\boldsymbol{X}$ (44-46).

Using FP models for prognosis

In prognosis the emphasis is to obtain individual absolute risk predictions to inform treatment decisions. Individual predictions at time, t, can be obtained from an FP model just as in Equation 1.6, by first transforming Equation 1.12 above to obtain the baseline survival function and then using Equation 1.13 below to obtain survival probabilities, $S(t)$.

$$S_0(t) = \exp(-\exp(\ln H_0(t)))$$
$$S(t) = S_0(t)^{\exp(\boldsymbol{\beta X})}$$

Equation 1.13

1.6 Phases of prognostic model research

The following sections discuss the three established stages of prediction model research; (i) model development with internal validation, (ii) external validation, and (iii) impact evaluation (1, 9, 10, 26-28), and describes in detail the key statistical measures of model performance.

1.6.1 Model development

There are many statistical approaches and methods used for the development of prediction models and as such there is little agreement on the best strategy (27). There is however general consensus that good quality data and suitable sample size are required for all model development studies (48). In clinical research it is often the case that a more parsimonious model is preferred, owing to ease of implementation and face validity (26). The general aim is to build a model using a set of predefined predictors, which has good discrimination (able to distinguish between those at high and low risk of outcome), and calibration (agreement

between predicted and observed event rates) performance. Chapters two and three of this thesis focus on a review of published prediction models and a model development study, respectively. Below some important statistical considerations in model development are discussed, though many others exist (34, 38).

Model structure and candidate predictors

An appropriate regression framework should be selected based on the outcome for which predictions are required, for example dichotomous outcomes (such as presence or absence of disease) may be modelled using a logistic model, and time-to-event outcomes (such as time to death) could be modelled using either Cox or a (flexible) parametric survival model. Consideration must also be given to the specification of the intercept/ baseline hazard, especially when there are multiple centres or studies (clusters) within the development data; options include a fixed, proportional or random-effects baseline hazard which can be assumed across studies. This choice is also effected by the proposed implementation method for the model in practice. These issues will be revisited later in the thesis (in chapters 3 and 4), as they are unique issues arising from the use of clustered data for prediction modelling.

A set of candidate predictors for potential inclusion in the model should be predefined and should include predictors for which there is previous evidence or biological plausibility (27, 34). A systematic review of the literature is perhaps the best approach to identifying a set of candidate predictors, though there will often be more predictors than can sensibly be used in the model (27). In this case data reduction strategies are useful, to exclude candidate predictors which are highly correlated with each other for example (38). It is unwise to select candidates based on univariable selection methods, which can lead to either not identifying

important predictors or exclusion of predictors which are associated with the outcome after adjustment for confounders (other predictors) (38, 49).

Functional form of continuous predictors

It is very common in the literature for continuous predictors to be dichotomised at some arbitrary cut point, perhaps the median or, even worse, some data-driven 'optimal' cut-off (50-52). Continuous predictors should not be arbitrarily cut as this leads to a loss of statistical power to detect the true association between the predictor and outcome (50, 53). Categorisation also leads to a loss of prognostic information; splitting the predictor into categories assumes that those either side of the cut-off have very distinctly different prognosis, while those within a category are assumed to have similar prognosis, both strong assumptions (51). A better approach is to investigate whether the predictor has a linear or non-linear form, where the later could be modelled by some transformation (e.g. restricted cubic splines or fractional polynomials) of the original predictor (54, 55). Consideration should be taken with regard to the interpretation of more complex non-linear terms, as well as the models applicability and face validity.

Data quality and missing data

It is important that both predictors and outcomes have consistent definitions and measurement methods within the data used for development, that mimic those that will be available in practice to ensure the model's applicability. Problems often arise when there are multiple clusters with differences in definitions or measurement methods, and when predictors or outcomes with significant measurement error are included. These issues may reduce the predictive performance of the model (27).

Missing data is highly prevalent in medical research, and many statistical methods have been developed to handle this under certain assumptions (56, 57). Multiple imputation is preferable to a complete case (CC) analysis; firstly because exclusion of patients with missing predictor information reduces power and secondly, because as patient data is rarely missing completely at random, a CC analysis can lead to bias in predictor effect estimates (58). Finally predictors with large amounts of missing data may not be useful in practice for example because measurement is costly and the predictor is not routinely collected.

Selection of predictors for the final model

There are various methods for selection of predictors for the final model, though currently no agreement on the best approach (27, 38, 59). One such method includes all candidate predictors in a full model, regardless of their significance; this method potentially avoids overfitting and selection bias (38). Other approaches use automatic stepwise procedures, in which predictors are selected for inclusion or exclusion from the model based on significance tests at a pre-specified significance level. Such methods are data-driven and as such are susceptible to overfitting and inclusion of predictors based on spurious effects (and omission of genuine predictors by chance). Backward elimination, which aims to reduce a full model, is preferred to forward selection which tests predictors for inclusion from the null model (27, 60), because the former is based on effects that are fully adjusted for other factors.

1.6.2 Internal validation

The performance of a model measured in the development dataset is described as the apparent performance and is often optimistic (34). The use of selection procedures (e.g. to choose predictors or their functional form) in model development studies commonly results

in selection bias and overfitting, where predictor effects are likely overestimated leading to such optimism (38, 48). Overfitting is of greatest risk when the development dataset is small and where predictors have weak effects (27). Apparent performance is only likely to provide a valid estimate of model performance when the development dataset is extremely large (34). Conversely, without correction for overfitting in small datasets a developed model is highly likely to give poorer performance in new patients. Therefore, internal validation methods should be routinely included as part of any model development study, to quantify the optimism in model performance measures (see section 1.6.4). Some internal validation methods are now described below.

Split-sample validation

A common internal validation approach in the literature is to randomly dichotomise the dataset into 'training' and 'test' samples, for development and validation of the model, respectively. So called split-sample validation is not recommended as it often shows optimistic performance of the model, because the validation sample is very similar to the development sample as the data was only split at random (34). It has been shown that in small datasets the results of split-sample validation can vary depending on the split (48, 61, 62). Split-sample validation also raises questions as to whether there is adequate sample size for model development; in smaller datasets it may be more efficient to use all data for development and then rather apply a validation method that utilises the whole data such as cross-validation or bootstrapping (described below).

A potentially more useful variant is the non-random split-sample validation, where the dataset may be split by some clustering factor such geographical location, centre, study or time (28,

33, 63, 64). This may be considered as an intermediary to external validation, as the validation data may be distinctly different from the development data in some key characteristics (33). However similar problems as with split-sample validation may arise in small samples, and so this approach is best reserved for situations when the development data remains extremely large even after some clusters are removed. Temporal validation, using the same or similar patients at a different time point allows assessment of how model performance may change over time, for example changes in case-mix over time have been seen to cause severe calibration drift as in the original EUROSCORE model (65).

Cross-validation

Cross-validation improves on split-sample validation by using the whole dataset for both development and validation of the model. A common approach involves randomly splitting the dataset into 10 groups of equal size, though any number of groups may be used. The model is then developed using 9 of the splits, and validated in the 10^{th} split. This process is repeated so that 10 models are developed and tested in 10 validation samples. The performance measures calculated in each validation sample are then averaged to give the overall performance of the model. This approach can be performed at the individual level, where the model is developed on all but one patient (used for validation), however this is often computationally intensive and does not yield greater accuracy (38). The cross-validation approach is potentially most appealing in the context of meta-analysis, where a study could be removed for validation, and then the model developed on the remainder, with this process repeated across all cycles of the omitted study (20, 66). This 'internal-external cross-validation' approach will be utilised heavily in Chapter 3.

Bootstrap validation

Bootstrapping allows for use of the whole dataset through resampling with replacement from original dataset, meaning sample size can be maintained as opposed to splitting methods which reduce the development sample (38, 67). Bootstrap validation accounts for all uncertainty in the model development strategy, and allows estimation of the amount of optimism in the final model. When the degree of optimism is estimated it can be used to modify the developed model, uniformly shrinking predictor effects overestimated due to overfitting. Shrinkage of predictor effects in the final model can improve the potential performance in new patients (see section 1.6.4). Similar adjustment factors can be calculated for apparent performance statistics (e.g. C-statistic), giving better estimates of the models performance after accounting for optimism.

The bootstrap procedure begins by developing a model in the original data and calculating its apparent performance in the original dataset. A bootstrap sample is then created by sampling '*n*' patients with replacement from the original data. The model development strategy is repeated in the new bootstrap sample; it is critical that all steps of the original process are repeated in the new sample, as this accounts for all uncertainty in the development strategy (33, 38, 68). The performance of the model in the original dataset is calculated and then subtracted from the apparent performance of the model in the bootstrap sample to give the estimated optimism in performance. A new bootstrap sample is generated and the optimism calculated many times, for example 1000 bootstrap samples may be taken. The optimism calculated from each bootstrap sample is then averaged and subtracted from the apparent performance of the original model to give the optimism-corrected performance estimate.

1.6.3 External validation

It is well known that overfitting in development data often leads to poor performance of models in new patients outside of the development dataset (34, 38). Despite the use of internal validation to estimate potential optimism, evaluating the generalisability of the model requires external data (28). Even after shrinkage, the developed model could perform poorly in new populations due to poor methodology used in the development study (26, 28). However, a key reason for differences in model performance between development and validation samples is differences in case-mix, such as clinical setting (e.g. primary versus secondary care) or predictor-outcome distributions (e.g. adult versus child population) (26). Using a model in a new external population is essentially attempting to extrapolate the model, and as such reasonable performance may be seen where validation case-mix overlaps with the development cohort (26).

As discussed above, some instances of non-random split-sample validation may be arguably considered as external validation, for example where the developed model is validated in patients from a different country. In this way external validation is possible as part of a development study (33). Alternatively external validation may be performed on an existing published model using new patients, and performed by independent researchers, at a different time (26, 33, 34, 64, 69). Ideally external validity of the model should be tested multiple times in different external populations to assess its performance given variations in case-mix (70). Where multiple external datasets are available performance statistics can be summarised across datasets using meta-analysis methods as described later (see section 1.8.2) (19-21, 24, 71).

1.6.4 Model performance statistics

Measuring model performance both internally and externally requires statistical measures of both discrimination and calibration. Various statistics have been proposed to measure both properties (72), and the below sections discuss some of the most commonly used statistics within this thesis, which focuses on time-to-event prognostic models.

Discrimination

The discrimination performance of a model refers to the models ability to separate between patients with and without the outcome of interest by assigning higher risk probabilities to those who will have the outcome. A models discriminative performance is highly dependent upon the case-mix variation in the dataset, with greater discrimination (separation) seen in datasets with wider variation (70, 73). Given this it is common for the discrimination of a model to vary across different validation samples, further motivating the use of meta-analysis methods to summarise heterogeneity in performance across studies, centres or locations (24). Some new statistical measures of discrimination have been proposed to identify and adjust for heterogeneity in performance, but these are not considered in this thesis (73, 74).

Discrimination is commonly reported in terms of the C-statistic, sometimes referred to as the concordance index, is equivalent to the area under the receiver operating characteristics (ROC) curve for logistic regression models. The C-statistic is defined as the proportion of patient pairs (one with and one without the event), for which the model correctly assigns a higher predicted risk to the patient with the outcome (38). This calculation is complicated by censoring in time-to-event models, where Harrell's C-statistic can instead be used, which excludes any pairs which cannot be ordered. It is not possible to order pairs where either;

both patients have been censored, or both patients have an event at the same time point, or where one patient survives beyond another patients censoring time (38, 67). A C-statistic of one indicates perfect discrimination, whereas a value of 0.5 represents no discriminative ability beyond chance.

Royston's D-statistic is another measure of prognostic separation specifically proposed for survival curves (75). The D-statistic is strongly related to the standard deviation of the linear predictor; it gives the log hazard ratio between two groups defined by dichotomising the linear predictor at the median (75). Royston's R^2_D gives a measure of explained variation based on the D-statistic, to give an interpretation similar to R^2 in linear regression models (75).

Calibration

Calibration measures the agreement between expected (model predictions) and observed event probabilities. Calibration performance should ideally be evaluated across the risk spectrum, and it is common to consider deciles of predicted risk (68, 76). It can be used to indicate how much the model under or over-predicts the absolute risk of outcome. Poor calibration may be observed in external validation datasets, and in this situation recalibration or updating methods may help to improve performance. The use of recalibration using meta-analysis techniques is the focus of chapter 4 of this thesis (64, 69). Calibration performance is often neglected in validation studies and poorly reported in the literature (77).

In this thesis calibration performance is measured using both the difference in, and ratio of, the expected (E) and observed (O) event probabilities, E-O and E/O, respectively. Observed event probabilities can be calculated by averaging the outcome variable for logistic models,

or the survival probabilities estimated by the Kaplan-Meier method for survival models. Expected probabilities are calculated as the average of the models predicted probabilities. In survival analysis expected and observed probabilities are usually calculated at specific time-points (34). Perfect calibration is represented by a value of one for the E/O statistic, or zero for the E-O statistic.

Ratio or difference in E and O?

The E-O statistic is an appealing measure as it is easily interpretable as the absolute difference between the observed and predicted event probabilities (34), and can be visualised simply for survival models in a calibration plot as the difference between the Kaplan-Meier and model prediction curve over time. The ratio of E/O is less easily interpreted, as two cases with the same absolute difference in calibration may result in two very different E/O statistics depending on the size of the probabilities. For example, take an example where E=0.95 and O=0.9, the absolute difference is E-O=0.05 meaning the model predicts event probabilities 5% higher than observed for the population as a whole, and the ratio is E/O=1.05. Now take a second example with E=0.1 and O=0.05; here, E-O=0.05 again showing the model predicts event probabilities 5% higher than observed for the population as a whole, however the ratio E/O=2. The latter is considerably different than the previous 1.05 value, even though the absolute difference in risk is the same.

Other commonly used measures of calibration include the calibration slope and calibration-in-the-large (CITL), which are more frequently referred to in the case of logistic prediction models. In logistic regression the CITL is defined as the difference between the numbers of

predicted and observed events. It can be calculated either directly, or by fitting a logistic model with the LP fixed as an offset predictor (i.e. $\beta = 1$), giving α as the CITL;

$$Logit(p) = \alpha + \beta(LP)$$

Equation 1.14

The CITL for the logistic model is similar to the E-O statistic discussed above for survival models, but gives an estimate of calibration across the whole risk spectrum, unlike in the case of survival models where E-O must be calculated at specific time points.

The calibration slope for logistic regression relates to the slope of a calibration plot in which expected risk probabilities are plotted against observed risk probabilities (see Figure 1.5), across groups of patients with similar predicted risks (typically presented in deciles) (34). The calibration slope can be calculated using a logistic model as in Equation 1.14, where the LP is fitted (not as an offset) and the corresponding regression coefficient represents the calibration slope. Figure 1.5 illustrates four calibration plots with different calibration slopes, which relate to situations in which; (i) there is perfect calibration (i.e. E=O), (ii) there is some systematic miscalibration, in this case over prediction, (iii) there is large variation in predicted risks, potentially due to overfitting during model development, and finally (iv) there is large miscalibration, potentially seen when validating a model in a new population with very different prevalence of the outcome (71).

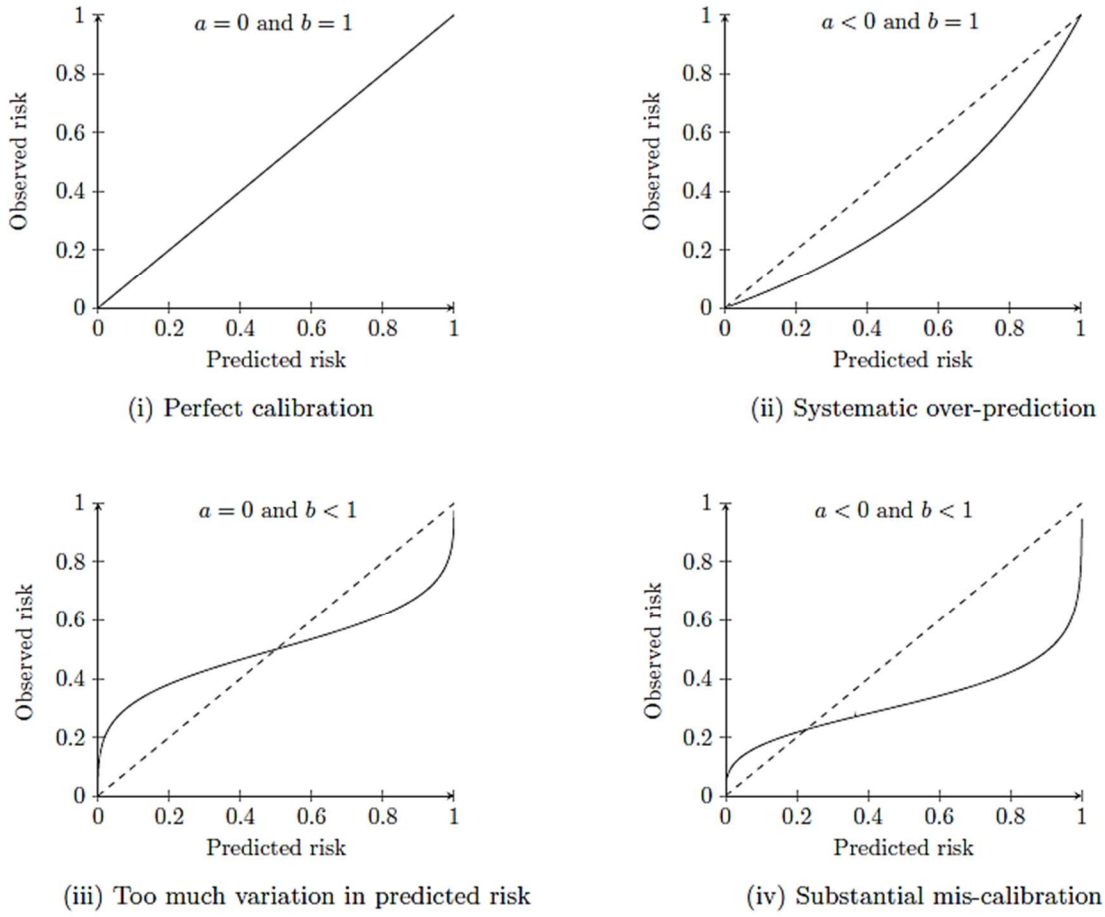


Figure 1.5 - Examples of calibration plots for logistic prediction models (71). Where the dashed line represents perfect calibration of E=O, and a=CITL and b=calibration slope.

1.6.5 Impact studies

The implicit aim of prediction models is to provide accurate predictions of patient outcome probabilities, to supplement decision making for clinicians and to ultimately improve patient healthcare and outcomes (1, 78). Well-developed models, which have shown promise through external validation and are easily applied in practice, should be assessed for their impact on patient outcomes, clinician decision-making and cost-effectiveness (26). Studies evaluating such outcomes are referred to as 'impact studies'.

Impact studies aim to measure the effect of using a model versus not using a model (usual care), and therefore are suited to a randomised controlled trial design (79). The impact of using a prediction model on patient outcomes and cost-effectiveness often requires long-term follow-up and is highly dependent on the uptake of the model by clinicians (26). It may therefore be more practical to conduct an initial study on behavioural change in clinicians, to assess the models uptake, with these studies requiring no patient follow-up. Randomisation by centres using a cluster design is preferred to avoid biases introduced by the clinicians within centres (26).

Impact studies may show a greater effect when the model is used to guide decision making by suggesting treatment strategies for patients falling within certain risk categories, as opposed to studies in which the absolute risk probability for the individual is provided alone and decision making is more subjective for the clinician (79). Similarly studies evaluating the impact of risk predictions provided automatically based on computerised systems may show greater performance (80). There are relatively few impact studies published in relation to the number of model development and validation studies published (9), but one good example is the STarTBack trial (81). The STarTBack trial randomised 1537 patients with back pain to either non-stratified usual care, or stratified care pathways based on patients predicted prognosis using the STarTBack tool (low, medium, high risk) (81). The trial found significant improvements in disability and quality of life measures in the stratified care group, as well as cost benefits (81). Chapters 5 and 6 of this thesis begin to move toward model implementation and impact, looking at meta-analysis methods for summarising and identifying the best thresholds for use of a model from published literature.

1.7 The TRIPOD statement

The TRIPOD (Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis) statement is a 22 item checklist, aiming to improve the quality of reporting and methods used for primary prediction model studies (33, 68). The checklist applies to all model development, validation and updating studies, and to all clinical settings. TRIPOD aims to better inform researchers in aspects of primary model studies including the design and analysis of such studies, and the interpretation of models and their generalisability. In terms of methodology, the accompanying explanation and elaboration paper provides detailed guidance on issues in development and validation including; handling of missing data, modelling of continuous predictors, selection of predictors for inclusion in the model, types of validation studies, model performance measures, and updating strategies (68).

The TRIPOD guidelines were developed to address the history of shortcomings in terms of both reporting and methodology in the prediction model literature (77, 82-84). TRIPOD will hopefully lead to future improvements in the reporting and methodology of prediction model studies, through uptake of the guidelines by the research community and the support of academic journals. While such improvements will likely take a substantial amount time, there has been some evidence to suggest that similar guidelines have improved the reporting quality in the literature. For example for the CONSORT statement for reporting of randomised trials (85), and for the STARD statement for reporting of diagnostic test accuracy studies (86, 87).

1.8 Systematic reviews and meta-analysis of prognosis and prediction studies

1.8.1 Systematic reviews

There is an abundance of published research investigating potential prognostic factors and prediction models, making systematic review and evidence synthesis highly desirable to inform clinical decision making and establish evidence-based research in prognosis. Traditionally systematic reviews summarise the available evidence on a particular treatment, though there is increasing interest in synthesis of prognostic factor effects or prediction model performance (8, 9, 71). Much research has been conducted on the methods for reviews of prognostic factors and prediction models including topics such as; search filters (88-90), study design (8, 9), data extraction (91), critical appraisal of published evidence (92-94), and synthesis of model performance (71). In this thesis, chapter 2 focuses on a systematic review of published prediction models, and utilises a preliminary version of the upcoming PROBAST tool for assessing risk of bias in primary prediction model studies (92, 95).

Systematic reviews of prediction models have previously shown the poor methodological and reporting quality of primary development and validation studies (77, 82-84), as well as the overflow of development studies and lack of respective validations (9, 82). The PRISMA (preferred reporting items for systematic reviews and meta-analyses), and PRISMA-IPD statements provide guidance on important reporting issues, which should be taken into account in any review and synthesis of prognostic factor or model studies (96, 97). Also the TRIPOD (Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis) statement aims to improve the quality of reporting and methods used for primary

prediction model studies, which should improve the quality of prognosis research in the future (33, 68).

This thesis aims to use meta-analytic methods for prognostic factor and prediction model research, with the aim of improving identification of important predictors, as well as prediction model performance, transportability and external validity. In the next sections the statistical methods for evidence synthesis are described for both aggregate data and IPD meta-analyses, as will be used in this thesis.

1.8.2 Traditional meta-analysis

Traditionally meta-analyses summarise the evidence on a single treatment effect obtained from study publications, typically as measured in a randomised controlled trial (RCT). However, it has become increasingly pertinent to summarise the association between potential prognostic factors and patient outcomes, due to the large number of often small studies, reporting conflicting results, in order to inform clinical practice. The increased interest in stratified medicine has also increased the use of meta-analysis to investigate potential effect-modifiers, or interactions, where treatment may have greater benefit or harm in particular subgroups of patients; for example it has been shown that Tamoxifen is only beneficial in estrogen receptor (ER) positive patients (32). Finally, meta-analysis is increasingly being used to summarise the external performance of prediction models, which is the focus in this thesis and the basis of the below model descriptions (19-21, 24, 71, 98).

Given multiple validation studies reporting estimates of model performance statistics (e.g. Harrell's C-statistic), a formal meta-analysis may be performed in order to summarise and

compare performance across studies (19, 20). Either a fixed-effect or random-effects meta-analysis may be used; a fixed-effect model assumes model performance is the same in each study, which is highly unlikely. Therefore, it is more appropriate to use a random-effects model as this allows for heterogeneity in performance, for example due to case-mix variation as discussed previously (see section 1.6.3) (70, 73).

Fixed and random-effects meta-analysis models

The statistical notation for fixed-effect and random-effects meta-analysis models are now given. Let $\hat{Y}_i$ be the estimate of a particular performance statistic of interest where $i = 1\ to\ I$ validation studies, and let $\hat{S}_i^2$ be the associated variance of $\hat{Y}_i$. A fixed-effect meta-analysis assumes that the true model performance is the same in all studies and that $\hat{Y}_i$ are estimates of this common performance θ as given by (99);

$$\hat{Y}_i \sim N\left(\theta, \hat{S}_i^2\right)$$

Equation 1.15

Maximum likelihood (ML) estimation using the inverse variance method is the most common approach to estimate the pooled effect, where a weighted average is calculated as (100);

$$\hat{\theta} = \frac{\sum_{i=1}^{I} \hat{Y}_i w_i}{\sum_{i=1}^{I} w_i}$$

Equation 1.16

Where;

$$w_i = \frac{1}{\hat{S}_i^2}$$

Equation 1.17

And therefore;

$$var(\hat{\theta}) = \frac{1}{\sum_{i=1}^{I} w_i}$$

Equation 1.18

Conversely, a random-effects meta-analysis allows for between-study variation in the performance statistic, assuming that the different studies are estimating different but related underlying performance statistics as given in Equation 1.19 below (99);

$$\hat{Y}_i \sim N\left(\theta_i, \hat{S}_i^2\right)$$
$$\theta_i \sim N(\theta, \tau^2)$$

Equation 1.19

This model assumes that $\hat{Y}_i$ follows a normal distribution around the i[th] study's true performance, θ_i, and that θ_i itself is normally distributed around an average performance, θ and a between-study variance τ^2. The between-study variance quantifies the heterogeneity across studies, and there has been much debate over how best to estimate τ^2 (101-104). Previous studies have shown that using restricted maximum likelihood (REML) estimation may better account for the uncertainty in the estimated τ^2 as ML is known to underestimate the between-study variance (105-110). Equation 1.19 reduces to Equation 1.15 when between-study variance is zero. The inverse variance method can be used to calculate the pooled effect estimate as in the fixed effects model, but now with weights calculated with the estimated between-study variance included;

$$w_i = \frac{1}{\hat{S}_i^2 + \hat{\tau}^2}$$

Equation 1.20

It should be noted that the random-effects model assumes between-study normality of true effects, which may not always hold for different model performance statistics. Current

research has suggested some transformations for discrimination and calibration performance statistics to approximate normality, but further research is still needed (24, 71, 111).

Usually of most interest will be the estimated $\hat{\theta}$ and its 95% confidence interval, which is commonly derived by $\hat{\theta} \pm 1.96\, SE(\hat{\theta})$, where $SE(\hat{\theta})$ is the standard error of $\hat{\theta}$. Other methods have been suggested to account for the uncertainty in the estimated between-study variance τ^2, with current research favouring the Hartung-Knapp-Sidik-Jonkman (HKSJ) method for calculating 95% confidence intervals (102, 104). Also of interest may be an approximate prediction interval, such as a 95% prediction interval (see Equation 1.21). A *t* distribution is used in the calculation of the prediction interval, rather than a normal distribution, to allow for the additional uncertainty in τ^2 (112, 113).

$$\hat{\theta} \pm t_{0.025}\sqrt{\tau^2 + SE(\hat{\theta})^2}$$

Equation 1.21

The prediction interval infers the potential model performance in a new population similar to those included in the meta-analysis (112, 113). A narrower prediction interval implies more consistent performance in new external populations, and is thus desirable if the model is to be generalizable outside of a few local settings. This equation is only an approximation (114), and perhaps the ideal, more natural framework for predictive inferences is a Bayesian framework, though this is outside the scope of the thesis.

Issues with traditional meta-analysis

The use of traditional aggregate data meta-analysis in the prognosis field is hindered by widespread poor reporting and publication bias. Sources of heterogeneity are therefore rife

in meta-analyses of prognostic factors, interaction effects and model performance. Primary studies may use different statistical analysis methods, adjust for different sets of predictors and report different effect measures such as odds or hazard ratios, which may be adjusted or unadjusted effect estimates. There is often poor quality of primary studies, with many small studies, with no protocol, and no clear aims. Differences in treatments used, disease stage, methods of measurement and cut-offs used for factors or models, all leads to large heterogeneity in the meta-analysis diluting the evidence base and conclusions. However many of these issues can be addressed if IPD are available and an IPD meta-analysis performed.

1.8.3 Individual participant data (IPD) meta-analysis

Advantages over aggregate data meta-analysis

IPD meta-analysis uses the raw patient level data to calculate and synthesise the effect of interest from each study. Many of the issues discussed above seen in traditional meta-analysis such as poor reporting and publication bias can be addressed using an IPD meta-analysis (115-117). Having IPD allows re-analysis of the data using consistent choices for analysis in each study such as; choice of adjustment predictors, choice of cut-offs for continuous predictors, handling of missing data, length of follow-up, examination of non-linear trends, and assessment of modelling assumptions. It is also possible to check the results of the original study publications by calculating the aggregate data directly within each study. Where model performance is of interest any unreported statistics (such as calibration slopes, E/O, or C-statistics), can be calculated manually where IPD is available. Importantly having IPD allows greater investigation of the causes of between-study heterogeneity, and potential effect-

modifiers (interactions) as discussed earlier (see section 1.8.2) by avoiding ecological bias (118).

Statistical approaches to an IPD meta-analysis

IPD meta-analyses can be performed using two different approaches; a two-stage or one-stage IPD meta-analysis (119, 120). The two-stage method first analyses each study separately to obtain the aggregate data needed for the second stage, in which a standard fixed-effect or random-effects meta-analysis model is used as described above. The one-stage method instead analyses all the IPD from all studies in one model allowing for the clustering of patients within studies (16). Usually results are similar for the two methods, though there are a number of reasons why a two-stage and one-stage analysis may yield different results (120, 121).

Issues with IPD meta-analysis

Despite the many benefits of using IPD meta-analysis over aggregate data meta-analysis, the IPD approach still has many challenges. There are many logistical challenges facing an IPD meta-analysis study; in particular, it is often time-consuming and costly in terms of obtaining, cleaning and analysing the data. Despite extensive efforts to obtain IPD from all identified study authors, IPD may still be unavailable for some studies leading to an availability bias in the analyses (122). Further, issues of publication bias are just as relevant in an IPD analysis as for a traditional aggregate data meta-analysis.

As with traditional meta-analysis many issues remain for IPD analyses due to primary study deficiencies. Even with the raw patient level data some differences between studies cannot be addressed; for example, differences in outcome definitions, methods of measurement,

treatment strategies, or available predictors within each study. Continuous predictors may have been categorised by the study authors, making reconstruction of the continuous predictor impossible. Where predictors have not be measured in particular studies results in systematic missingness, in recent years a number of studies have developed methods to handle systematically missing predictors using multivariate meta-analysis and imputation, though this is not considered in this thesis (123-130).

1.9 Current challenges facing prediction model research

This chapter has discussed key aspects of the design and methods used in prognosis research, and prediction modelling studies in particular. This thesis aims to make a contribution toward improved methodology in prognosis research. Therefore, in this section the thesis topics are motivated by summarising the existing challenges facing the field, with particular attention given to prognostic model research as this forms the main focus of the majority of subsequent chapters in the thesis.

In recent years an ever increasing number of prediction models have been published, with many published models developed to answer the same research questions (9, 77, 82, 84, 131-133). For example a recent review by Damen et al. found 363 prediction model development studies all predicting cardiovascular disease (CVD) risk in the general population (82), while Perel et al. found 102 competing models predicting outcomes in patients with traumatic brain injury (TBI) (84). Evidence also suggests that many of these primary development and also validation studies are of poor methodological and reporting quality (77, 133-136). As discussed earlier, this has led to the creation of the TRIPOD statement which aims to improve

the reporting of primary studies for prediction model development, validation and updating in the future (33, 68).

Despite the increasing number of development studies there has not be a corresponding increase in validation and impact studies evaluating existing models (9). Where external validation studies have been conducted there is evidence of poor quality and in particular there has been an emphasis on model discrimination, with calibration performance of the model often not reported (77). Calibration is a key feature of model performance as discussed above (see section 1.6.4), with model miscalibration leading to invalid individualised risks, and subsequently to poor or even harmful clinical decision making (65, 76).

Development of new models based on poor performance of existing models at external validation, or worse without any external validation of existing models, is counterintuitive to evidence based medicine and simply research waste (64, 69). Therefore, rather than aiming for new models, new research should aim to validate existing prediction models and compare competing models head-to-head in new external datasets to identify the best performing model for future impact studies (137-140). Another approach to reduce future research waste is to tailor models to new external populations, through recalibration or updating methods (34, 64, 69, 141-145). Recalibration methods are discussed in detail in chapter 4 of this thesis, and involve adjusting an existing model to the characteristics of individuals in a new population (146). In this way recalibration and updating methods combine evidence from previous patients used to develop the existing model, with information on new patients (64).

However, once an existing model has been adjusted (even in the simplest manner), it can be considered a new model in need of external validation before use in new patients external to the update population (147). This implies that some cycle of constant model updating and immediate validation is needed, and points toward the use of a more dynamic prediction approach. Future prognosis research may look to develop the existing methodology for dynamic prediction, so that models learn from each new patient (or measurement) they are applied to in a similar way to how a doctor learns, making ever more informed decisions with experience. Current methods include the joint modelling of longitudinal measurements (e.g. biomarker levels over time) to inform prognosis of a time-to-event outcome (e.g. myocardial infarction) (148, 149).

Given the advent of 'big data' and IPD meta-analyses of large datasets, methods for externally validating, comparing, recalibrating and updating prediction models should become easier in the future (21, 82). These datasets will often contain clustering of patients by centre, study or country, which naturally lend such data to meta-analytic analyses, but with much greater opportunity (and statistical power) to investigate heterogeneity (21). Between-study heterogeneity in model performance can be interrogated further in terms of patient case-mix, and particularly heterogeneity in predictor effects between-studies.

Finally in the interests of reducing research waste and with the wealth of competing models available, future research may look to improve methods for synthesis of the models themselves. Previous research has proposed the aggregation of prediction models, forming a kind of 'meta-model', as an amalgamation of the available models regression coefficients (23). Current research suggests model performance is improved by model aggregation in small

external datasets, while model development may be preferable in larger datasets with significant between-study heterogeneity, though it should be noted that updating methods were not considered in these studies (23, 150).

1.10 Aims and outline of the thesis

Given the increasing number of prognosis studies, and the availability of big datasets and IPD from multiple studies, it is clear that meta-analysis methods will play a critical role in the development and evaluation of prognostic factors and, in particular, prognostic models in the coming years. This represents an exciting opportunity, which forms the focus of this thesis.

The broad aim of the thesis is to apply and develop methods for evidence synthesis of prognosis research, in particular to improve individualised predictions from prognostic models developed and/or validated using meta-analysis techniques. The key aims are:

- To review the methodology and reporting of prognostic models in a specific clinical setting, to illustrate the importance of a systematic review and meta-analysis of prognostic models, and the difficulties of applying meta-analysis without IPD;
- To use IPD from multiple studies to develop and validate a new prognostic model in a clinically relevant setting;
- To develop and illustrate evidence synthesis methods that externally validate the performance of an existing prognostic model using IPD from multiple studies;
- To extend such methods to identify how to best recalibrate and update an existing model when its external validation performance is otherwise heterogeneous;

- To develop and examine, through simulation, a novel method for improving meta-analysis of prognostic studies when studies do not provide IPD but do provide the predictive accuracy of a continuous factor at different thresholds.

The thesis therefore includes a mixture of clinical application and methodology development, and the chapters are briefly outlined below.

Chapter 2 aims to systematically review all available evidence on prognostic models for predicting the risk of a recurrent VTE. The review discusses many aspects of prognostic model development and validation through a critique of the existing literature in the clinical area. The identified models differ in many key areas including; selection of predictors, handling of missing data, patient selection, statistical analysis and model validation. Risk of bias assessments suggest models require further validation. Recommendations for further research are provided, and the review motivates the work in chapter 3.

Chapter 3 builds on the previous chapter which reviewed the available evidence and identified a number of existing models with many methodological issues. The aim is to develop a new prognostic model using IPD from seven studies, using meta-analysis techniques to account for the clustered nature of the data. In particular an internal-external cross-validation (IECV) approach is used to maximise the use of the data for development and validation of the model. Model development aims to overcome the methodological issues identified in the previous chapter, and uses the IECV framework to examine the model's calibration and discrimination performance across multiple settings. The work arising from Chapters 2 and 3 has been published in Health Technology Assessment and BMJ Open (95, 98).

Chapter 4 aims to investigate how model recalibration methods help to improve model performance when IPD from multiple validation studies are available. Four options are examined to recalibrate an existing flexible parametric survival model in breast cancer across multiple centres and countries by: (i) shifting the baseline hazard by a constant, (ii) re-estimation of the shape of the baseline hazard, (iii) adjustment of the linear predictor as a whole (calibration slope), and (iv) adjustment of individual predictor effects. IPD meta-analysis is used to examine calibration and discrimination performance across studies for each of the strategies, to ascertain if and how they improve performance and reduce heterogeneity. Recommendations are given for those using IPD meta-analysis for external validation on the use of an existing model in new populations. This work has been presented at the 37th International Society of Clinical Biostatistics Conference (ISCB), and the Royal Statistics Society 2016 International Conference, and is being drafted for submission to Statistics in Medicine.

Chapter 5 tackles the situation when IPD are not available, and aims to develop a new method to deal with missing (partially reported) threshold information in test accuracy meta-analysis, where a single prognostic factor (predictive test) is used to inform the prognosis of an outcome for an individual patient. For continuous tests, primary studies usually report predictive test accuracy results at multiple thresholds, but the set of thresholds used often differs. Without IPD, this creates missing data when performing a meta-analysis at each threshold. A standard meta-analysis (NI: No Imputation) ignores such missing data. A Single Imputation (SI) approach has been proposed to recover missing threshold results using a simple piecewise linear interpolation. In this chapter a new method is proposed that performs Multiple Imputation of the missing threshold results using Discrete Combinations (MIDC), to

address short-comings of the SI method. Stata software code is developed for others to use, and an example is given that shows how the MIDC method may give very different results to the NI method.

Chapter 6 covers an extensive simulation study which aims to evaluate the statistical properties of the new MIDC method developed in chapter 5. The study is designed to evaluate the performance of new MIDC method in comparison to the previously proposed SI method and when missing data is ignored (NI). Several scenarios are considered, including varying the amount of missing data, the missingness mechanism and the assumed spacing of reported thresholds. The findings indicate that the MIDC method gives best performance in general, unless extreme unequal spacing is apparent. The work arising from Chapters 5 and 6 has been submitted for publication in Research Synthesis Methods.

Finally, chapter 7 concludes with a summary of the key findings and recommendations from the thesis, and outlines the limitations and areas for further research.

CHAPTER 2: A SYSTEMATIC REVIEW OF PROGNOSTIC MODELS FOR RECURRENT VENOUS THROMBOEMBOLISM (VTE) POST TREATMENT OF FIRST UNPROVOKED VTE

2.1 Introduction

In this chapter, the new research for this thesis begins by identifying and examining the quality of prognostic models in the important clinical area of venous thromboembolism (VTE), which is the third most common cardiovascular disease after heart attack and stroke. VTE is a chronic condition with estimated incidence at 1 per 1000 person years (151-153), and often presents as deep vein thrombosis (DVT), with some patients suffering an embolism in the lungs known as a pulmonary embolism (PE). An initial VTE developed in the presence of a known provoking factor may be termed “provoked”, while those developed in the absence of clinical risk factors may be termed “unprovoked” (153, 154). There are several known pre-disposing risk factors including surgery, trauma, hormone intake, pregnancy and prolonged immobility (153, 155). Such provoking factors can be considered as acquired risk factors because they are transient; that is, while they increase the risk of an initial VTE, they are temporary and when they are removed the patient is at a low risk of recurrence, for example post-surgery (153-155).

The aim of therapy for VTE is twofold: initially to prevent extension of the acute thrombosis, and secondarily to prevent both recurrence and long term sequelae such as post-thrombotic syndrome and pulmonary hypertension. Current treatment comprises initial management with heparin, usually low-molecular weight heparin for a minimum of five days, overlapping with oral anticoagulant therapy (usually warfarin in the UK) until the International Normalised

Ratio (INR) is above two. It is usual to treat an initial VTE for a minimum of three months however the optimum duration of therapy beyond this is unclear (156, 157). Treatment with Novel Oral Anticoagulants (NOAC) is a new alternative treatment to heparin and warfarin.

Due to the transient nature of provoking factors, patients with a first unprovoked VTE are at much higher risk of recurrent VTE (approaching 30% at five years after cessation of therapy) as the cause is unknown (153, 155). Prevention of recurrent VTE poses a difficult clinical decision problem; a balance must be struck between the risks of recurrent thrombosis if anticoagulant treatment is stopped versus the risks of bleeding associated with continued anticoagulation therapy (153, 156).

Therefore it is important to identify individuals with a high risk of VTE recurrence compared to the risk of major bleeding on anticoagulation, in order to inform treatment strategies. However, the population of patients with unprovoked VTE is highly heterogeneous and risk of VTE recurrence varies considerably across individuals (37, 39, 40). Therefore there is much interest in developing prognostic models for VTE recurrence. As described in Chapter 1, a prognostic model is a statistical equation that predicts an individual's outcome risk based on a weighted combination of multiple predictors (e.g. age, sex, biomarkers) (9). A key stage of prediction model research is model development. This uses a dataset to identify important predictors and then develops the model equation; it usually also examines the model's apparent performance in this same data, possibly using resampling techniques to adjust for optimism (internal validation). The next stage is external validation, which uses data external to the model development data and its source, and examines whether the model predictions are accurate in independent data from the same or another (related) setting. External

validation is crucial as model performance is usually over-optimistic when considered only in the development dataset (9, 33, 34). Validation typically focuses on a model's predictive performance as measured by discrimination (i.e. the model's ability to separate those with and without the outcome) and calibration (i.e. the agreement between the model's predicted risk and the observed outcome risk) as described in chapter 1.

2.1.1 Aims of this chapter

A reliable prognostic model is needed for the unprovoked VTE population, in order to inform clinical and patient decision making with regard to treatment strategies (11), in particular whether or not to extend treatment beyond the initial period (e.g. 3 months) with oral anticoagulants (OAC) to prevent recurrent VTE. However, there are multiple published studies describing the development and validation of prognostic models for VTE recurrence.

Therefore, in this chapter the aim is to perform a systematic review to identify and summarise studies developing or validating a prognostic model for individual VTE recurrence risk following cessation of therapy for a first unprovoked VTE. Through the identification of existing studies the review will help to determine whether reliable prognostic models exist and, if not, what further research is needed within the field. In particular, the review appraises the quality of evidence for and against each existing model, to help clinicians and other practitioners to better understand their strengths and weaknesses (11), allowing more informed decisions to be made on which (if any) models to use in practice. A protocol for the review was registered with PROSPERO (CRD42013003494) and published in Systematic Reviews (158), and the findings of the review were published in BMJ Open in 2016 (95).

This chapter now begins by describing the methods of the review, followed by detailing the results and conclusions. It should be noted that a team of clinical and methodology experts were involved in the review on a supervisory level. However, all components of the review were primarily conducted by the PhD candidate (Joie Ensor), including the searches, selection of relevant studies, data extraction, qualitative and quantitative summaries, and recommendations.

2.2 Methods

The objectives of the review were to:

- Identify relevant articles that described either development or validation of a prediction model predicting the risk of recurrent VTE or adverse outcome following cessation of therapy for a first unprovoked VTE
- Summarise the quality (risk of bias) of identified studies
- Qualitatively summarise the content of the models identified and their predictive performance
- If considered appropriate, use meta-analysis to quantitatively summarise (pool) predictive performance across studies

2.2.1 Search strategy to identify relevant studies

The following bibliographic databases were searched: Cochrane Library (Wiley) (including the Cochrane Database of Systematic Reviews, DARE, HTA Databases and CENTRAL Register of Controlled Trials), MEDLINE (Ovid) 1950- July 2014, MEDLINE In - Process & Other Non-Indexed Citations (Ovid) to date and EMBASE (Ovid) 1980- July 2014. Searches used index

terms and text words that encompassed the patient group supplemented by terms relating to recurrence or adverse outcome and prognostic factors; an example of the search strategy as used for MEDLINE is presented below;

```
Database: MEDLINE (Ovid) 1946 to July Week 3, 2014
Search strategy:
1. exp Venous Thromboembolism/
2. Pulmonary Embolism/
3. exp Venous Thrombosis/
4. (vte or dvt or pe).ti,ab.
5. deep vein thrombosis.ti,ab.
6. pulmonary embolism.ti,ab.
7. venous thrombo$.ti,ab.
8. or/1-7
9. (recurrence or recurr$ or re-occur$).ti,ab.
10. Recurrence/
11. exp Death/
12. (death$ or mortality).ti,ab.
13. Mortality/
14. clot$.ti,ab.
15. Hypertension, Pulmonary/
16. pulmonary hypertension.ti,ab.
17. post thrombotic syndrome.ti,ab.
18. PTS.ti,ab.
19. or/9-18
20. “Predictive Value of Tests”/
21. predict$.ti,ab.
22. exp Risk/
23. risk$.ti,ab.
24. prognos$.ti,ab.
25. or/20-24
26. exp Anticoagulants/
27. (anti-coagul$ or anticoagul$ or warfarin or acenocoumarol
or coumadin or coumarin or phenprocoumon or sintrom or
sinthrome or jantoven or marevan or waran or nicoumalone or
dicoumarol or dicumarol).ti,ab.
28. (phenindione or dabigatran or ximelagatran or apixaban or
rivaroxaban or edoxaban or azd0837 or ly517717 or ym150 or
betrixaban or idraparinux).ti, ab.
29. or/26-28
30. 8 and 19 and 25 and 29
```

Publicly available trials registers were also searched, such as ClinicalTrials.gov, UK Clinical Research Network Study Portfolio Database (UKCRN), WHO International Clinical Trials Registry Platform and the metaRegister of Controlled Trials (mRCT). Reference lists of all included papers were checked and subject experts were contacted. No restrictions on publication language were applied.

In addition, abstracts from the Conference Proceedings Citation Index (CPCI) were searched in order to capture studies that were not yet fully published.

2.2.2 Inclusion criteria

Study Design: Studies of any design (e.g. cohorts, RCTs) or systematic reviews that developed, compared or validated a prognostic model (or clinical prediction rule based on a model) utilising multiple (at least two) predictors to predict the risk of recurrent VTE or adverse outcome (mortality or bleeding) following cessation of therapy for a first unprovoked VTE. The decision to focus on multivariable models was a consensus amongst the wider study team, who deemed VTE recurrence as a multi-faceted problem, and thus unlikely to be explained by just a single predictor.

Patient group: Relevant patients were those aged ≥ 18 years with a first unprovoked VTE where the patient had received at least three months treatment with an OAC therapy. Studies with mixed populations (including those outside of remit) were included provided that appropriate data for the defined group of relevant patients was extractable.

Setting: Studies in any setting were included.

Potential prognostic models: Studies were included if they reported a prognostic model utilising multiple predictors to predict the risk of recurrent VTE or adverse outcome following cessation of therapy for a first unprovoked VTE, in the defined patient group of interest. A prognostic model was defined as a combination of at least two predictors within a statistical model (e.g. a multivariable regression model), used to predict an individual's risk of outcome (e.g. VTE recurrence).

2.2.3 Study selection

Study selection followed a two-step process. Titles and (where available) abstracts were initially screened by two reviewers (Joie Ensor and David Fitzmaurice) independently, using predefined screening criteria. These were broadly based on whether studies, 1) included patients with a first unprovoked VTE, who received a minimum of three months OAC therapy, and 2) developed or examined prognostic models in relation to individual prediction of VTE recurrence or other adverse outcomes (mortality or bleeding).

Full texts of any potentially relevant articles were then obtained and two reviewers independently applied the full inclusion criteria (see APPENDIX A: Chapter 2 Appendices). Any discrepancies between reviewers were resolved by discussion or by referral to a third reviewer. Portions of non-English language studies were translated where necessary to facilitate study selection and subsequent data extraction. The study selection process was documented using the PRISMA flow diagram. Any relevant systematic reviews identified were screened to identify any further primary studies. Reference management software (Endnote) was used to record reviewer decisions, including reasons for exclusion.

2.2.4 Data extraction

In those articles deemed relevant, data extraction was then conducted independently by two reviewers (Joie Ensor and Kym Snell) using an in-depth piloted data extraction form. Disagreements were resolved through discussion or referral to a third reviewer.

Data extraction included the following elements:

- Study characteristics (e.g. sample size, country, year)
- Study design characteristics (e.g. design, length of follow-up)
- Patient characteristics (e.g. summaries of age, sex, family history, treatment details in the sample)
- Candidate predictors considered and their definitions (e.g. any thresholds used for continuous predictors, methods of measurement, timing of measurement post cessation of therapy)
- Outcome measures (e.g. recurrence of VTE, mortality, bleeding)
- Statistical methods employed and how predictors included in the analysis were handled (e.g. continuous vs. dichotomised).
- Prognostic model details, including: the final model equation and included predictors; how the model was developed and how it can be used to obtain an individual's risk probability; and any internal and external validation performance statistics for model performance (including discrimination and calibration) together with their confidence intervals.

2.2.5 Assessment of study quality (risk of bias)

The quality (risk of bias) of any studies developing or evaluating a prognostic model was assessed by piloting an early version of PROBAST (Prediction study Risk Of Bias Assessment Tool), a tool for assessing risk of bias and applicability of prognostic model studies, that was nearing completion and ready for piloting when this review was undertaken (92). PROBAST defines risk of bias as "any flaw or shortcoming in the design, conduct or analysis of a primary study that is likely to distort the predictive performance of a model." In particular, bias that would lead the reported calibration and discrimination of a model to be systematically wrong (beyond the play of chance).

Particular elements were considered in the following domains:

- Patient selection, such as
    - what study design was used (e.g. prospective),
    - if appropriate inclusions and exclusions were used, and
    - whether patients had similar disease presentation, or if this was accounted for in analyses
- Outcomes, such as whether
    - the outcome definition was pre-specified,
    - included predictors were excluded from the outcome definition,
    - the same definition and assessment was used for predictors and outcomes in all patients, and
    - the outcome was determined blind to predictor information
- Predictors, such as whether

- the same predictor definitions were used for all patients,
- predictors were measured blinded to outcome data,
- all predictor information was available at the time the model was intended for use, and
- non-linear associations for continuous predictors were considered and, if undertaken, predictor categorisation was not data-driven

- Sample size, such as
  - whether there was a pre-specified sample size consideration for model development accounting for numbers of events and multiple comparisons in selection of predictors,
  - whether all enrolled patients were included in analyses, and
  - how much data was available for external validation
- Missing data, including whether
  - there was adequate reporting on completeness of data, and
  - multiple imputation was considered
- Statistical analysis, such as
  - handling of continuous predictors,
  - selection of possible predictors irrespective of univariable analyses,
  - whether weights assigned to predictors in the final model's statistical equation related to the same regression coefficients as from the fitted model in the development data
- Internal and external model validation

- Whether model validations in terms of predictive performance were reported and how these were obtained; in particular, whether calibration and discrimination statistics were presented, and, during internal validation, whether over-fitting and optimism was evaluated and accounted for (e.g. using bootstrapping or shrinkage techniques)

2.2.6 Summarising identified evidence

For each unique model identified, the evidence available was summarised using the extracted data. In particular, each model was narratively summarised in terms of; the model development and validation methodology, the included predictors and how they were coded, the specification of the model and how it could be used, whether the model was validated internally and externally (and if so how), and the reported performance of the model in terms of calibration and discrimination. The PROBAST evaluation was used to determine the risk of bias of the model (that is, whether the model is likely to work as intended for the VTE population of interest), with model's classed as low, moderate, or high risk of bias.

The consistency of development methods used and main findings were examined to identify whether studies at higher risk of bias produced different results and conclusions to those considered to be at low risk of bias.

If multiple studies were found that validated the same prognostic model, it was planned to meta-analyse estimates of calibration (e.g. Expected/Observed events) and discriminatory (e.g. C statistic) performance using a random-effects meta-analysis (113, 159), to summarise

the model's average performance across different settings and its potential performance in a future setting.

2.2.7 Relevant articles identified outside of search dates

Before publication of the review in BMJ Open, a further search was performed beyond the initial literature search period, in case further studies had been published. Two relevant studies were identified; these were published in February 2015 (160) and September 2015 respectively (161); both of these will be discussed in detail later as evidence found outside the systematic review searches (see section 2.3.4).

2.3 Results

2.3.1 Quantity of research available

Searching of bibliographic databases resulted in 13,516 records identified after automatic removal of 1,879 duplicates. A further 2,747 duplicate records were manually removed, leaving 10,769 records to be screened for inclusion. Screening of titles and abstracts identified 10,485 records irrelevant to the review question. Full text articles were sought for eligibility assessment, three articles were unobtainable from the British library (162-164) and a further three articles were unable to be translated into English (165-167) out of 19 non-English language articles (i.e. 16 were translated). Of the 278 full text articles assessed for inclusion, 258 articles were excluded with;

- 91 articles excluded as discussion or review articles that did not develop or update a prognostic model,

- 150 articles were excluded based on issues related to the model (e.g. not for individual prediction, emphasis on the effect of a single predictor etc.),
- 3 articles were excluded based on the study population, and
- 14 were excluded based on both population and model issues (see Figure 2.1).

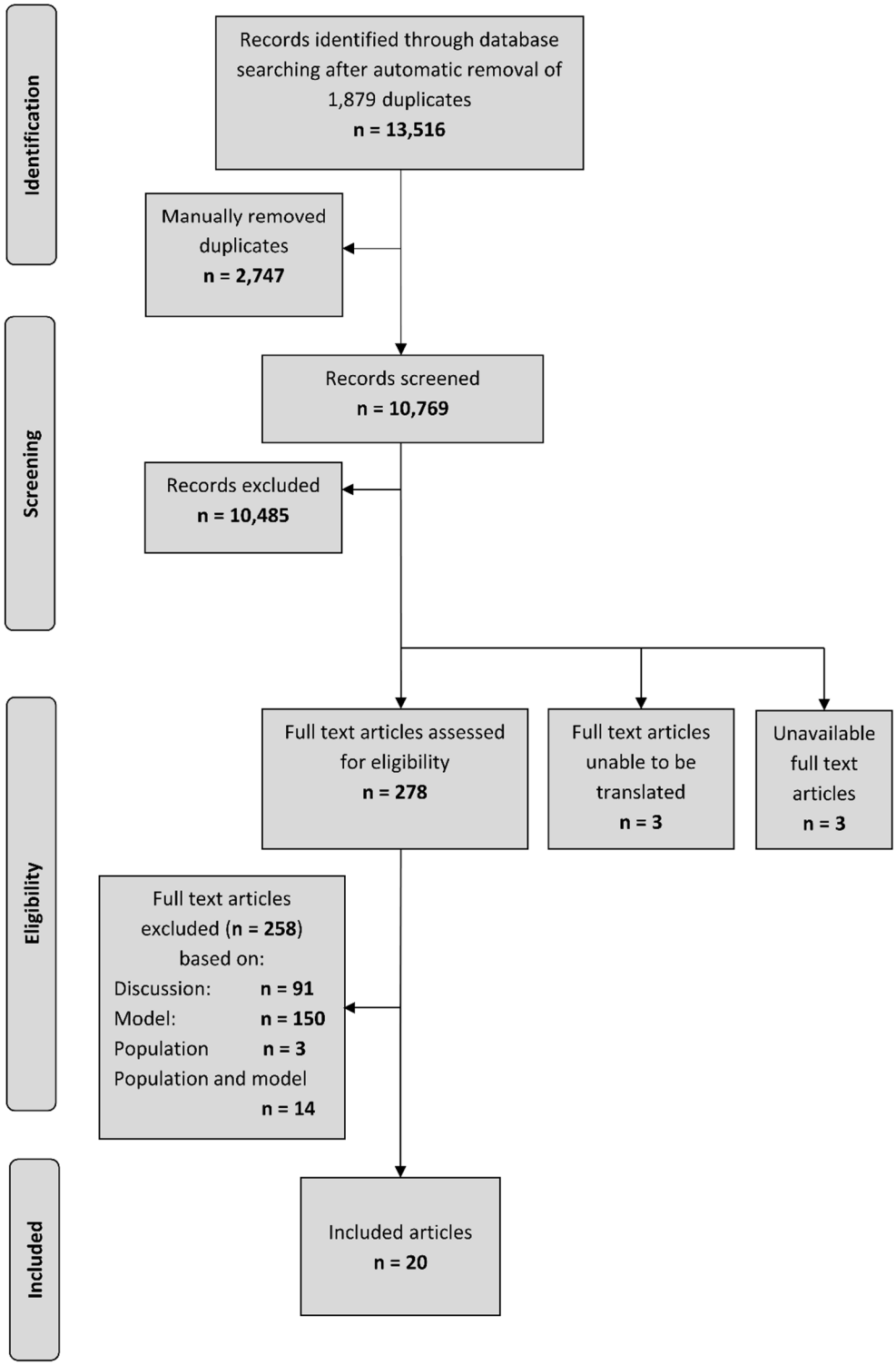


Figure 2.1 - PRISMA flow diagram showing the quantity of research available.

Twenty articles therefore met the inclusion criteria after screening, comprising seven on-going studies (168-174), eight conference abstracts (175-182), one project record referring to the project this work forms part of (98, 183), and four full text peer-reviewed articles (37, 39, 40, 184).

The authors of the 15 conference abstracts and on-going studies were contacted to seek additional information. Based on author responses 13 of the 15 abstracts/on-going studies were associated with the four full text articles already identified. The authors of the remaining two articles (which were both abstracts) did not respond to further enquiry and so no further publications could be found to supplement the available abstracts (168, 176). One, a study by Raskob et al. (176) is based on data from the EINSTEIN extension study (185), and aimed to identify a subgroup of patients at high and low risk of recurrent VTE. Further information regarding the study was unavailable from the included abstract, therefore it was unclear whether a prognostic model was developed and if individual recurrence risk could be predicted from such a model. The second abstract relates to the ongoing VISTA study (168), discussed later in the article.

The remainder of this chapter focuses on summarising and critiquing in detail three of the full text articles included in the review (and their 13 associated abstracts). The fourth full text article was an update to one of the earlier models and is given special attention later (see section 2.3.3). First, a brief introduction to the full text articles and the models developed is given. Throughout this chapter these articles will be referred to using the name of the corresponding model developed (i.e. HERDOO2, Vienna, and DASH).

HERDOO2 model (39)

Rodger et al. used conditional logistic regression to develop a prognostic model for use as a clinical decision rule (see Figure 2.2). This suggested that a female patient with less than two predictors (post-thrombotic signs (either leg hyperpigmentation, edema or redness), D-dimer level ≥ 250 µg/L, BMI ≥ 30 kg/m^2 or age ≥ 65 years) could potentially safely discontinue OAC therapy after five to seven months of initial OAC therapy for an unprovoked VTE. A low risk (< 3% annual recurrence risk) group of males could not be identified in the study and therefore Rodger et al. recommended that all male patients continue OAC therapy (39).

Vienna prediction model (40, 184)

Eichinger et al. used a Cox proportional hazards model to develop a prognostic model including sex, site of index event and D-dimer as predictors (see Figure 2.2). A nomogram based on the prognostic model was derived to allow easy implementation of the model and can be used to calculate patient's cumulative recurrence rate at 12 and 60 months from cessation of therapy, with estimated 95% confidence intervals (40). Another full text article included in the review describes an update to the proposed Vienna model (see section 2.3.3), by recalculating the model at 3, 9 and 15 months after cessation of therapy using new measurements of D-dimer levels at these time points. Eichinger et al. used a dynamic prediction approach in the updated model and adapted a Fine-Gray model to allow for the competing risk between recurrence and death (in some of those who restart therapy) (184).

DASH score (37)

Tosetto et al. used a Cox proportional hazards model to develop a prognostic model including predictors for abnormal D-dimer levels (+2 score), age ≤ 50 years (+1 score), male sex (+1

score) and hormone use (-2 score) (see Figure 2.2). This proposed score can be used to calculate patient's cumulative recurrence risk at one, two and five years from cessation of therapy, with estimated 95% confidence intervals. Tosetto et al. suggest that a combined DASH score ≤ 1 would indicate an annual recurrence risk < 5% and therefore indicate that a patient could potentially stop OAC therapy, conversely a DASH score > 1 would indicate annual recurrence risk > 5% and thus suggest patients should potentially continue OAC therapy (37).

Linear predictors of models included in the review

A linear predictor (LP) describes the combination of the included predictors and their estimated regression coefficients as described in chapter 1. The following describes the linear predictors for each of the included models.

HERDOO2

The HERDOO2 model is a logistic regression model for which the LP was defined as follows, where the regression coefficients represent log odds ratios;

LP = –3.9717 + (1.2977 x BMI ≥ 30 kg/m2) + (0.6473 x post-thrombotic signs) + (0.9155 x D-dimer ≥ 250 µg/L) + (0.8084 x age ≥ 65 years)

Vienna

The Vienna model is presented as a nomogram which is based on a Cox regression model, meaning that the regression coefficients in the LP below represent log hazard ratios.

LP = (0.64 x Male) + (0.96 x PE) + (0.73 x Proximal DVT) + (0.24 x D-dimer (per doubling))

DASH

The DASH model is a score system based on the optimism corrected coefficients of a Cox regression model, with the following linear predictor;

LP = (0.96 x D-dimer (abnormal)) + (0.43 x age ≤ 50 years) + (0.58 x Male) + (-1.05 x Hormone use in women (at time of initial VTE))

The scores used in the final rule were calculated by doubling the above regression coefficients and then rounding to the nearest integer.

Figure 2.2 - Linear predictors of prognostic models included within the review

2.3.2 Quality assessment and critical appraisal

Population characteristics

The population characteristics of the three study populations were broadly similar across predictors measured in all studies (see Table 2.1). The median age of patients in the DASH population was somewhat higher than that of the HERDOO2 and Vienna study populations, and the Vienna study included longer follow-up compared to the other studies, both of which could affect estimates of predictor effects in the models.

Patient selection & outcomes

All of the three articles developed models based on data collected using a prospective design (see Table 2.2), which is ideal for prognostic modelling as predictor information can be collected blind to patient outcome. Across all three articles, recurrent VTE (at various predicted time points) was the primary outcome (see Table 2.2), and was objectively confirmed and independently adjudicated. Detection bias was limited in all three articles by pre-specification of outcome definitions, with the same definition and assessment used for all patients (within each study), meaning systematic differences in the determination of outcomes were avoided.

The inclusion/exclusion criteria used in the three articles is summarised in Table 2.3, and common criteria included the exclusion of patients with high-risk thrombophiliac conditions, patients < 18 years old, and patients treated with < 3 months OAC therapy.

Table 2.1 - Summary patient characteristics of included model studies

Model	HERDOO2				Vienna		DASH			
Measurement statistics used	**Mean (SD) or Freq (%)**				**Median (25th, 75th percentiles) or Freq (%)**		**Median or %**			
Patient characteristic	**n**	**Recurrence**	**n**	**No recurrence**	**n**	**All**	**n**	**Recurrence**	**n**	**No recurrence**
Age (years)	91	53.6 (14.8)	555	52.3 (17.9)	929	54 (43, 63)	239	63	1579	61
Male proportion	91	63 (69.2)	555	269 (48.5)	929	562 (60)	239	69.40%	1579	48.60%
Site (Distal DVT) proportion	91	NA	555	NA	929	164 (17.7)	239	NA	1579	NA
Site (Proximal DVT) proportion	91	NA	555	NA	929	327 (35.2)	239	NA	1579	NA
Site (PE) proportion	91	NA	555	NA	929	438 (47.1)	239	NA	1579	NA
BMI (kg/m^2)	91	30.3 (7.6)	555	28.9 (7.1)	909	27.1 (24.4, 30.1)	^	27.2	^	27.2
D-dimer (µg/L)$^{\Phi}$	91	383 (738)	555	294 (314)	832	355 (236, 558)	239	67.7%*	1579	42%*
Factor V Leiden proportion	91	19 (20.9)	554	81 (14.6)	916	224 (24.4)	239	NA	1579	NA
Duration of OAC (months)	91	5 to 7	555	5 to 7	929	6.6 (6.1, 8.0)	239	6.7	1579	6.8
Duration of follow up (months)	18 (1, 47)$^{\#}$				43.3 (14.7, 78.5)		22.4			

*** DASH reported the percentage with abnormal D-dimer, defined as ≥ 500ng/mL**

^ BMI data available for 802 subjects, no reporting of number of subjects by event status

 # Follow-up for HER DOO 2 presented as mean (range)

$^{\Phi}$ D-dimer measured in ng/mL within the DASH article

NA – The information was not provided for these fields. In particular, both the HERDOO2 and DASH studies did not include patients with distal DVT index events a priori. And the DASH study did not provide figures for the proportion of patients with Factor V Leiden, but the percentages of patients with thrombophilia were 23.4% and 20.9% for recurrence and non-recurrence respectively.

Table 2.2 - Study characteristics

Model	HERDOO2	Vienna	DASH
Year of publication	2008	2010	2012
Country	Four countries (Unspecified)	Austria	Austria, Canada, Italy, Switzerland, UK, USA
Study setting	12 tertiary care centres, patients enrolled between October 2001 and March 2006	Recruited from 4 thrombosis centres in Vienna between July 1992 and August 2008	Patient-level meta-analysis of previously published studies (11)
Study design	Prospective cohort study	Prospective cohort study	Individual patient data from 7 prospective studies
Clinical outcome	Recurrent VTE	Recurrent VTE	Recurrent VTE
Key prediction time points (months)	12	12, 60	12, 24, 60
Total sample size	646	929	1818
Events	91	176	239

Table 2.3 - Study inclusion/exclusion criteria

Model	HERDOO2	Vienna	DASH
Inclusion criteria	First unprovoked VTE Received OAC 5-7 months No recurrent VTE on treatment	First unprovoked VTE Age ≥ 18 Received OAC ≥ 3 months	First unprovoked VTE Including thrombophilic blood abnormalities where there were no other VTE risks
Exclusion criteria	Age < 18 Deficiency in antithrombin, protein C or S Presence of lupus anticoagulant Already discontinued OAC Geographically inaccessible to follow-up Not proximal DVT or PE index event	Deficiency in antithrombin, protein C or S Presence of lupus anticoagulant Presence of cancer	Known antiphospholipid antibodies Antithrombin deficiency Not proximal DVT or PE index event

All articles only included patients with a first unprovoked VTE, but definitions of unprovoked varied somewhat (see Table 2.4). The HERDOO2 and DASH models both included patients with hormone intake at time of index event, while the HERDOO2 model also included pregnancy associated VTE at index event within its definition of unprovoked VTE. The DASH model study justifies including hormone intake as unprovoked because some evidence suggests hormone therapy is a weak predictor for VTE recurrence (37, 186). However, evidence suggests that these risk factors are acquired (154), and inclusion of patients outside the unprovoked population might therefore lead to biased conclusions about predictor effects.

Table 2.4 - Unprovoked VTE definition across studies

Model	HERDOO2	Vienna	DASH
Not provoked by:			
Trauma	X	X	X
Surgery	X	X	X
Cancer	X	X	X
Pregnancy	-	X	X
Immobility	X	-	X
Hormone intake	-	X	-

Predictors

The three studies investigated a wide variety of candidate predictors, including clinical and laboratory predictors. There was some overlap between models (see Table 2.5), with D-dimer, age and sex being the most commonly included predictors. The Vienna model avoided the categorisation of continuous candidate predictors, while the DASH model investigated patient age in pre-specified quartiles, to allow for non-linear associations between age and recurrence risk. The HERDOO2 model in contrast performed chi-squared testing to identify the optimal threshold to dichotomise every continuous predictor under consideration.

Data-driven analyses are known to incite reporting biases, where optimal thresholds are reported without any clinical meaning (187). Dichotomisation of continuous predictors is also methodologically poor, as it seeks to separate patients risk into two categories, treating all those above the threshold as having the same constant risk (and similarly for those below the threshold), which is unrealistic in practice (187).

All models also allowed for site of index event in some way. Both the HERDOO2 and DASH models excluded patients with distal DVT index events from their studies (37, 39), which is important risk stratification in itself (i.e. both models are not applicable to patients with distal DVT events). Only the Vienna model included such patients and adjusted for site of index event as a predictor in the model (see Table 2.5). The Vienna models predicted risks reflect the low risk of recurrence associated with distal DVT index events, and provides an estimate of risk (where the other models do not) which may be a helpful tool in consultation with patients and confirm treatment decisions (11).

Table 2.5 – Predictors included in final model

Model	HERDOO2	Vienna	DASH
Predictors included:			
D-dimer	X	X	X
Age	X	-	X
Sex	-	X	X
BMI	X	-	-
Post thrombotic signs	X	-	-
Site of index event	-	X	-
Hormone therapy	-	-	X

Sample size

The HERDOO2 model was markedly underpowered. The study collected information on 69 predictors and considered at least 36 of these; however, there were only 91 recurrent events (see Table 2.2), meaning around 2.5 events per predictor (EPP) assuming all patients had complete data for candidate predictors. Previous evidence has suggested that an EPP < 10 can lead to bias in estimates of predictor effects and their standard errors, as well as the coverage of confidence intervals, with EPP=2 showing severe biases (188, 189). This may then cause over-fitted models (i.e. models that include inappropriate predictors or predictor effects that are too large) and subsequently misleading predictions. The Vienna and DASH models investigated 15 and 14 candidate predictors respectively, with 176 and 239 total events respectively (see Table 2.2). Following the same rule of thumb (EPP < 10) (188, 189), and assuming complete predictor availability for all patients (i.e. no missing data), the Vienna (just) and DASH models therefore had sufficient numbers of events to assess the predictors of interest with appropriate statistical power (see Figure 2.3). Recent evidence has questioned the EPP rule described above, highlighting deficiencies in the original studies and suggesting that there is no one-size fits all rule when it comes to sample size requirements for prediction model development (190). Further research is needed in this area.

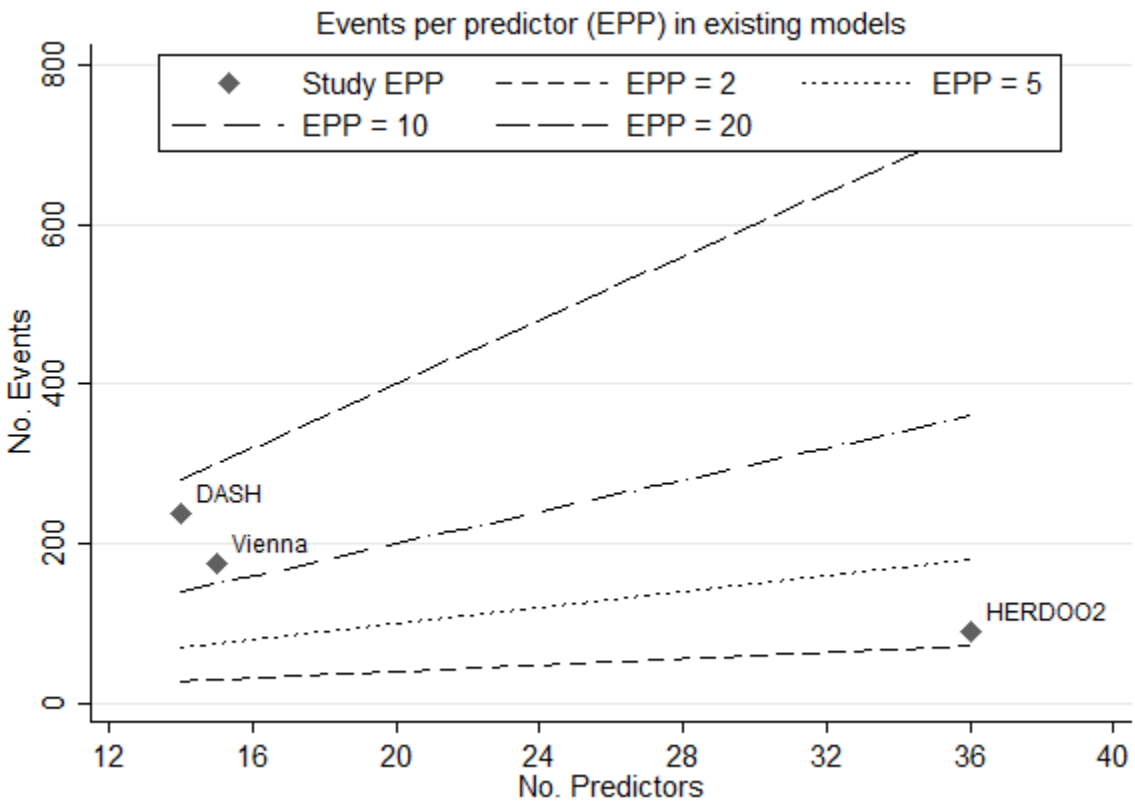


Figure 2.3 - Events per predictor (EPP) for included studies, based on total sample size and number of predictors. NB: lines represent number of events required to maintain EPP=x for given number of predictors.

Missing data

All of the three included studies suffered from some degree of missing predictor information, and used a complete case analysis to overcome this issue. The presence of missing predictor data would further reduce the apparent EPP discussed above (see Figure 2.3). Each study also used a selection procedure meaning more predictors were considered, resulting in a higher proportion of missing predictor data. For example the Vienna prediction model considered peak thrombin as a predictor, for which 300 out of 929 patients had missing predictor information (40). Similarly, the DASH model considered predictors including BMI, for which

only 802 out of 1818 patients had complete predictor information (37). This means that the predictor selection process included a massively reduced sample size compared to the data used toward the final model, which may have increased the chance of spurious predictor-outcome relationships (see Figure 2.4).

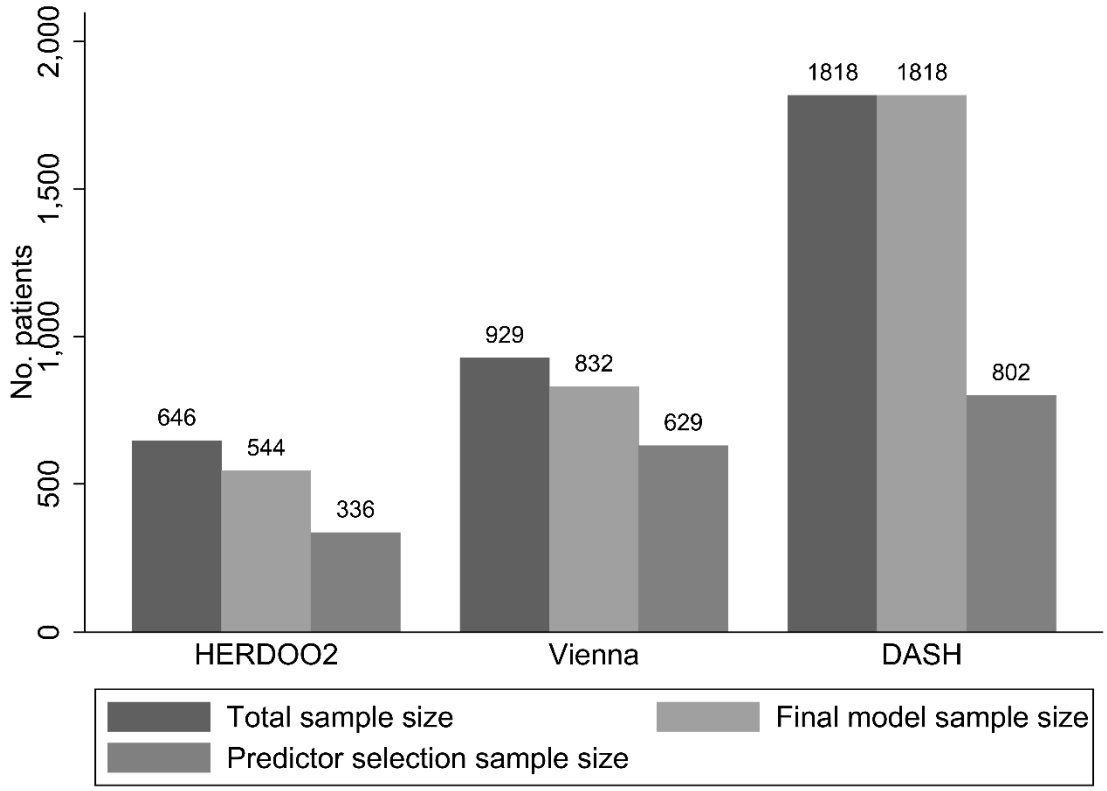


Figure 2.4 - Final model sample size compared to total & selection sample size. Final model sample size=total sample minus patients with missing information in any predictor included in the final model; Predictor selection sample size=total sample size minus patients with missing predictor information in any predictor considered for inclusion in the model using a selection procedure.

No methods to assess the impact of this missing predictor information were used, and in the Vienna and DASH models the number of missing recurrent events was not reported, so no

assessment of the impact on statistical power (nor EPP) could be made accurately. A complete case analysis in the presence of missing data may not represent the entire population, and reduces sample size giving lower power and making predictor effects only pertinent to a specific subgroup of the population with no missing data. Multiple imputation (MI) can be used as a sensitivity analysis to assess the impact of missing data on the performance of the model. MI using chained equations imputes missing predictor information from a posterior distribution based on the observed data (57); it increases sample size, power, and may improve the generalisability of the model. However, this was not used in the development of any of the articles.

Statistical analysis

A Cox proportional hazards model was used to develop both the Vienna and DASH models, which appropriately accounts for censoring of patients in the analysis of time-to-event outcomes such as recurrence (41). The HERDOO2 model used a conditional logistic model for analyses, which does not account for the censoring of patients over time and the variable lengths of individual follow-up (34).

All studies recruited patients from different centres or countries (see Table 2.2), however only one (DASH model) stratified by source in their analyses. Stratification accounts for heterogeneity in the baseline recurrence risk in different patient groups. Ignoring the clustering of patients within centres or countries could lead to poor model calibration (where model predictions do not closely fit observed recurrence rates) and/or biased predictor effects (16, 17, 20), and thus could diminish performance in a new setting. The DASH model did not propose how to implement the model in practice; where models are stratified there a several

options for implementation and so it is important to be clear which one is to be used in practice (19, 20, 34) (e.g. use a single intercept related to one of the centres).

The HERDOO2 model excluded predictors from multivariable analysis where univariable analysis yielded p-values ≥ 0.2; this predictor selection strategy was therefore completely data driven, which could lead to potential bias in results, with predictor effects that may be important in combination being excluded. The DASH model also excluded some predictors from multivariable analysis on the basis of univariable results. Univariable analyses are not recommended for decisions about inclusion of predictors in a multivariable model (49).

Both the Vienna and DASH models used bootstrapping and shrinkage methods to adjust predictor coefficients for optimism (see chapter 1), but the HERDOO2 development did not account for optimism in predictor estimates, despite the extremely low EPP. The use of optimism correction methods reduces over-fitting by reducing the magnitude of predictor effects, to help ensure the model performance is more accurate in a new patient population (33, 38).

The specification and application of the proposed models was described in various ways across the studies. Both the Vienna and DASH models were presented well, with an indication of how the predictors are combined to calculate a patient's recurrence risk at a specific time point. Both provided cumulative recurrence risk at specific time points after cessation of therapy including an estimate of the uncertainty surrounding these estimates (95% confidence intervals). This information could be used to direct the decision making process, informing clinicians and patients of the individual's level of risk, and therefore allowing individualised

treatment strategies. However, neither reported any estimate or parameterisation of the baseline hazard or survival function, which would be required for full external validation of the model in a new setting (144, 145, 147).

Conversely the HERDOO2 model derived a clinical decision rule splitting patients into those with less than two predictors (from their model), and those with greater than two predictors, suggesting that one group should continue OAC therapy, while the other could safely stop. The HERDOO2 model does not allow fully individualised risk estimates at specific time points; it only indicates that fewer than two predictors would indicate a < 3% annual risk of recurrence. This therefore does not allow clinicians or patients to make decisions based upon their preference of recurrence risk threshold, limiting the applicability of the decision rule if a value other than 3% was of interest.

Model validation

Model performance was evaluated using internal validation in all the studies, but none reported an external validation (37, 39, 40). Internal validation was reported in terms of both calibration and discrimination with both the Vienna and DASH models presenting both (though not for the simplified Vienna nomogram, which constitutes the final model). The HERDOO2 model presented neither calibration nor discrimination statistics. The performance statistics reported are given in Table 2.6. Apparent C-statistics (which represent the discriminatory performance within the development data without adjustment for optimism) are between 0.65 and 0.72 for the different models, indicating moderate discrimination ability; however apparent performance is likely to be optimistic (34). The Vienna model also presented a bootstrap adjusted C-statistic (which accounts for optimism) of 0.646 for

predictions at five years post cessation of therapy, indicating a small reduction after accounting for optimism. The Vienna and DASH models also provided a bootstrap optimism-adjusted calibration slope (or uniform shrinkage factor), which showed moderate calibration performance of 0.88 for the Vienna model, and strong performance of 0.97 for the DASH model (with 1 indicating perfect calibration). Both the Vienna and DASH models used their shrinkage factor to adjust the predictor effect values in their final model, to adjust for the optimism.

External validation is the true indication of model performance in the wider population, as a model validated within its development dataset will give optimistic performance statistics (9, 28, 147). External validation studies are currently being undertaken to validate both the HERDOO2 model (169-171) and the Vienna model (168), which will provide a more robust indication as to the overall performance (in terms of calibration and discrimination) of these models in new patient populations where they are intended for use. The REVERSE II study is a randomised trial aiming to compare the impact (on patient outcomes) of using the HERDOO2 model to decide on cessation of OAC therapy, compared to standard practice (169, 171). The VISTA study is a randomised trial comparing the use of the Vienna model to decide on treatment duration, with usual care where treatment duration is based on physicians judgement (168).

Table 2.6 - Internal validation performance statistics

Model	Calibration Slope*	Apparent discrimination^	Bootstrap adjusted discrimination#
HERDOO2			
Model for use (Score)	-	-	-
Development model (Beta terms)	-	-	-
Vienna			
Model for use (Nomogram)	-	-	-
Development model (Beta terms)	0.88	0.651	60 month = 0.646
DASH			
Model for use (Score)	-	0.71	-
Development model (Beta terms)	0.974	0.72	-

*** Bootstrap calibration slope**

^ C-statistic based on development data

 # C-statistic based on bootstrap internal validation

2.3.3 Update to the Vienna prediction model

The authors of the Vienna model also recently developed an update to the original Vienna model, with the aim of predicting recurrence risk at later time points using updated D-dimer measurements (184). New D-dimer measurements were taken at 3, 9 and 15 months post cessation of therapy, with analyses showing a slight decrease in hazard ratios for the effect of log D-dimer over time (though the 95% confidence intervals remained similar) (184). Three new nomograms were developed for use in practice to predict recurrence risk at 12 and 60 months from time of new D-dimer measurement. A web based calculator was also developed by the authors allowing prediction of recurrence risk at any integer month after baseline (3 weeks) and up to 15 months post cessation of therapy.

The updated model was adjusted for optimism using leave-one-out resampling (cross-validation) to calculate shrinkage factors for 3, 9 and 15 months of 0.79, 0.81 and 0.7, indicating moderate calibration of the model but reduced performance compared to the original Vienna model (optimism-adjusted calibration slope = 0.88). Discriminatory performance for 5 year predictions at each new time point showed a small reduction in performance compared to the original model (optimism adjusted AUC values were, 0.61, 0.61 and 0.58, for 3, 9 and 15 months, compared to AUC=0.646 for the original model) (40, 184). The updated Vienna model expands the earlier model by allowing dynamic prediction of recurrence risk over time, but while the earlier Vienna model has recently been externally validated (160), this model has not been externally validated to date, and shows inferior internal validation performance statistics compared to the original model.

2.3.4 Relevant articles identified outside of review search dates

Subsequent to the completion of the initial review searches (1950 – July 2014), two additional highly relevant studies were identified (160, 161). The first was an external validation of the Vienna prediction model using IPD from five studies, which aimed to assess the performance of the Vienna model in terms of both discrimination and calibration in a new population (15, 160).

The study reported that the derivation and validation populations were homogeneous after removal of patients with provoked VTE and those with missing predictor information (160). Discrimination was calculated using the C-statistic for comparison to the original Vienna model, with a C-statistic in the validation cohort of 0.626 compared to 0.646 (the optimism

adjusted discrimination – see Table 2.6) for the derivation data, indicating a reduction in the discriminatory performance of the model in a new setting.

The true calibration of the model in the validation data could not be assessed without the baseline hazard function (76, 144, 145, 147). As the original Vienna model was developed using a Cox model which does not parameterise the baseline hazard function, this meant that assumptions about the shape of the baseline hazard function had to be made (144, 145, 147, 160). The authors recalibrated the Vienna model assuming a Weibull distribution; however, this is akin to a new model development or updating of the model because the baseline hazard is refitted in the new data. Therefore this new model (including new baseline hazard) would itself require further external validation (144, 145, 147). As the authors could not use the Cox model directly to predict survival probabilities (due to the lack of baseline hazard function), they could only assess calibration using risk groups defined by the prognostic index alone, to make predictions for groups of patients within the validation data (76, 147). Comparison of observed and expected survival probabilities in five risk groups showed a general trend for the Vienna model to under predict the risk of VTE recurrence at 12 months post-cessation of therapy (160). It should be noted that the study did not validate the simplified Vienna nomogram proposed for use in practice (147, 160).

The second study identified after the initial review period was an external validation of the updated Vienna model (see section 2.3.3) in a prospective multicentre cohort study (161). The study aimed to validate the updated model in elderly patients over 65 years old, and assessed the models performance in terms of discrimination and the proportion of recurrent events between high and low risk patients defined by the model. The study found no difference

between the proportion of recurrences in the low vs. high risk groups (where recurrence risk <6.2% 12 month was defined as low risk). Discriminative performance was poor at both 12 and 24 months, with C-statistics of 0.39 (95% CI 0.25, 0.52) and 0.43 (95% CI 0.31, 0.54) respectively.

The study suffered from a very low number of events, 17 and 26 by 12 and 24 months respectively. Therefore the conclusions of the study should be interpreted with caution, as it is known that small validation samples tend to show poor calibration and discrimination performance, with current recommendations indicating that validation sample sizes should be a minimum of 100 events and preferably ≥ 200 events (76, 191). Also there were several distinct differences between the derivation patient population for the updated Vienna model and the validation sample used by Tritschler et al. which naturally led to heterogeneity in model performance. In particular the validation study used a much older population (median (IQR) age 74 (69, 79.8) versus 54 (43, 63) in the derivation population). This also led to differences in D-dimer levels, with the elderly patients in the validation study having much higher D-dimer levels (median (IQR) D-dimer 1022 (607, 1755) versus 355 (236, 558) in the derivation population). Further to this, women in the validation study appeared to have much greater risk of recurrence than men, which is discordant with current evidence suggesting that men are between 1.5 to 2 times more likely to suffer a recurrence (37, 39, 40, 192, 193). These differences in baseline characteristics may mean that the predictor effects in the updated Vienna model were miscalibrated when applied in this new population, leading to the poor performance seen in the validation study. However, performance of the Vienna model in populations similar to the originally intended population may show adequate performance

and so further validation is needed in external populations with more similar case-mix to assess the models generalisability (28).

2.3.5 Quality assessment and risk of bias summary of HERDOO2, Vienna and DASH models

Quality assessment based on the early version of the PROBAST tool showed that there was evidence throughout the included studies of a moderate to high risk of bias (see Table 2.7), predominately because of a lack of external validation (see Table 2.6 and Table 2.7). The HERDOO2 model development suffered high risk of bias, due to some markedly poor methodological choices, including the choice of analysis model, substantially underpowered analyses, data-driven categorisation of predictors, lack of adjustment for optimism, and poor presentation of the model for use (see Table 2.7). In contrast, the Vienna prediction model and DASH score were considered generally methodologically sound in terms of their development. Both had statistical power to investigate their candidate predictors, accounted for optimism in their selection procedures, and the Vienna study assessed continuous predictors without categorisation and loss of information (though the DASH study did categorise continuous predictors). Both studies presented their proposed models more clearly than the HERDOO2 model; indicating the recurrence rate associated with predictor values and the uncertainty around those estimates. However, predictions were only provided for particular, discretised values of risk; for example both models provide predictions for only a small selection of time points (Vienna model for 12 and 60 months post therapy, DASH score for 1, 2 and 5 years from cessation of therapy), and only provide 95% confidence intervals for a small selection of predicted annual recurrence rates.

Despite being of generally good methodological quality for development, both Vienna and DASH were classed at moderate risk of bias due to a lack of sufficient external validation (see Table 2.7). The DASH score has received no external validation, and any such future validation should account for the method of implementation, which was not proposed by the authors. The Vienna model has now been externally validated (as discussed above), but issues remain because: (i) validation performance was shown to be lower than expected and uncertainty was high (160); (ii) a new Weibull baseline hazard component was added to the model, which itself requires additional validation (144, 145, 147); (iii) the nomogram version of the model, which is intended for use in practice, was not validated (147); and (iv) validation of the updated dynamic Vienna prediction model in a new population also indicated poor performance. Thus, until further external validation is undertaken and the results of on-going validation studies are available, the true performance in new populations cannot be ascertained.

Table 2.7 - Quality considerations for included studies

Model	HERDOO2	Vienna	DASH
Use of a selection procedure?	Yes	Yes	Yes
Adjustment for optimism in selection procedure?	No	Yes	Yes
Events per predictor > 10?	No	Yes	Yes
Appropriate type of model?	No	Yes	Yes
Modelled continuous predictors as linear/non-linear?	No	Yes	No
Considered multiple imputation to handle missing data?	No	No	No
Adjustment for optimism in internal validation?	Yes	Yes	Yes
Reported discrimination?	No	Yes*	Yes
Reported calibration?	No	Yes*	Yes*
Were final model predictor weightings related to regression coefficients?	Yes	Yes	Yes
Internal validation?	No	Yes*	Yes
External validation?	No	Yes*	No
Overall risk of bias classification?	**High**	**Moderate**	**Moderate**
Key reason for decision	No external validation/ Several quality issues	External validation	No external validation

*** Not for the nomogram/score used in practice**

2.4 Discussion

This systematic review of prognostic models for VTE recurrence risk, identified four full text articles developing three independent prognostic models (37, 39, 40, 184). A critique of the included studies described and identified the strengths and weaknesses of the studies with a particular focus on methods of patient selection, outcome reporting, predictor selection, sample size, model development and validation.

Firstly, a key finding was the different definitions of unprovoked VTE across the included studies (see Table 2.4). The Vienna model study excluded patients with index events provoked

by use of female hormones, such as the oral contraceptive pill or hormone replacement therapy (HRT), while the HERDOO2 and DASH studies defined index events related to hormone use as unprovoked. Risk factors consistently defined as provoking across the studies included; surgery, trauma, immobility and pregnancy (see Table 2.4). The use of varying definitions to describe the unprovoked population creates confusion as to which population the proposed models are applicable to. Further research in developing prognostic models to predict recurrence risk in an unprovoked population should aim to use a standard, consistent definition for the population, excluding patients with acquired/removable risk factors (154), to ensure that model predictions are reliable for intended patients. Given the definition of unprovoked VTE used in the DASH and HERDOO2 studies, the proposed models may not be applicable within an unprovoked population (153, 154).

Across the included studies various predictors were included within the proposed final models, with sex, site of index event and D-dimer level (post therapy) being included consistently within all three models, indicating strong evidence of an association with recurrence risk (see Table 2.5). As such any future model development should consider including these predictors, as they appear prognostic for recurrence risk, and thus evaluate new predictors in addition to these. Indeed the discrimination performance shown in current models was moderate at best and therefore any new model would ideally include additional predictors to improve the discriminatory performance statistically, though a parsimonious model may better facilitate implementation in practice (11, 27). While it has been discussed that the effect of D-dimer as a predictor may be dependent on the method/assay used, previous research has investigated the link between variability in D-dimer assays and

recurrent VTE, and found that varying assays do not alter the prognostic value of D-dimer in predicting recurrence (15).

After evaluation of the models' development and validation criteria, all models were labelled with at least a moderate risk of bias (see Table 2.7). This was mainly due to a lack of robust external validation, which is essential as prognostic model performance is known to be optimistic when evaluated on the same data used to develop the model (9). The HERDOO2 model development was classed at high risk of bias, as - alongside no external validation - it had methodological concerns, including the choice of analysis model, substantially underpowered analyses, data-driven categorisation of predictors, lack of adjustment for optimism, and the presentation of the model for use (39). The Vienna model and DASH score were methodologically sound, as they had adequate statistical power to investigate their candidate predictors, accounted for optimism in their selection procedures, the Vienna model assessed continuous predictors without categorisation and loss of information, and both presented their proposed models clearly (37, 40). However, until further external validation is performed, the true performance in new populations cannot be ascertained.

The new external validation study for the Vienna model adds important information on the applicability of the model in practice. The study shows that the ability of the model to identify those at high and low risk of recurrence is weaker in a new population outside of the derivation dataset (160). However, it is important for the Vienna model to undergo further validation, because the validation study related to the fitted model (i.e. the prognostic index from the fitted Cox model), and not the nomogram (which potentially used a simplified set of regression coefficients) which was recommended for use. The updated dynamic Vienna

prediction model has now also been externally validated in an elderly population, which showed poor discriminatory performance, but suffered from small validation sample sizes and large variations in case-mix from the original models intended population (161). Therefore, it may be a concern that the current randomised trial to evaluate the impact of using the Vienna model in clinical practice is premature (168); that is, reliable external validation performance should ideally be established first, before examining impact (26, 64, 79).

This is the first systematic review identifying prognostic models for VTE recurrence risk in the unprovoked population, and as such it is a strength of the study that a robust systematic methodology was used, which yielded a large amount of potential research, making it unlikely that any relevant study was not included. An important limitation of this review is that the conclusions and quality classifications for the prognostic models discussed in this article are based on the reporting standards of the original articles. Further, it was not possible to perform a quantitative analysis of the identified articles due to a lack of homogeneity in many areas, including the included predictors, model structure and study populations (23, 71).

In conclusion, currently available models to predict risk of recurrent VTE in an unprovoked population have several limitations. In particular, sufficient external validation has not yet been performed for two of the available models and this review indicates that further validation studies are required before the models are implemented in practice. Even then the impact of the models on clinical decision-making and, crucially, patient outcomes should be evaluated through a randomised trial, ideally, or health economic modelling exercise (9, 26, 64, 79, 98). Any new models should try to build on the existing work, ensure external validation in multiple populations, transparency in reporting of model development as outlined in the

TRIPOD statement (33, 68), and finally improved statistical analyses to ensure model predictions are more robust.

The next chapter in this thesis aims to develop and externally validate a new prognostic model in this field, as part of a project commissioned by the HTA. Given the findings of this review, chapter 3 will aim to build on the previous evidence in terms of both their strengths and weaknesses. The new model development will consider the previous evidence for and against important predictor effects, and the most common definitions for unprovoked VTE diagnosis. In particular chapter 3 will aim to address the shortcomings of previous models in terms of the reporting and application of the model, ensuring that future research can use the full model for prediction, validation and head-to-head comparison with competing models.

CHAPTER 3: DEVELOPMENT OF A PROGNOSTIC MODEL USING META-ANALYSIS METHODS: PREDICTING RISK OF RECURRENT VTE IN THE UNPROVOKED POPULATION

3.1 Introduction

3.1.1 Background

As highlighted in chapter 2, prevention of recurrent VTE is a challenging clinical decision problem, which must balance the risks of recurrent thrombosis if OAC therapy is stopped versus the risks of bleeding associated with continuing treatment. This has been highlighted in recommendations from the 9th ACCP antithrombotic guidelines (156), which particularly highlighted this issue of balancing the risks of recurrence and bleeding in unprovoked population. The guidelines suggested that those suffering an initial unprovoked DVT should be treated with different lengths of anticoagulation therapy dependent upon their bleeding risk (156). Those at low to moderate risk of bleeding are suggested to have extended treatment over three months of therapy, while those at high risk are recommended to have a further three months of therapy beyond this (156).

Previously the emphasis from a clinical perspective has been to identify those patients at sufficiently high risk of recurrence to justify continuing therapy. More recently the emphasis has shifted to identifying those patients at sufficiently low risk of recurrence to justify cessation of therapy. This reflects an appreciation of the importance of risk of recurrence, with recurrent events being fatal in approximately 5% to 9% of patients (40).

The current UK guidelines from the British Committee for Standards in Haematology (BCSH) (157) state that all patients with a proximal DVT or PE should be treated for at least 3 months, in line with ACCP guidance. In terms of extending treatment beyond three months it is stated that therapy should be continued if the risk from recurrence on stopping treatment is greater than the risk from anticoagulant-related bleeding. However, these opposing risks are not easily predicted in an individual. In a patient with an average risk of warfarin-related bleeding the annual risk of recurrent VTE that would favour continued anticoagulant therapy has been estimated to be between 3% and 9% (157).

In terms of identifying those patients who may require longer duration of therapy the BCSH guidelines identify that patients with unprovoked venous thrombosis have an annual risk of recurrence of more than 9% in the first year after stopping treatment (157). As this risk exceeds the risk of warfarin-related bleeding, the BCSH recommend that patients with a first unprovoked or recurrent episode of proximal DVT or PE should be considered for long term anticoagulation (157). The issue is not straightforward, however, as whilst the cohort risk for patients with a history of unprovoked venous thrombosis may be greater than 9% as suggested by the BCSH, individual patients risk of recurrence is highly heterogeneous.

The BCSH guidelines illustrate this through identification of a lower annual risk in patients with a normal D-dimer result after completion of initial warfarin therapy compared to those with an elevated D-dimer (3·5% vs. 9%) (157). D-dimer is a breakdown product of fibrin, the principal constituent of a venous clot, and has been mainly used within the context of diagnosis of VTE. Within the diagnostic context a normal d-dimer result, defined usually by the laboratory, effectively rules out an acute VTE when used in combination with a clinical risk

score (194). Risk of recurrence has also been related to the presence of post-thrombotic syndrome and male sex (193, 195, 196).

As most recurrences are easily preventable using anticoagulant therapy, it is of great importance that patient characteristics associated with risk of recurrence are identified, so that patient therapy can be stratified. Stratification of patients with unprovoked VTE according to their recurrence risk might be achieved on the basis of clinical predictors such as gender, comorbidities, or weight; or by measuring laboratory markers of thrombophilia such as factor V Leiden, the prothrombin 20210A mutation, natural coagulation inhibitor deficiencies, elevated coagulation factors, and antiphospholipid antibodies (39, 40, 170, 193). More recently efforts have been made to utilise global coagulation markers, including D-dimer, as prognostic tools (39, 40).

Prognostic models are useful tools in the area of VTE recurrence because the population is highly heterogeneous and therefore it is useful to have a mechanism to predict individuals' risk rather than arbitrarily categorise patients when deciding upon treatment strategies (39, 40). As discussed in chapter 1 a prognostic model combines multiple predictors to predict the risk of a patient with particular characteristics having an event within a specified time. Individual risk predictions can help to inform clinical and patient decision making with regard to treatment strategies, in this scenario whether or not to extend treatment with oral anticoagulants (OAC) to prevent recurrent VTE.

3.1.2 Aims of this chapter

The systematic review in Chapter 2 highlighted several applicability and methodological issues in existing models (95). There was a lack of consistent and appropriate definitions for a first unprovoked VTE, for example with some studies not considering hormone intake to be a provoking risk factor (37, 39). Several methodological issues were also identified including: mishandling of continuous predictors in analyses, underpowered analyses and poor presentation of final models for use in practice. Some of the existing models identified by the systematic review had not been externally validated (to date), and though internal validation had often been performed, external validation is needed to indicate true performance of a model in practice.

The research in this chapter will build on the findings of the systematic review, with the aim to develop and externally validate a new prognostic model for the prediction of individual recurrence risk following cessation of therapy for a first unprovoked VTE. Individual participant data (IPD) will be utilised from multiple cohort studies in order to develop a new prognostic model based on multiple predictors, and simultaneously to externally validate the developed model. The aim is to provide a final, validated prognostic model which allows individualised recurrence risk prediction, which could be used to inform patient care as part of an evidence based approach.

3.2 Methods

The research that follows was conducted by the PhD candidate, Joie Ensor. The work was directed by statistical and clinical supervisors, and input was received throughout the project

from a wider collaborating group. The aims and methods for data collection, patient inclusion, model development and model validation are now described.

3.2.1 Identifying, obtaining & cleaning IPD

IPD from multiple cohort studies was identified for the project through external collaborators in Canada, who had already produced such a database (which will be referred to as the 'Recurrent VTE' (RVTE) database throughout this chapter). Agreement on the sharing of this IPD was made with the database holder, clearly stating the intended use of the data for this project and agreeing appropriate recognition for those that originally collected the data to be used.

The RVTE database contained seven trials investigating an association between D-dimer, measured after anticoagulation was stopped, and VTE recurrence. It included a total of 1634 patients with a first unprovoked VTE; the median follow-up time post-treatment is 22 months and there are 230 recurrent events post-treatment. The database had key two benefits: (i) the availability of D-dimer values, which clinical members of the project team thought may add considerable predictive value, and (ii) the seven trials in the database allowed internal-external cross-validation (20, 66), a novel way to develop a model whilst also examining its performance in external data, within the framework of an IPD meta-analysis.

3.2.2 Population at baseline and outcome of interest

What defined a relevant population?

For the purposes of this research, and based on the advice of clinical collaborators, an initial VTE was defined as unprovoked where there was no history in the previous three months of any of the following risk factors;

- Major surgery,
- Lower limb trauma,
- Use of combined oral contraceptive pill or hormone replacement therapy (HRT),
- Pregnancy,
- Significant immobility, or
- Active cancer

Unprovoked patients were therefore selected from the RVTE database by excluding those patients with a history of any of the above provoking risk factors within the last three months.

Baseline characteristics of the RVTE database

The population characteristics of the seven trials in the RVTE database at baseline were summarised using means and standard deviations for continuous variables, and using counts and percentages for categorical variables.

Baseline patient characteristics were summarised firstly for the whole database and secondly by individual trial. A summary of the number of recurrent events, total patients, as well as the median and longest follow-up for each of the seven trials was also presented to describe the

recurrent events within the RVTE database. The percentage of missing data within the whole database was also presented by each candidate predictor.

Outcome of interest

The outcome of interest was the recurrence of a VTE following cessation of therapy for a first unprovoked VTE.

3.2.3 Available candidate predictors

Seven candidate predictors were available within the RVTE database and all were considered for inclusion in the prognostic model; these were:

- Age (Years),
- BMI (kg/m^2),
- Gender (Female/ Male),
- Site of index event (Distal DVT/ Proximal DVT/ PE),
- Treatment duration before cessation of therapy (Months),
- D-dimer level post cessation of therapy (ng/mL), and
- Lag time between cessation of therapy and measurement of D-dimer (Days)

All candidate predictors were continuous except for gender and site of index event which were categorical, with gender being dichotomous (Male/Female) and site of index event having three categories: proximal deep vein thrombosis (DVT), distal DVT, and pulmonary embolism (PE). Patient age and BMI were measured at cessation of therapy (15). Ultimately, BMI was not considered, for reasons described below.

3.2.4 Issue of different start-points and the need for two models

Most predictors were available at the cessation of therapy. However, D-dimer was measured after some lag time from cessation of therapy, to allow for the effects of therapy on D-dimer to diminish. The average lag time was around 37 days post therapy within the RVTE database, while the standard lag time is around 30 days post therapy (197). To address this, the project team agreed that two models should be developed: a **pre D-dimer** model (start-point at cessation of therapy) and a **post D-dimer** model (start-point at the time of D-dimer measurement) (see Figure 3.1).

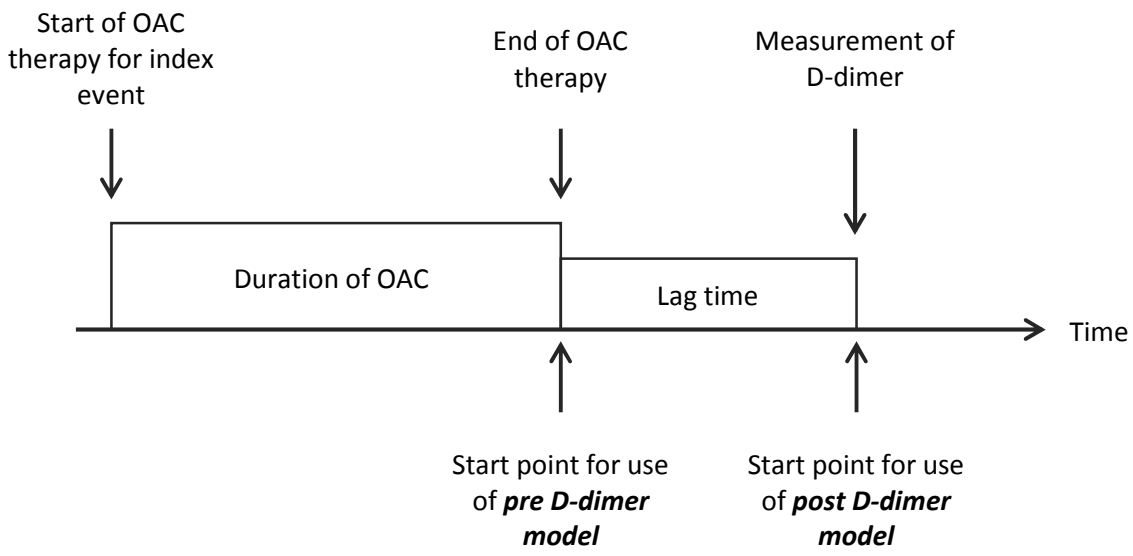


Figure 3.1 – Timeline of patient therapy and start points for pre and post D-dimer use

Post D-dimer model: start-point when D-dimer measured

The primary analysis aimed to utilise D-dimer post cessation of therapy to potentially improve the predictive performance of the prognostic model, as the predictive ability of D-dimer is well documented (15, 40, 198-204). Such a model could be used to inform a decision on extended duration of therapy in patients who have *already* stopped therapy for a given lag time. All seven candidate predictors could be selected for inclusion in the model.

Pre D-dimer model: start-point at cessation of therapy

A secondary analysis was to develop a prognostic model which could be used to predict individual's risk of recurrent VTE at the time of cessation of therapy. As such, candidate predictors for potential inclusion were only age, gender, BMI, site of index event and treatment duration. Such a model could be used to obtain individual recurrence risk predictions at the exact time when cessation of therapy is being considered. These predictions could then inform decisions on whether to continue or stop therapy (alongside other information including the risk of bleeding and patient preference).

3.2.5 Univariable (unadjusted) summary of candidate predictors

The univariable (unadjusted) association between each predictor and recurrence was assessed using a Cox proportional hazards model (41), assuming right censoring as described in chapter 1. A simple linear trend was assumed for continuous predictors during univariable analysis, though more complex trends were investigated during predictor selection for the multivariable model (see section 3.2.6). This assessed the impact of each predictor individually in relation to recurrence. A summary table of the univariable association with recurrence for each candidate predictor was presented, including the estimated hazard ratio with 95% confidence intervals and the corresponding p-value.

3.2.6 Development of prognostic model

Model structure

As the outcome of interest was time-to-event (time to recurrence), prognostic models were developed using a flexible parametric (FP) survival model, fitted using the methods of Royston and Parmar (43, 47). As described in chapter 1, FP models allow firstly, a risk score to be

calculated for an individual patient, which is the combination of parameter estimates (log hazard ratio estimates) from the model combined with the individual patient's predictor values; and secondly, the probability of recurrence by particular time points to be estimated for an individual patient, by utilising the risk score alongside the estimated baseline survival in the population.

The FP models were fitted using maximum likelihood estimation via the stpm2 (205) command in Stata 12.1, with extension to random effects (frailty modelling) as required (206). Model assumptions (e.g. proportional hazards) and model fit were suitably checked throughout (see APPENDIX B: Chapter 3 Appendices).

Modelling baseline hazard

A key part of model development in the Royston Parmar framework is to estimate the baseline hazard. Firstly the spline complexity for the baseline hazard which best fit the available data was investigated visually and through model fit statistics, considering possible degrees of freedom (d.f.) ranging from 1d.f. to 5d.f.

Comparisons were made between models with different d.f. using the Akaike information criteria (AIC) and Bayes information criteria (BIC) statistics, with smaller values preferred. The AIC and BIC provide a measure of how well the model fits the data, whilst penalising models with greater complexity (207). The AIC and BIC are somewhat subjective in isolation and therefore should be compared as a difference relative to the lowest value. As a guideline when comparing models a difference of two or less (in AIC or BIC) would provide strong evidence of an appropriate model fit, differences between four and seven weaker evidence, and

differences greater than ten essentially no evidence of a strong model fit (207). For example when comparing a model with additional predictors to a model without, a difference of less than two in the AIC or BIC would suggest that the additional predictor is not required to improve the model fit.

Accounting for clustering by trial-specific baseline hazards

Recall that the RVTE database contained seven trials, and so accounting for clustering of patients within trials is potentially important. During model development a comparison of the baseline hazards across the trials was carried out. If the shape and magnitude of the baseline hazard was similar between trials a simplistic model would consider using a common baseline hazard for all seven trials. This could be achieved by stacking all seven trial datasets into one large dataset and ignoring the clustering of patients within trials, thereby calculating a single baseline hazard. However, ignoring the clustering of patients within trials is known to create bias in the predictor-outcome associations (16). Therefore, if the baseline hazard between trials did not appear similar, clustering of patients within trials was accounted for by allowing for any between-trial heterogeneity in the baseline hazard across trials (16). This was achieved using FP models with a random-effect on the baseline hazard, thereby producing a weighted mean baseline hazard and an estimate of between-study variability around this mean. This approach thus allows a separate baseline hazard for each trial and estimates the distribution of these (proportional) baselines across trials (20). The average baseline hazard was taken as the baseline hazard to be used in the final model, though it was recognised that large between-study variability in the baseline may affect calibration of the model in some

populations. This would be investigated using internal-external cross-validation (see section 3.2.7), and motivates statistical methodology development in Chapter 4.

Predictor selection & specification

In order to identify a suitable set of predictors to be included in the prognostic model, the multivariable fractional polynomial algorithm (MFP) described by Sauerbrei and Royston (54, 59) was used. The MFP algorithm selects predictors and their transformations as appropriate using a backward selection process; a nominal alpha of 0.15 was chosen to warrant exclusion from the model so as to be more inclusive. Furthermore, patient age was considered to be of clinical and prognostic importance a priori, and thus forced to remain in the model, regardless of its significance. The MFP algorithm allows continuous variables to be modelled appropriately using fractional polynomials for non-linear trends (59), as opposed to being categorized, which has been discussed throughout the literature as suboptimal, for example leading to a loss of power (33, 51, 52, 68, 187).

Also considered for inclusion was a potential interaction effect between the candidate predictors, patient age and D-dimer level, based on clinical judgement which suggested that because D-dimer levels increase with age a compound effect may therefore be plausible (208, 209). Potential time-dependent predictor effects (non-proportional hazards) were also evaluated for the final models (see APPENDIX B: Chapter 3 Appendices).

Handling missing data and exclusion of BMI

Complete case data was used in the development of all models. As a sensitivity analysis, and under the assumption of a missing at random (MAR) mechanism, multiple imputation by

chained equations (MICE) was used to impute missing values of patient level data for the predictors included in the final model (210) (57) (56). Model coefficients were compared to those of the complete case as a sensitivity analysis. To make MAR more plausible all available predictors were included within the imputation model, as well as the outcome (observed recurrences) and the baseline hazard as suggested by White et al. (57). To give an indication as to whether missing data was indeed MAR, summary statistics for population characteristics were compared between complete cases and those with missing information. The reproducibility of the imputation results was examined using the Monte Carlo (MC) error as discussed by White et al. (57).

As a rough guide, the number of imputed datasets should equal the largest proportion of incomplete data observed within individual trial populations (57); in this analysis 48% was the largest proportion of incomplete data (see APPENDIX B1: Summary characteristics of the RVTE database), resulting in 50 imputed datasets being used. In the RVTE database the patient BMI predictor had substantial missingness, and was not recorded for any patient in three of the seven trials. Therefore given the need to recognise the clustering of patients within the same trial, it was not deemed sensible to impute these missing values using the data available from the other trials in which BMI was recorded (211). In particular, when undertaking this project, methods for imputation of such systematically missing predictors was not well developed, although subsequently methods are emerging (123, 124, 127-130). Therefore in the primary analyses for both models (Pre and Post D-Dimer models), utilising all seven trials data, it was decided to exclude BMI as a candidate predictor due to the amount of missing data (see section 3.3.1).

Sample size considerations

When undertaking this work, there was no general consensus about sample size required for prognostic model development. Previous research suggests a general rule of thumb of at least ten events for each candidate predictor considered in a prognostic model (189). For the primary model, there were seven candidate predictor effects (age, gender, site of VTE, D-dimer post-treatment, lag time and treatment duration) for consideration, but some of these were continuous predictors, which could have potentially required non-linear modelling (e.g. fractional polynomials) that would increase the number of predictor effects to be estimated (e.g. if age + age^2 is included, then 'age' relates to two predictors).

The RVTE database included seven trials in total, with 1634 patients with follow-up information post-treatment, and 230 of these had a recurrence; there was good follow-up (median 22 months) and nearly all patients had complete data on all seven candidate predictors. During the internal-external cross-validation procedure (see section 3.2.7 for description) one study is excluded for model development in each cycle, which meant that between 1196 and 1543 patients, and between 161 and 221 recurrences, were available for the development phase of the prognostic models. Thus, there were at least 23 (= 161 events divided by 7 candidate predictors) events for each of the seven candidate predictors, which is considerably greater than the minimum ten per predictor rule, and gave adequate scope for modelling of non-linear trends and clinically plausible interactions as necessary. The events per predictor was lower for the Post D-dimer model (due to missing predictor information), with at least 18 events per predictor in the IECV procedure. Therefore the available sample size for the development of the prognostic model was deemed suitable.

Assumption checks & sensitivity analyses

Continuous predictors were assumed to be normally distributed and this assumption was checked using graphical methods. After inspection of the distribution of candidate predictors a natural log transformation was applied as necessary to achieve approximate normality (prior to the use of the MFP algorithm). Influence of individual data points was assessed by plotting leverage residuals against fitted data. The proportional hazards assumption for each predictor was tested using scaled Schoenfeld residuals plotted against the variable of interest. Plots of Martingale residuals against continuous predictors were used to assess their functional form. Deviance residuals were used to identify outliers. When running the final model sensitivity analyses were performed by excluding any outlying values and checking the robustness (accuracy) of the model to these.

3.2.7 Internal-External Cross-Validation (IECV)

IECV framework

The model development strategy outlined in Section 3.2.6 was implemented within the framework laid out by Debray et al. (20) for developing, implementing and evaluating clinical prediction models using an IPD meta-analysis (IPD from multiple studies). This approach adapts the Internal-External Cross-Validation (IECV) procedure first described by Royston et al. (66) whereby N-1 trials are iteratively selected from the N total trials in the IPD meta-analysis, and the prognostic model is developed within this subset of trials, leaving the remaining trial for external validation of the model (see Figure 3.2). Thus N different models are derived (one for each cycle) and each is subsequently externally validated in the other omitted trial. In this manner, it is possible to investigate (across all permutations of the

excluded trial) whether model performance remains consistent when applied in another trial's population that is not included during model development (external validation). It is important to ensure adequate sample size for model development within each cycle of the IECV, therefore for this study the largest trial (Eichinger et al.) was always included in the model development set of studies.

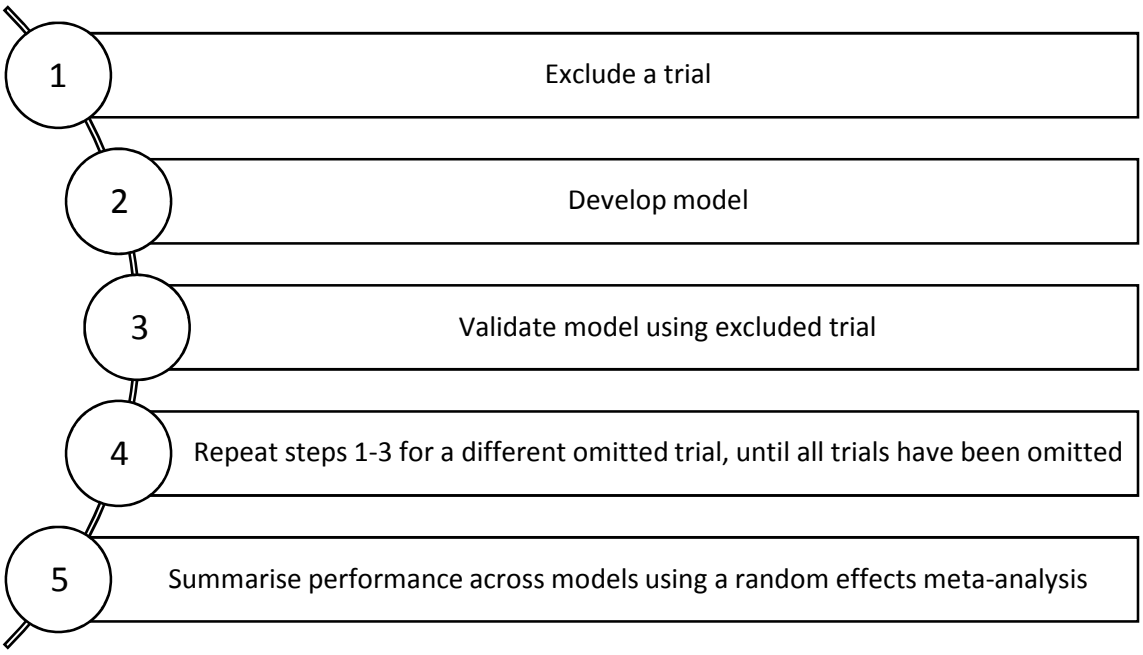


Figure 3.2 - Schematic of Internal-External Cross-Validation (IECV) approach

Validation performance statistics

For each cycle of the IECV approach, an FP model was developed and estimated according to Section 3.2.6, thereby producing a model with an average baseline hazard and risk score equation for the included predictors. The performance of this model was then assessed in the excluded validation trial based on both its discrimination and calibration (72).

The discriminatory ability of the developed model (to distinguish between those who will and those who will not have a recurrence) was examined in the external dataset using Harrell's C statistic (212, 213) with bootstrapping (1000 resamples) used to obtain 95% confidence intervals. Larger C statistics indicate a greater degree of separation in a prognostic models risk score, with a C statistic of 1 showing perfect discrimination and a value of 0.5 showing no discrimination beyond chance.

Calibration of the developed model was assessed by comparing observed (O) and predicted (E) probabilities of recurrence over time, both visually and statistically. To do this, in the external dataset each individual's predicted probability (from the developed model) of recurrence was calculated over time, and the population-average of these predicted survival curves was then plotted against the Kaplan-Meier curve of observed event risk over time in the population. Excellent calibration would be revealed by the Kaplan-Meier and predicted survival curves matching closely. To quantify differences in the curves for each group at a particular time-point, the observed recurrence-free probability (from the Kaplan-Meier curve) and the predicted recurrence-free probability were calculated, and then their difference (E-O) calculated with 95% confidence intervals. A difference of zero would indicate perfect calibration.

Meta-analysis to summarise performance

Development and validation was repeated across all cycles of the IECV, each time excluding a different trial from model development for external validation (see Figure 3.2). Therefore across all cycles, *n* of each validation statistic was obtained (*n* discrimination statistics, and *n* calibration statistics at each time-point, etc.). For each statistic, a random-effects meta-

analysis was undertaken to summarise the performance across all cycles of the IECV (see chapter 1 for a description of random-effects meta-analysis). This analysis weights by the inverse of the variance of each omitted study's estimated statistic plus the estimated between-study heterogeneity in the true statistic value. The model was estimated using the method of moments (DerSimonian and Laird), giving an estimate of the average statistic, the between-study heterogeneity in the statistic, and an approximate 95% prediction interval for the statistic in an external validation population (112, 113). Good prognostic models would have excellent average estimates for each calibration and discrimination statistic, and ideally have little or no heterogeneity in the statistic across different external validation populations.

Production of final model after completion of IECV

If the meta-analysis showed that the performance of the model produced by the IECV approach was consistently good across all cycles (e.g. with calibration close to perfect at all time-points), then a final model was developed using all the trial data combined, with clear guidance for how to use the model to make individual predictions of recurrence risk over time. However, where model performance was not consistently good across each cycle, trials in which the model performed badly in external validation were identified and investigated for any unusual features. Potential differences in case-mix across the trials may lead to poor validation such as different methods of predictor measurement or different treatment strategies (21). Where trials with poor validation were identified a model based on a set of IPD excluding these trials was considered. Similarly, if particular time-points performed poorly in terms of calibration, then the focus was restricted to those time-points which performed well.

Finally, where a suitable model was identified, performance of simpler versions were also examined, to check whether adequate model performance could be achieved with fewer included predictors, so as to ensure the simplest and most easily applicable, yet accurate model for clinical practice was derived (34).

3.2.8 Comparison to existing prognostic models

It was planned to examine the performance of any existing prognostic models or decision rules identified by the systematic review (see Chapter 2) in the RVTE database, if the necessary predictors within these models were available in the database. In particular, head-to-head comparisons of any such models would add important information on which models performed best and could be used for further evaluation for example in impact studies. However such comparison was not possible; firstly because the majority of models identified by the review contained predictors not recorded in the RVTE database, and secondly because the holders of the RVTE database did not permit validation of the Vienna model, as there was an ongoing study to validate this model in the database (160).

3.3 Results

3.3.1 Exploratory analysis of RVTE database

Exploratory analysis identified some extreme values of patient age and BMI predictors (patient ages of zero, and BMI values lower than 10) which were removed from the dataset as erroneous data. D-dimer levels, lag time and treatment duration were all found to have a strong positively skewed empirical distribution, and a log transformation was therefore applied in order to approximate normality (see APPENDIX B1: Summary characteristics of the

RVTE database and APPENDIX B2: Exploratory analysis figures). Patients with treatment durations above 1000 days were removed as this was considered erroneous data based on clinical expertise. These exclusions due to extreme values of predictors led to eight patients being excluded; age equal to zero (no. excluded=1), BMI less than ten (no. excluded=3) and treatment durations greater than 1000 days (no. excluded=4).

As previously mentioned there was a large amount of missing data across the trials for patient BMI, with around 57% of BMI data missing over the whole database, with systematic missing in three trials leading to exclusion of BMI as a candidate predictor. Across the trials there was also some missing data on D-dimer levels and lag time, with 15% and 11.4% missing respectively. No missing data was present for age, gender, treatment duration or site of index event variables.

3.3.2 Pre D-dimer model

The Pre D-dimer model development identified only two important predictors; patient gender and site of index event. Model validation through the IECV approach showed very poor performance of the model in terms of discrimination. Random-effects meta-analysis of C-statistics from each cycle of the IECV, gave a pooled C-statistic of 0.56 (95% CI: 0.51, 0.6). A 95% prediction interval accounting for heterogeneity between trials suggested that the C-statistic for the model used in a new setting could vary anywhere between 0.49 and 0.62, which represents a potentially broad range of performance from discrimination no better than chance to a higher but still rather weak level. The poor performance is potentially to be expected given that the model only includes two predictors, which may not adequately describe variation in the population.

Given the very poor discriminatory performance of the Pre D-dimer model, it was decided that it would not be clinically useful and so the focus of the remainder of this chapter is on the development and validation of the Post D-dimer model. A summary of the validation performance and presentation of the final Pre D-dimer model is provided in the appendices (APPENDIX B5: Pre D-dimer model validation performance and APPENDIX B6: Final pre D-dimer model), with further details published elsewhere (98).

3.3.3 Post D-dimer model: Development and validation

The results of the development and validation of the Post D-Dimer model for prediction of the risk of VTE recurrence are now described below. Candidate predictors available for the post D-dimer model were age, gender, site of index, treatment duration, D-dimer, and lag time (the number of days from cessation of therapy to measurement of D-dimer) (see section 3.2.4). This involved six of the seven trials within the RVTE database; the Baglin trial could not be included because lag time was not recorded. For the IECV approach, due to its large size, the Eichinger trial was always included in the model development set of studies. This ensured a large sample size for model development in each cycle of the IECV. Thus, in summary, there were six studies and seven candidate predictors, with five cycles of the IECV approach conducted.

Complete case data

The complete case data for the development of the post D-dimer was somewhat different to the original RVTE database described above (see APPENDIX B1: Summary characteristics of the RVTE database). Eight patients with unrealistic values were excluded based on the exploratory analysis conducted previously (see section 3.3.1). There was substantial missing

data for both D-dimer and lag time predictors; 243 patients were excluded from the analysis based on missing D-dimer levels, while a further 183 patients were excluded based on missing lag time data (despite having recorded D-dimer levels). The whole of the Baglin (198) trial had to be excluded from the complete case analysis due to missing (unspecified) lag time data. These exclusions led to a reduction in overall sample size to 1200 patients, and a reduction in the number of included events down to 161 recurrent events (see Table 3.1).

Table 3.1 - Summary of baseline characteristics and candidate predictors for the complete-case data used for development of the post D-dimer model

Trial	Palareti 03	Palareti 06	Poli	Tait	Eichinger	Baglin	Shrivastava	All
Recurrences/Total	31/280	23/268	12/81	17/99	69/387	-	9/85	161/1200
Follow-up (months)								
Median	20.8	20.2	19	21.9	28.5	-	26.2	21.6
Longest	31.4	37.2	49	41.6	114.8	-	51.2	114.8
Candidate factors								
Age (Years)*	70.1 (12.3)	65.5 (13.52)	64.5 (14.2)	60.9 (13.8)	54.1 (15)	-	54.9 (12.8)	61.7 (15.2)
BMI (kg/m^2)*	-	-	-	29.1 (6.2)	27.9 (4.8)	-	32.3 (7.2)	28.8 (5.7)
Treatment duration (Months)*	7.5 (6.2)	11.9 (12.3)	13 (11.03)	5.8 (0.9)	8.2 (11.2)	-	8 (5.3)	8.9 (9.9)
D-dimer$^{\#}$ (ng/mL)	500 (310, 1012.5)	540 (330, 855)	247 (201, 356)	544 (306, 991)	345 (230, 551)	-	350 (200, 660)	417.5 (257, 747)
Lag time$^{\#}$ (Days)	28 (24, 33)	31 (29, 35)	30 (30, 30)	26 (22, 35)	21 (16, 27)	-	48 (30, 227)	29 (22, 33)
*Gender^*								
Female	128 (45.7)	99 (36.9)	28 (34.6)	36 (36.4)	146 (37.7)	-	22 (25.9)	459 (38.25)
Male	152 (54.3)	169 (63.1)	53 (65.4)	63 (63.6)	241 (62.3)	-	63 (74.1)	741 (61.75)
*Site of index event^*								
Distal DVT	12 (4.3)	0 (0)	0 (0)	0 (0)	88 (22.7)	-	10 (11.8)	110 (9.2)
Proximal DVT	217 (77.5)	165 (61.6)	57 (70.4)	59 (59.6)	147 (38)	-	57 (67)	702 (58.5)
PE	51 (18.2)	103 (38.4)	24 (29.6)	40 (40.4)	152 (39.3)	-	18 (21.2)	388 (32.3)
Unspecified DVT	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	-	0 (0)	0 (0)

NB: *Mean (standard deviation); $^{\#}$Median (LQ, UQ); ^Count (percentage)

Univariable analysis

Initial univariable analyses were performed by fitting each candidate predictor against recurrence individually using a Cox proportional hazards model, so as to assess the association between each predictor and recurrence (ignoring clustering of patients within trials). Summaries of the univariable association between each predictor and recurrence including the hazard ratio and a 95% confidence interval are presented below (see Table 3.2). These unadjusted results do not consider each predictor's prognostic association after adjustment for other predictors, which is more important for the prognostic model, but provide an initial summary.

Univariable analyses of the predictors considered in the post D-dimer scenario (see Table 3.2) show that unadjusted hazard ratios for patient age and treatment duration are close to 1, with hazard ratios of 1.003 (95% CI: 0.99, 1.01) and 1.199 (95% CI: 0.93, 1.52) respectively. As these are continuous predictors, the hazard ratios compared the change in rate of VTE recurrence for each 1-unit change in the predictor, and so hazard ratios close to 1 may actually have a large impact when multiplied by a large predictor value. However, confidence intervals for both predictors included 1, with large p-values, providing no statistical evidence that the unadjusted recurrence rate was affected by age or duration of treatment. Conversely, the effect of male gender appears to be significantly different from 1 with a hazard ratio of 1.56 (95% CI: 1.12, 2.21) indicating that the unadjusted recurrence rate is around 60% higher for men compared to women. Compared with distal DVT, both proximal DVT and PE have a greater than 5 fold increase in recurrence rate, with hazard ratios of 5.5 (95% CI: 2.02, 15.01) and 5.69 (95% CI: 2.06, 15.74) respectively. While gender and site of index event appear to

have significant prognostic value independently, multivariable analysis should be used to assess whether they retain prognostic value when adjusted for other predictors.

The effect of a patients D-dimer levels appeared to indicate an increase in recurrence rate of around 70% for every 1-unit increase in log D-dimer levels, with a hazard ratio of 1.716 (95% CI: 1.43, 2.06). Conversely the lag time between cessation of therapy and measurement of patient's D-dimer levels appeared to decrease recurrence rate by around 20% for every 1-unit increase in log lag time.

Table 3.2 - Univariable Cox regression analysis of the candidate predictors for the post D-dimer model

Candidate factors	Hazard ratio	Lower 95% CI	Upper 95% CI	P-value
Age	1.003	0.993	1.014	0.513
Treatment duration (months)	1.199	0.926	1.552	0.169
Gender				
Male	1.564	1.108	2.207	0.011
Site of index event				
Proximal DVT	5.498	2.015	15.007	0.001
PE	5.693	2.060	15.736	0.001
D-dimer (Log)	1.716	1.428	2.061	< 0.001
Lag time in days (Log)	0.824	0.627	1.083	0.166

Development of multivariable prognostic model

Baseline spline complexity

In order to consider the complexity (number of knots) required for the baseline spline function a series of preliminary models were fit with varying numbers of knots for the spline function. Comparisons were then made between the models using the AIC and BIC statistics, with smaller values preferred. While simply concerned with the complexity of the model there is no need for variable selection and so a full model is fit, assuming linearity for continuous predictors (46).

Table 3.3 shows the AIC and BIC values for proportional hazards models with between 1 and 5 degrees of freedom (d.f.) for each of the five models fitted (five cycles of the IECV), where the model is built using a derivation dataset based on five trials excluding the trial named in the column header. For simplicity at this stage, the clustering of patients within trials was ignored, and so the set of five trials used in each cycle of the IECV approach was analysed as one large dataset.

Given that lower values of the information criteria represent a better fit, it can be seen that in general across the five derivation datasets a model with 3d.f. minimizes the BIC. The lowest values of AIC vary between 3d.f. to 5d.f. across the derivation datasets, but the unit value of the AIC actually varied very little. The BIC often selects simpler models because increasing numbers of parameters carry greater penalties on the BIC (207), as such increasing the number of d.f. increases the internal knots used and so inflates the number of parameters giving higher BIC values.

The BIC criteria was consistently minimised by a model with 3d.f. except in one instance where the minimum value was very close (difference less than one) to the BIC for a 3d.f. model. The AIC was variable, with the minimum AIC most often occurring for models with 4d.f. however the unit value of the AIC was very close (difference no greater than three) to that for 3d.f. models. Given that 3d.f. minimised the BIC consistently and the minimal AIC values were close to 3d.f. models, as well as considering visually the shapes seen in the baseline hazards for each trial (see Figure 3.3), a complexity of 3d.f. was deemed appropriate for the post D-dimer model. This relates to a baseline spline function with four knots.

Table 3.3 - Comparison of degrees of freedom for baseline spline complexity across derivation datasets for the post D-dimer scenario

		External validation trial name				
	Degrees of freedom	**Cycle 1: Palareti 03**	**Cycle 2: Palareti 06**	**Cycle 3: Poli**	**Cycle 4: Tait**	**Cycle 5: Shrivastava**
AIC	1	974.0	1026.6	1141.8	1110.5	1173.2
	2	969.2	1020.0	1134.6	1106.6	1165.0
	3	964.7	1012.6	1128.9	1097.3	1154.4
	4	963.2	1012.2	1127.6	1094.0	1153.7
	5	962.8	1014.5	1129.4	1095.7	1155.5
BIC	1	999.8	1053.0	1168.8	1137.2	1200.5
	2	997.9	1049.3	1164.6	1136.3	1195.2
	3	996.2	1044.8	1161.9	1130.0	1187.7
	4	997.6	1047.3	1163.6	1129.6	1190.0
	5	1000.0	1052.5	1168.5	1134.3	1194.8

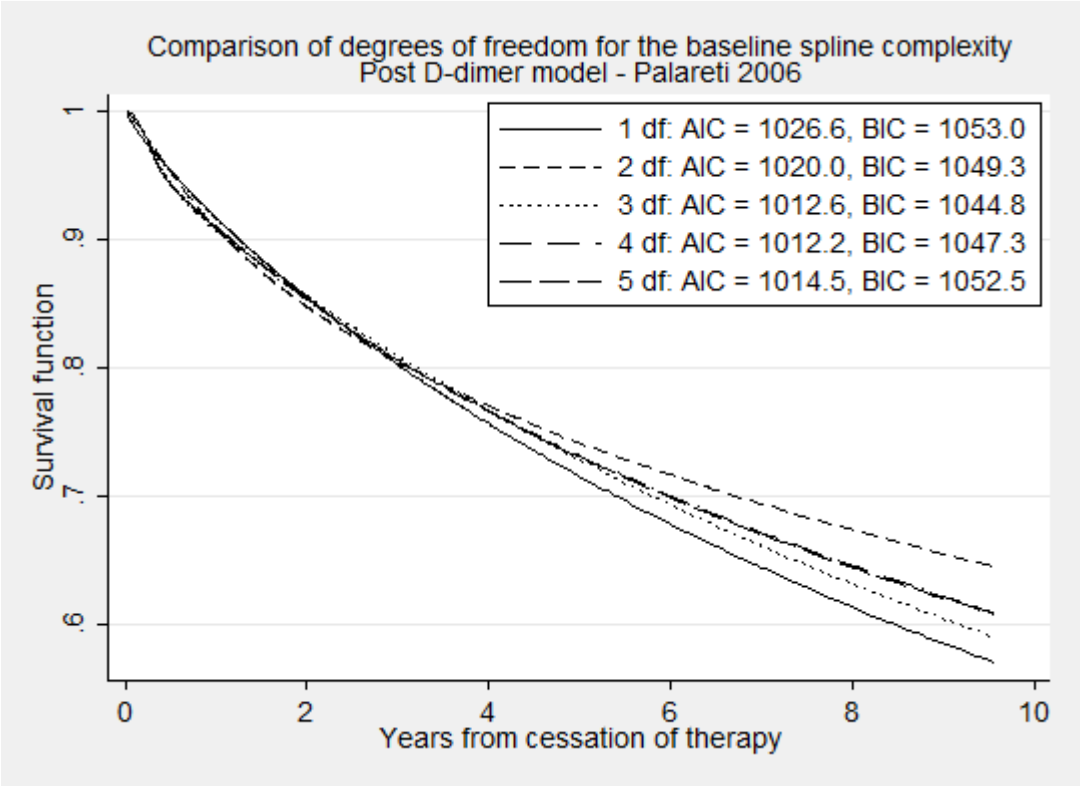


Figure 3.3 - Comparison of baseline spline complexity with differing numbers of internal knots (Example shown for development dataset excluding the Palareti 2006 trial)

Baseline hazard within trials

Investigation of the baseline hazard function using a null model (with no predictors) was also undertaken within each trial in the RVTE database, to ascertain whether the shape and magnitude of the baseline hazard in each trial was noticeably different. Examination of the baseline hazard functions within each trial (see Figure 3.4 and Figure 3.5) showed the shape of the baseline hazard across trials was similar, with a peak in hazard just under 1 year from cessation of therapy, and a fall in hazard thereafter. There was also a rise in the baseline hazard seen in the Poli (214) trial after two years from cessation of therapy, which was not seen in the other trials; however this was considered to be potentially due to the small number of individuals in this trial, as illustrated in Figure 3.5 by the large confidence interval surrounding the tail of the hazard function. While the shape of the baseline hazard across trials appeared to be homogenous, the magnitude of the baseline hazard varied across trials distinctly.

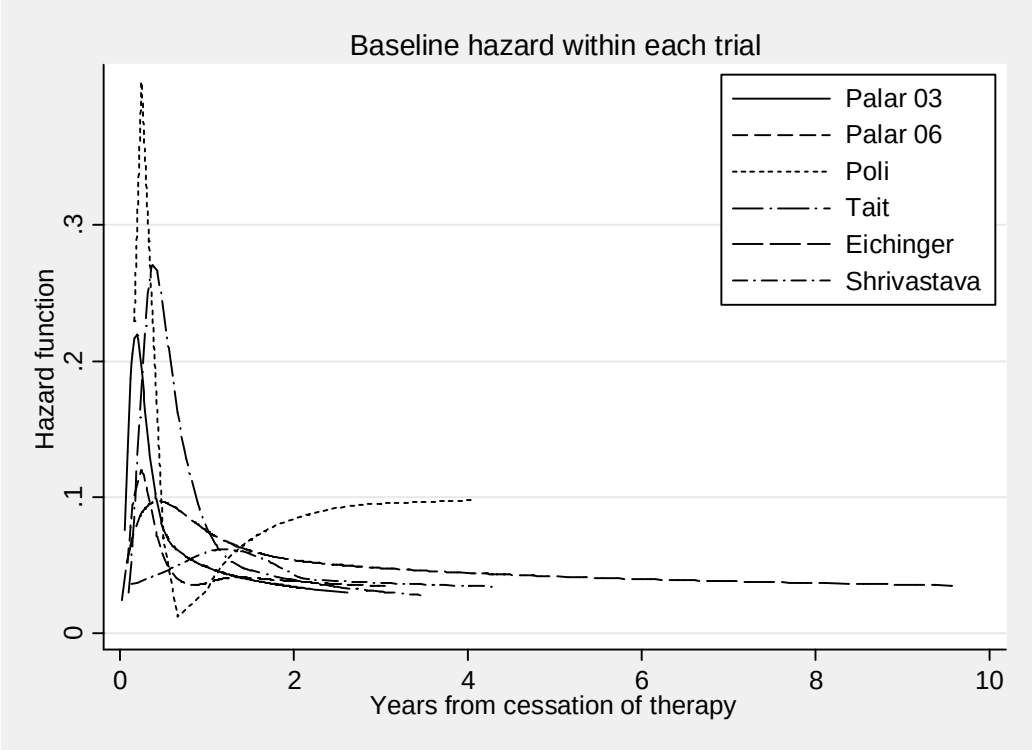


Figure 3.4 - Baseline hazard within each trial for the post D-dimer scenario (null model)

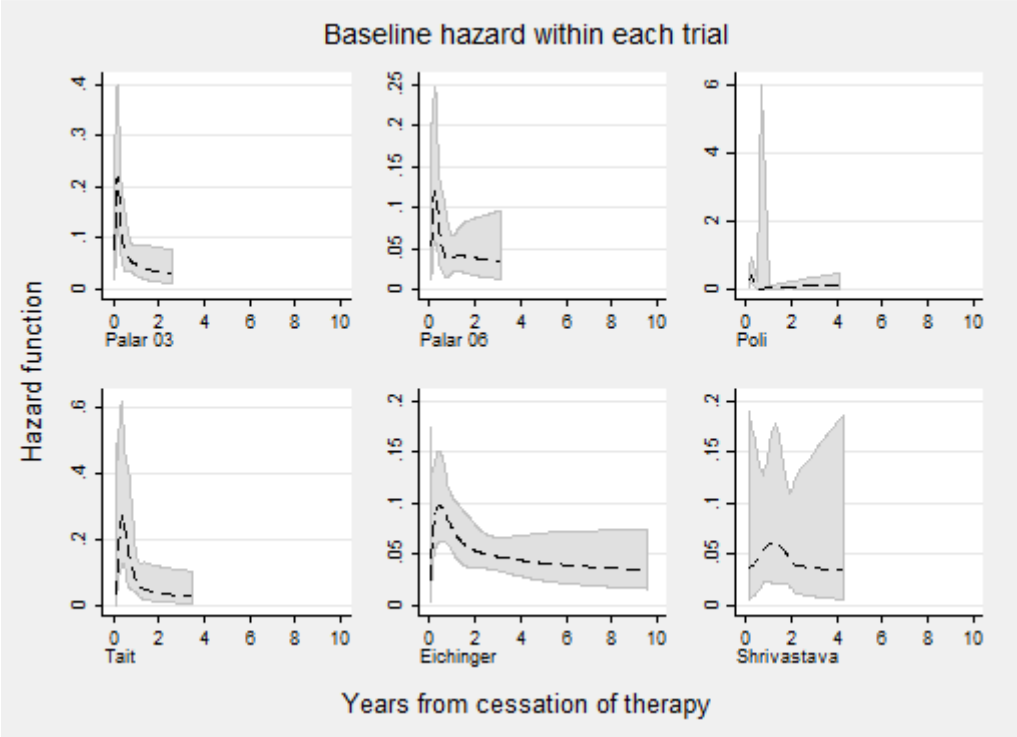


Figure 3.5 - Baseline hazard within each trial with 95% confidence intervals for the post D-dimer scenario (null model)

Given the shape of the baseline hazard appeared similar across trials, but there was variation in the magnitude, it was considered appropriate to develop the post D-dimer model by assuming *proportional* baseline hazards across trials and placing a random-effect on the baseline hazard. This therefore allowed estimation of an average baseline hazard across trials and allowed for variability in each trial's own baseline hazard away from this average.

Selection of predictors & model estimates during IECV cycles

Candidate predictors were entered into the multivariable fractional polynomial (MFP) algorithm of Sauerbrei and Royston (54, 59) (see section 3.2.6).

As previously discussed, given its large number of patients relative to the other trials, the Eichinger (203) trial was forced to remain in the development dataset throughout all cycles of the IECV approach; therefore no model was built without the Eichinger trial population, and subsequently no external validation was performed in the Eichinger trial. The trial was included in all models developed because it was the largest population available and therefore would have a large impact on any final model produced. Thus, although there were six trials available, there were only five cycles of the IECV approach for the post D-dimer model.

The results of predictor selection and parameter estimates at each cycle of the IECV approach are shown in Table 3.4. Treatment duration was not significant during predictor selection and so was excluded from the developed models in all cycles. Age was forced to be included, and the effect of age was estimated in the opposite direction to that estimated in univariable analysis. All other predictor coefficients were estimated to be similar in magnitude to those

seen during univariable analysis for the post D-dimer model, although 95% confidence intervals were altered due to adjustment.

Table 3.4 - Model regression coefficients and selected predictors for each IECV cycle for the post D-dimer model (Hazard ratios (Lower 95% CI, Upper 95% CI))

Candidate predictors*	Hazard ratios (Lower 95% CI, Upper 95% CI)				
	Cycle 1: Palareti 03 excluded	***Cycle 2: Palareti 06 excluded***	***Cycle 3: Poli excluded***	***Cycle 4: Tait excluded***	***Cycle 5: Shrivastava excluded***
Age (years)	-	-	0.99 (0.97, 1.01)	0.99 (0.97, 1.01)	0.99 (0.97, 1.01)
D-dimer (log)	1.90 (1.55, 2.32)	1.80 (1.48, 2.2)	2.05 (1.65, 2.56)	1.99 (1.6, 2.48)	1.95 (1.6, 2.39)
Lag time (log)	0.73 (0.55, 0.98)	-	0.75 (0.55, 1.02)	0.73 (0.53, 0.99)	-
Gender					
Male	1.88 (1.27, 2.77)	1.77 (1.22, 2.56)	1.68 (1.16, 2.44)	1.93 (1.34, 2.8)	1.72 (1.21, 2.44)
Site of index event					
Proximal DVT	5.21 (1.73, 15.64)	5.16 (1.79, 14.88)	5.53 (1.8, 16.95)	5.53 (1.92, 15.96)	5.47 (1.82, 16.44)
PE	5.37 (1.79, 16.12)	5.99 (2.03, 17.64)	5.99 (1.92, 18.73)	5.93 (2.05, 17.12)	5.21 (1.7, 15.96)
Constant	0.01 (0, 0.04)	0.02 (0.01, 0.05)	0.01 (0, 0.04)	0.01 (0, 0.04)	0.02 (0.01, 0.05)

** Treatment duration was not selected for inclusion in any cycle of the IECV*
NB: An empty cell indicates the predictor was not selected for inclusion in the model

Model validation in the IECV cycles

The final step of the IECV approach (see section 3.2.7) is to assess, in each cycle, the developed model's performance within the external validation trial. As the validation trial was excluded from model development the performance of the model within this dataset can be deemed as external validation. Model performance is now assessed in terms of both discrimination and calibration as described previously (see section 3.2.7).

Across all cycles, there were 92 events in the external validation data. Although the number of events in each cycle was considerably less, the results were pooled across cycles, and thus the total effective sample size was 92. Previously research has recommended a minimum of 100 events and 100 non-events to achieve sufficient power for external validation studies (76, 191, 215), and so overall the total events was akin to this, although not within each cycle.

Model performance statistics for the post D-dimer model developed in each cycle of the IECV approach are presented in Table 3.5, and show C-statistics ranging from 0.65 in the Poli (214) trial to 0.80 in the Shrivastava trial (216). Discrimination overall, across all validation trials, showed a pooled C-statistic from a random-effects meta-analysis (see Figure 3.6) of 0.69 (95% CI: 0.63, 0.75), which reveals moderately good discrimination on average. A random-effects meta-analysis was performed as there were expected to be different discriminatory effects within each validation trial, as opposed to one true C-statistic in all trials as assumed under a fixed-effect meta-analysis (21). Importantly the approximate 95% prediction interval for the C-statistic in a new population was 0.59 to 0.79, which suggests potentially large variability in discrimination performance across settings.

Table 3.5 - Summary statistics for discrimination and calibration of the post D-dimer model in each cycle of the IECV approach

	Estimate (95% CI)				
External validation trial	**Cycle 1: Palareti 03**	**Cycle 2: Palareti 06**	**Cycle 3: Poli**	**Cycle 4: Tait**	**Cycle 5: Shrivastava**
Recurrences/ Total patients	31/280	23/268	12/81	17/99	9/85
C-statistic	0.66 (0.55, 0.76)	0.66 (0.53, 0.78)	0.65 (0.48, 0.82)	0.67 (0.52, 0.8)	0.8 (0.68, 0.93)
E-O (6 months)	0.01 (-0.02, 0.04)	-0.03 (-0.05, 0)	0.06 (0, 0.13)	0.01 (-0.04, 0.07)	-0.03 (-0.06, 0)
E-O (1 year)	-0.01 (-0.05, 0.02)	-0.05 (-0.08, -0.02)	0.05 (-0.02, 0.12)	0.02 (-0.05, 0.08)	-0.05 (-0.09, -0.01)
E-O (2 year)	-0.04 (-0.08, 0)	-0.07 (-0.11, -0.03)	0.08 (-0.03, 0.19)	0.02 (-0.06, 0.1)	-0.04 (-0.11, 0.03)
E-O (3 year)	-0.09 (-0.13, -0.05)	-0.13 (-0.17, -0.09)	0.21 (-0.09, 0.51)	-0.03 (-0.12, 0.05)	-0.04 (-0.12, 0.05)

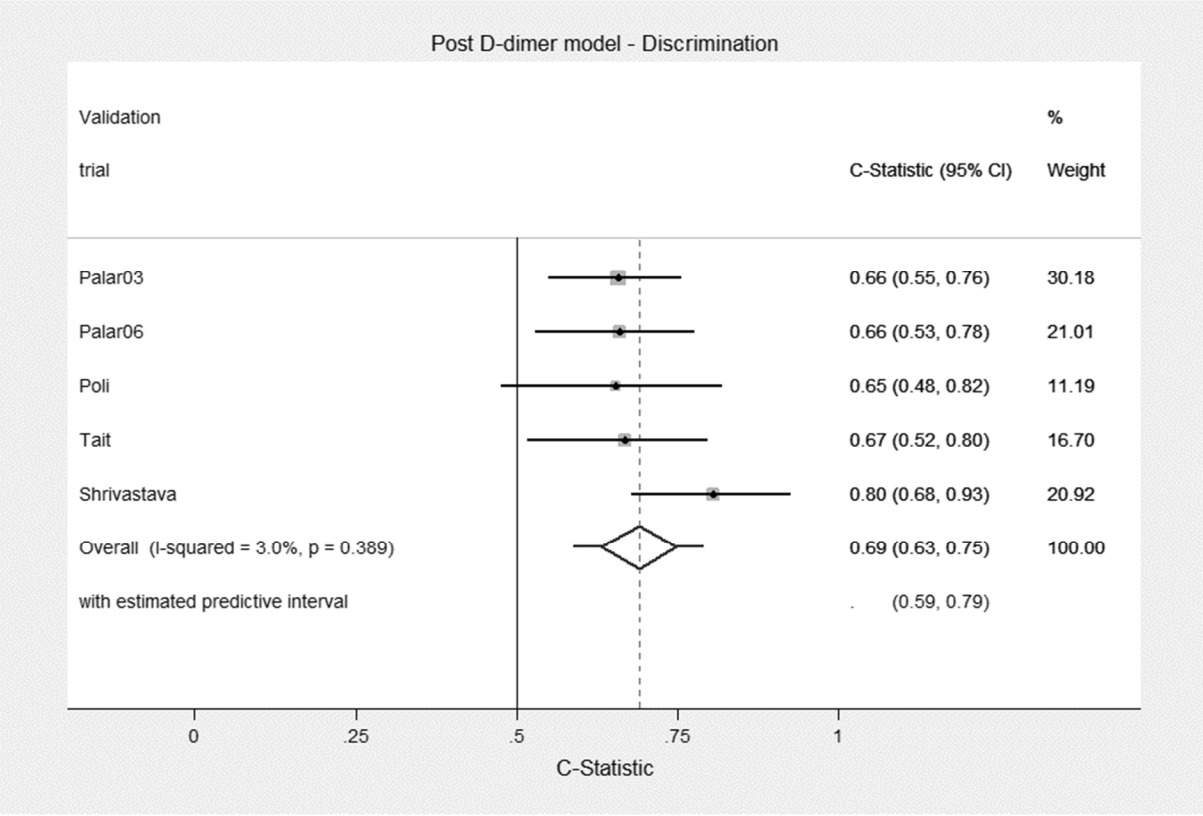


Figure 3.6 - Random-effects meta-analysis of discrimination performance as measured by the C-statistics obtained, for each cycle of the IECV approach for the post D-dimer model

Calibration for the post D-dimer model (see Table 3.5) was consistently strong across all cycles of the IECV up to 2 years post cessation of therapy. The E-O statistics were close to zero for

time-points up to about 2 years, but larger discrepancies were apparent thereafter, for example in the Palareti 2006 (204) and Poli (214) trials. The close relationship between the model's predicted recurrence-free survival probabilities (Expected, E) and the true observed recurrence-free survival probabilities (Observed, O) up to 2 years can be seen for each trial in Figure 3.7. There was no strong evidence of systematic miscalibration in all trials (i.e. any miscalibration could be in both directions, above or below the observed Kaplan-Meier curve).

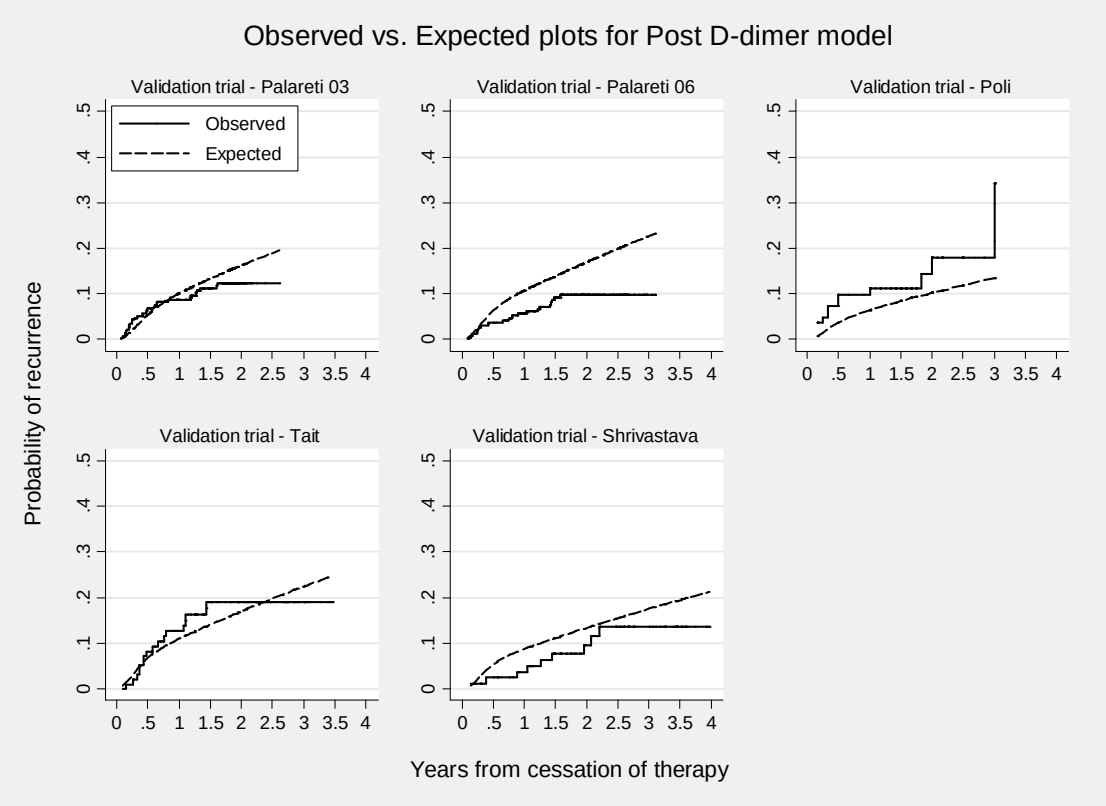


Figure 3.7 - Observed vs. Expected risk within the validation trial for each cycle of the IECV (The post D-dimer model)

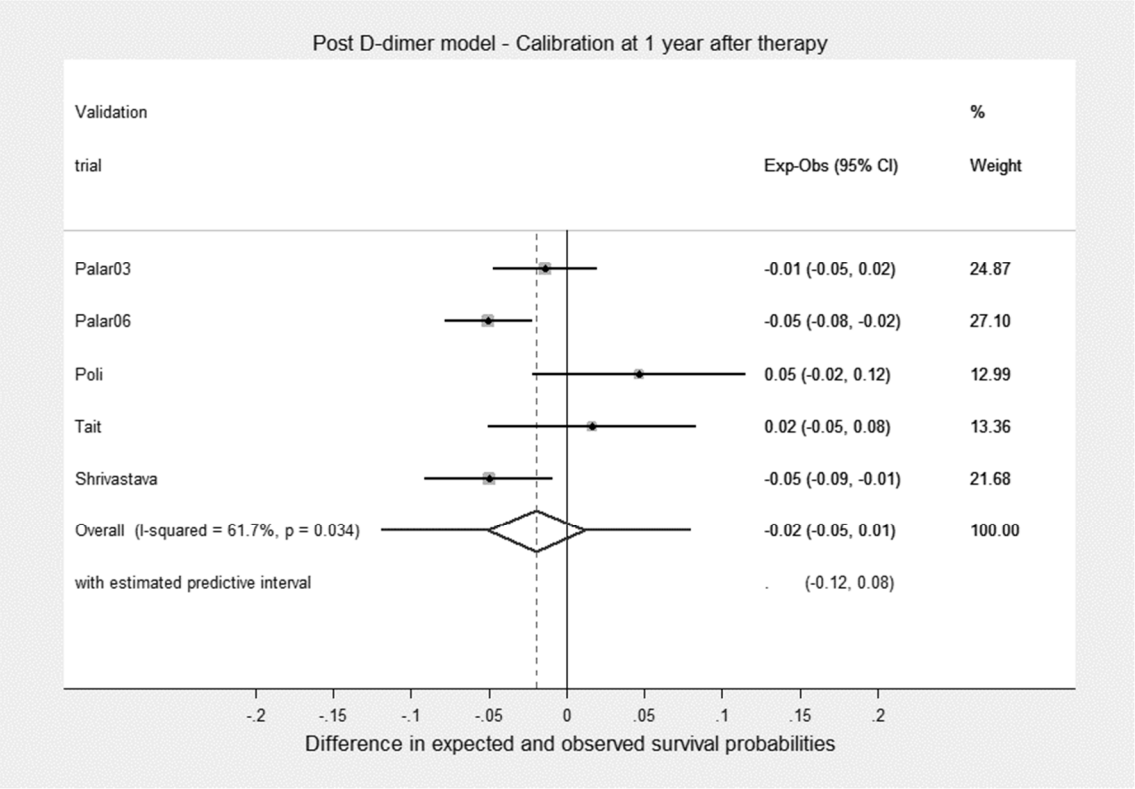


Figure 3.8 - Random-effects meta-analysis of calibration performance (at 1 year post therapy) within validation trials across IECV cycles (The post D-dimer model)

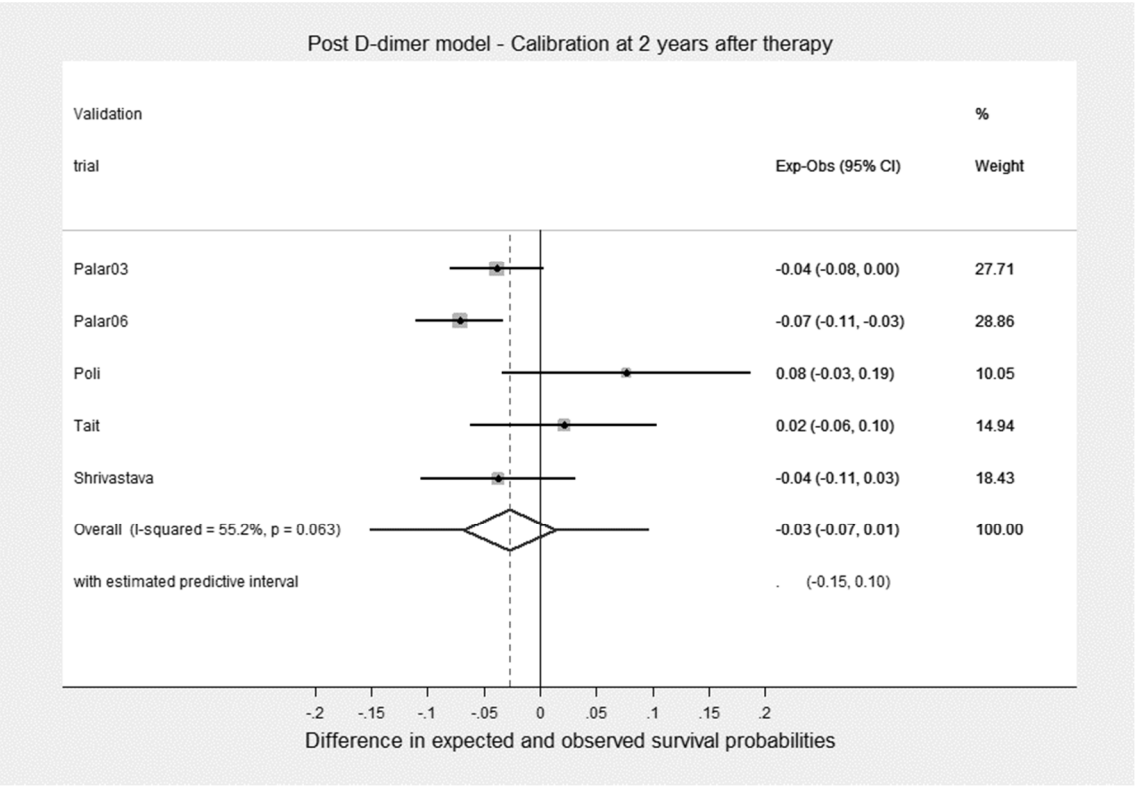


Figure 3.9 - Random-effects meta-analysis of calibration performance (at 2 years post therapy) within validation trials across IECV cycles (The post D-dimer model)

A random-effects meta-analysis of the calibration statistics at one year post cessation of therapy (see Figure 3.8) gave a pooled value of -0.02 (95% CI: -0.05 to 0.01), indicating close agreement on average in the validation trials. There was heterogeneity in calibration performance (I-squared = 61.7%), and the 95% prediction interval for the calibration at one-year in a new population is -0.12 to 0.08. The interval is wide partly due to the observed heterogeneity, but also partly reflecting the uncertainty in the between-study heterogeneity estimate (due to there being only five validation trials). Similar results can be seen for a random-effects meta-analysis of calibration statistics at two years post cessation of therapy (see Figure 3.9) showing consistent agreement on average in the validation trials at two years.

Note that meta-analysis was performed on the E-O scale, rather than E/O scale, to aid interpretation of any miscalibration on the absolute scale (see chapter 1 for a discussion on the interpretation of these statistics).

3.3.4 Final model: Post D-dimer model

The IECV approach identified that the post D-dimer model had a moderately good discrimination shown by an average C-statistic of 0.69 (similar to other published risk prediction models (24)) and good calibration on average across trials, especially up to 2 years, and given this it was deemed appropriate to produce a final D-dimer model based on all the trials combined (see section 3.2.7).

Thus model development proceeded with all six trials included. The specification and parameter estimates of this final post D-dimer model are now described, alongside sensitivity analysis evaluating some aspects of model fit.

Specification and parameter estimates

The final post D-dimer model was developed using the whole trial dataset, with potential candidate predictors including patient age, gender, treatment duration, site of index event, D-dimer and lag time as discussed previously (see section 3.2.4). A random-effect was placed on the baseline hazard to allow for between-trial heterogeneity. The MFP algorithm was used to perform predictor selection, as described previously (see section 3.2.6), with patient age, gender, site of index event, D-dimer and lag time (note the natural logarithm of D-dimer and lag time were used due to skewness as before) being selected for inclusion in the final post D-dimer model. Estimated hazard ratios remained similar to those seen through cycles of the IECV as expected (see Table 3.6). D-dimer was associated with a two-fold increase in recurrence rate for every one unit increase in log ng/mL. Log lag time was associated with a 25% reduction in recurrence rate, which is likely to reflect that healthier patients live longer, therefore the more time that passes before measuring D-dimer, the more likely patients remaining in the trial are healthier and therefore have a lower recurrence rate (see Table 3.6).

Table 3.6 - Specification and estimates of the final post D-dimer model fitted to all trial data

Predictor	Beta coefficient (95% CI)	Hazard ratio (95% CI)	P-value
Age	-0.0105 (-0.022, 0.0011)	0.99 (0.98, 1.001)	0.075
Gender			
Male	0.55 (0.19, 0.89)	1.72 (1.22, 2.44)	0.002
Site of index event			
Proximal DVT	1.74 (0.67, 2.79)	5.67 (1.96, 16.43)	0.001
PE	1.76 (0.68, 2.83)	5.79 (1.98, 16.94)	0.001
D-dimer (log)	0.7 (0.51, 0.89)	2.01 (1.66, 2.45)	<0.001
Lag time (log)	-0.29 (-0.58, 0.002)	0.75 (0.56, 1.002)	0.051

The estimated average baseline survival, $S_0(t)$ from this model is shown below in Figure 3.12 and allows practitioners to estimate the average baseline survival for a specific time point, which can be used to predict recurrence free survival probability using Equation 3.1. Section 3.3.5, below, details how to use the estimated baseline $S_0(t)$ in combination with the estimated predictor effects to make predictions over time for new individuals.

The apparent calibration of the model in the entire dataset was excellent, as expected due to the final model being developed on the same set of data (see Figure 3.10).

A plot of the recurrence probabilities over time in centiles of the distribution of the prognostic index was used (see Figure 3.11), to give an idea of what may happen to individuals at the fringes of the risk spectrum (46). It is clear from Figure 3.11 that while the 50^{th} centile corresponds roughly to the predicted curve seen in Figure 3.10, there is a marked increase in the probability of recurrence for those in the 90^{th} centile of the prognostic index. This separation reflects the good discrimination observed during the IECV approach, where the average C-statistic was 0.69 (see Figure 3.6).

Shrinkage consideration

IECV showed very slight miscalibration on average; for example, at 1-year the summary E-O was -0.02, and at 2 years was -0.03. To address this, some form of shrinkage was considered. However, the research team noted that there was heterogeneity in the direction of miscalibration across studies, and therefore choosing a particular shrinkage factor was problematic. Therefore, and given the miscalibration was only very small on average, no shrinkage was undertaken. However, this issue motivates subsequent methodology research in Chapter 4.

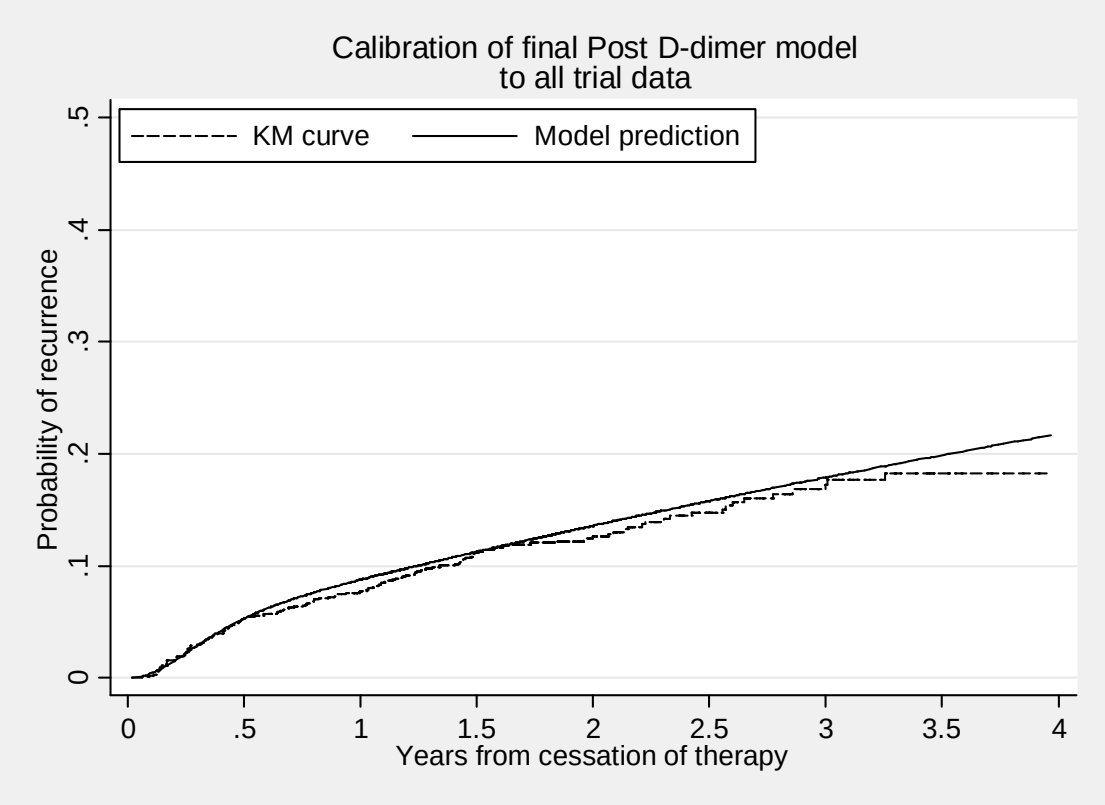


Figure 3.10 – Apparent calibration of the post D-dimer model fit to all trial data

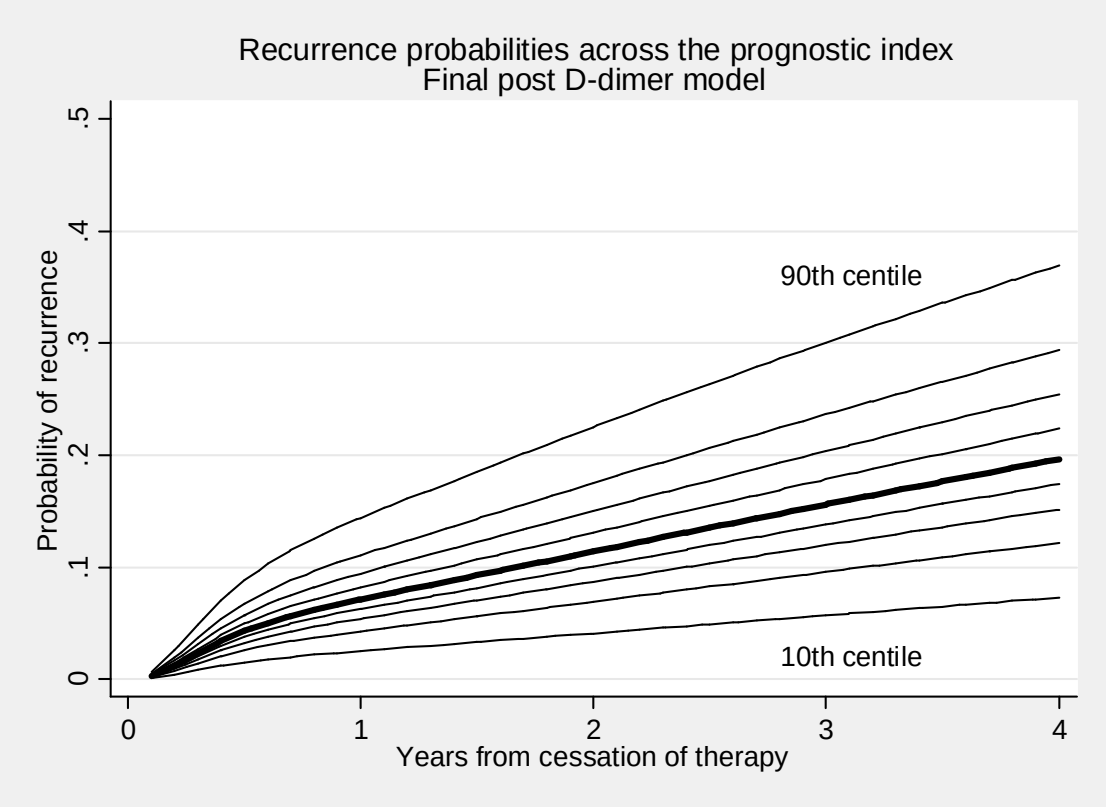


Figure 3.11 - Probability of recurrence across the risk spectrum (The post D-dimer model)

Model checking

During development of the post D-dimer model, a number of assumptions were made and only complete data were used. The robustness of the final model to these assumptions and other issues was investigated as appropriate and laid out in section 3.2.6. Checks for the proportional hazards assumption, outliers, leverage and the functional form of continuous predictors showed no issues for the post D-dimer model (see APPENDIX B4: Model checking results: Post D-dimer model).

Sensitivity analysis

To check the robustness of the complete data model to missing data, a multiple imputation approach was used as described in section 3.2.6. Results gave similar conclusions about the predictors to include and their magnitude of effect (see APPENDIX B3: Sensitivity analysis

results: Post D-dimer model). As the complete data model was already validated during the IECV approach, and it performed well in terms of calibration and discrimination, the Post D-dimer model as derived using complete data was retained as the final model.

Further to this the inclusion of possible interaction effects and time-dependent effects in the model was considered as discussed in section 3.2.6. Given the selected nominal alpha level for these analyses no additional predictors were included in the model (see APPENDIX B3: Sensitivity analysis results: Post D-dimer model).

Summary

Based on the external validation within each IECV cycle, on average across populations the performance of the post D-dimer model is expected to be between moderate and good in terms of discrimination and have good calibration up to at least 2 years post cessation of therapy. Thus including D-dimer and lag time appears beneficial for improved prediction of recurrence risk following cessation of therapy for a first unprovoked VTE. Performance may be improved by the inclusion of further predictors not available in the RVTE database, but given the reasonable pooled discrimination and pooled calibration identified across the IECV external validations, the model appears robust and potentially useful for informing clinical decisions, at least on average. Further research may look to reduce potential heterogeneity of model calibration performance across populations, especially in regard to later time-points.

3.3.5 Using the post D-dimer model to make predictions for new individuals: a detailed illustration of the model in practice

The final post D-dimer model has the potential to stratify the largely heterogeneous population of unprovoked patients, allowing for better decision making on duration of treatment for these high risk patients. This section now explains the practical application of the final post D-dimer model.

In order to predict an individual's risk of recurrence the beta coefficients must be combined with the baseline risk corresponding to the time that prediction is required for (see chapter 1). The equation to combine these parameters is given below (see Equation 3.1), along with the beta values from the post D-dimer model (see Equation 3.2). In Equation 3.1, $S_0(t)$ represents the average baseline (recurrence free survival) risk at time t, and $\boldsymbol{\beta x}$ represents the risk score for a patient as shown in Equation 3.2.

Equation 3.1 – Post D-dimer model equation to predict probability of recurrence free survival at time t

$$S(t) = S_0(t)^{exp(\boldsymbol{\beta x})}$$

Equation 3.2 - Risk score equation for the post D-dimer model

$$\boldsymbol{\beta x} = (-0.0105 \times Age) + (0.545 \times Gender{:}\, Male) + (1.735 \times Site{:}\, Proximal\ DVT) + (1.756 \times Site{:}\, PE) + (0.701 \times \ln D\text{-}dimer) + (-0.291 \times \ln Lag\ time)$$

Equation 3.1 allows the prediction of a recurrence free survival probability at a particular time point after cessation of therapy, meaning that the probability of recurrence by a specific time point, R(t), can also be predicted and is equal to:

Equation 3.3 - Post D-dimer model equation to predict probability of recurrence by time t

$$R(t) = 1 - S(t)$$

The average baseline risk at time t, $S_0(t)$, can be estimated for any time t (post cessation of therapy) by reading off its value from the predicted curve presented in Figure 3.12, and provided for specific time points (six months, one, two and three years) within Table 3.7.

Table 3.7 - Baseline (recurrence free) survival at particular time points to combine with patient specific predictor values for individual risk prediction (Post D-dimer model)

Model predictor	Time from cessation of therapy			
	6 months	1 year	2 years	3 years
$S_0(t)$	0.9996	0.9993	0.9988	0.9983

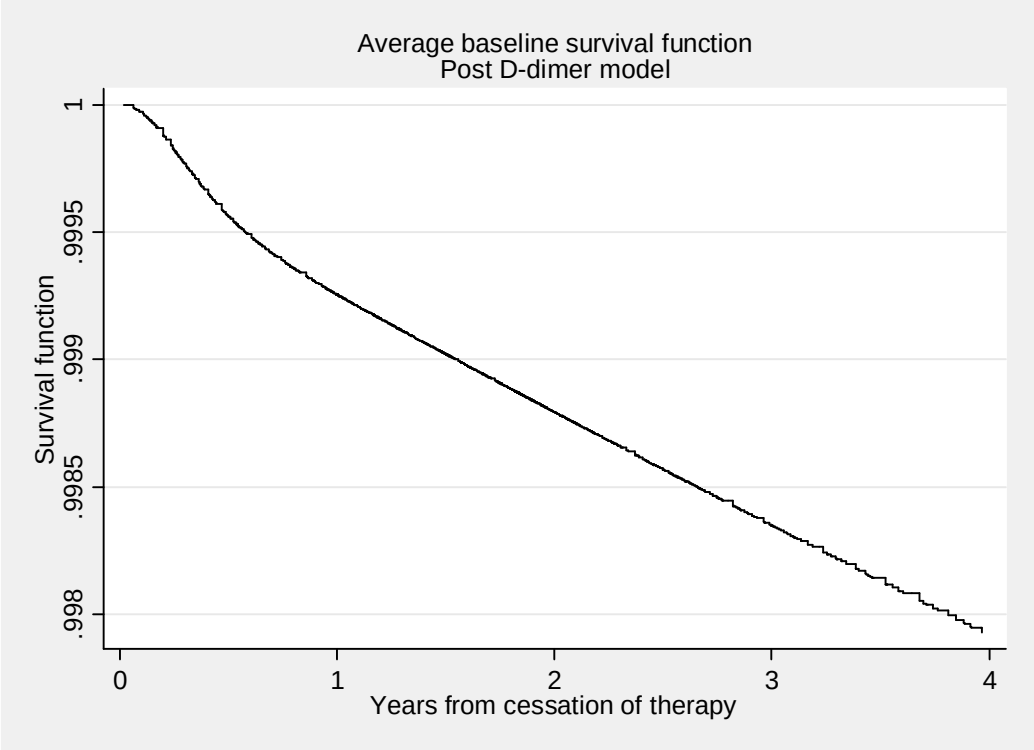


Figure 3.12 - Average baseline (recurrence free) survival function for the post D-dimer model

Example application of the model

As an example of the potential application of the post D-dimer model, three example patients were created using varying predictor information to illustrate patients at different risk of recurrence (see Table 3.8). For each of the continuous predictors (age, log D-dimer and log lag

time), the 25th, 50th and 75th percentile of the predictors distribution was used for patients A, B and C respectively, to reflect the RVTE database (see Table 3.1). All three patients were selected as male, and the site of index event was selected as distal DVT, proximal DVT and PE for patients A, B and C respectively. An example of the risk score created using these patient characteristics is presented for patient A in Equation 3.4. Both recurrence free survival probability and probability of recurrence were predicted at over time post cessation of therapy for patients A, B and C respectively (see Figure 3.13).

Equation 3.4 - Risk score equation for Patient A using the post D-dimer model

$$\boldsymbol{\beta x} = (-0.0105 \times Age\ (= 51)) + (0.545 \times Gender{:}\,Male\ (= 1)) + (1.735 \times Site{:}\,Proximal\ DVT\ (= 0)) + (1.756 \times Site{:}\,PE\ (= 0)) + (0.701 \times (Log)\ D\ dimer\ (= 5.55)) + (-0.291 \times (Log)\ Lag\ time(= 3.14))$$

Table 3.8 - Model parameters for three example patients and recurrence free survival/recurrence risk predictions using post D-dimer model

Model predictor	Patient A	Patient B	Patient C
Age (years)	51	64	74
Gender			
Male	1	1	1
Female	0	0	0
Site of index			
Distal DVT	1	0	0
Proximal DVT	0	1	0
PE	0	0	1
D-dimer (ng/mL)	275	417.5	747
Log (D-dimer)	5.55	6.03	6.62
Lag time (days)	22	29	33
Log (Lag time)	3.14	3.4	3.53

The post D-dimer model predictions are presented in Figure 3.13 and Figure 3.14, with recurrence free survival probability and probability of recurrence calculated using Equation 3.1 and Equation 3.3, respectively. Predicted recurrence free survival probability can be seen

to decrease over time for all three example patients (see Figure 3.13). The predicted $S_0(t)$ is markedly different between patient A and the other two patients, this is likely due to lower values of continuous predictors such as D-dimer, and also the low risk site of index event (Distal DVT) contributing little within the post D-dimer model (see Equation 3.1) to patient A's risk of recurrence. Smaller differences were observed between patient B and C, reflecting the similar effect seen for proximal DVT and PE index events (see section 3.3.4).

The predicted probabilities of recurrence free survival can be seen in Figure 3.13 at the intersection of the vertical gridlines with the predicted curves. For example the vertical gridline corresponding to one year from cessation of therapy, intersects with the predicted curve for patient A at the predicted probability of recurrence free survival (0.985) from Equation 3.1 using patient A's risk score (see Figure 3.13).

Similarly the predicted probability of recurrence over time from cessation of therapy can be predicted using the post D-dimer model (see Equation 3.3). The probability of recurrence is opposite to the probability of recurrence free survival, increasing over time from cessation of therapy (see Figure 3.14). The same trends seen in Figure 3.13 between patients A, B and C can be seen in Figure 3.14 as expected.

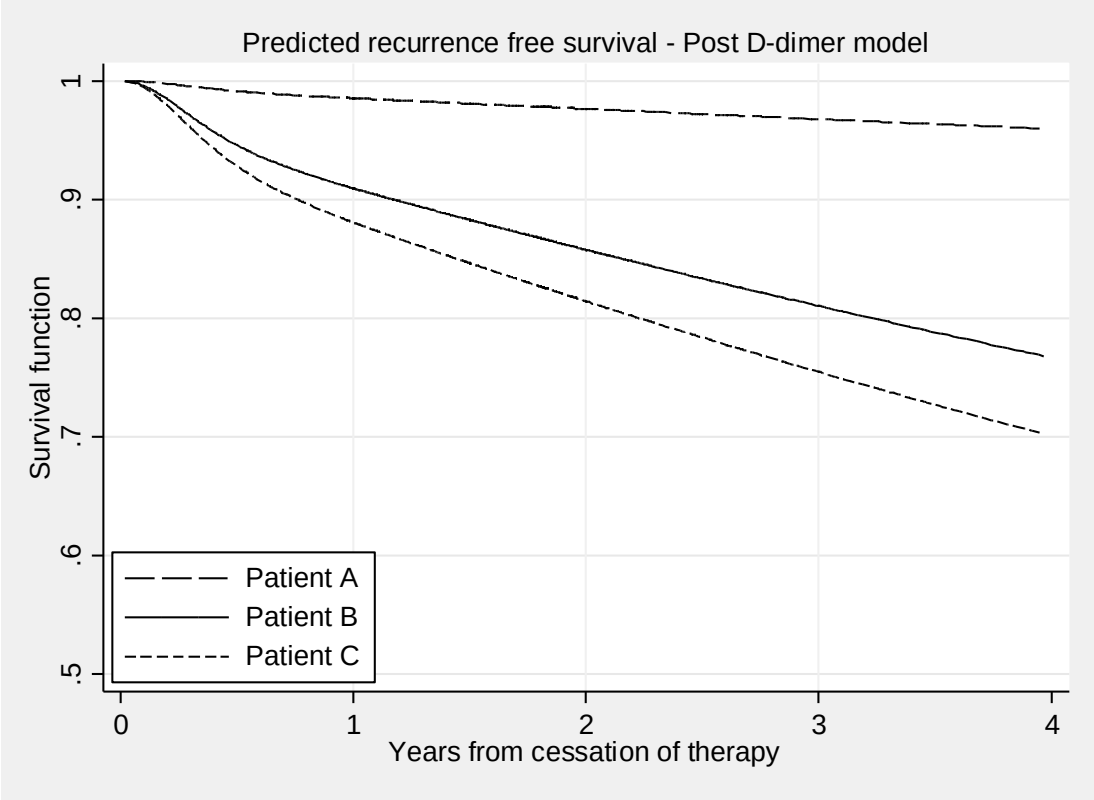


Figure 3.13 - Predicted recurrence free survival for three example patients using the post D-dimer model

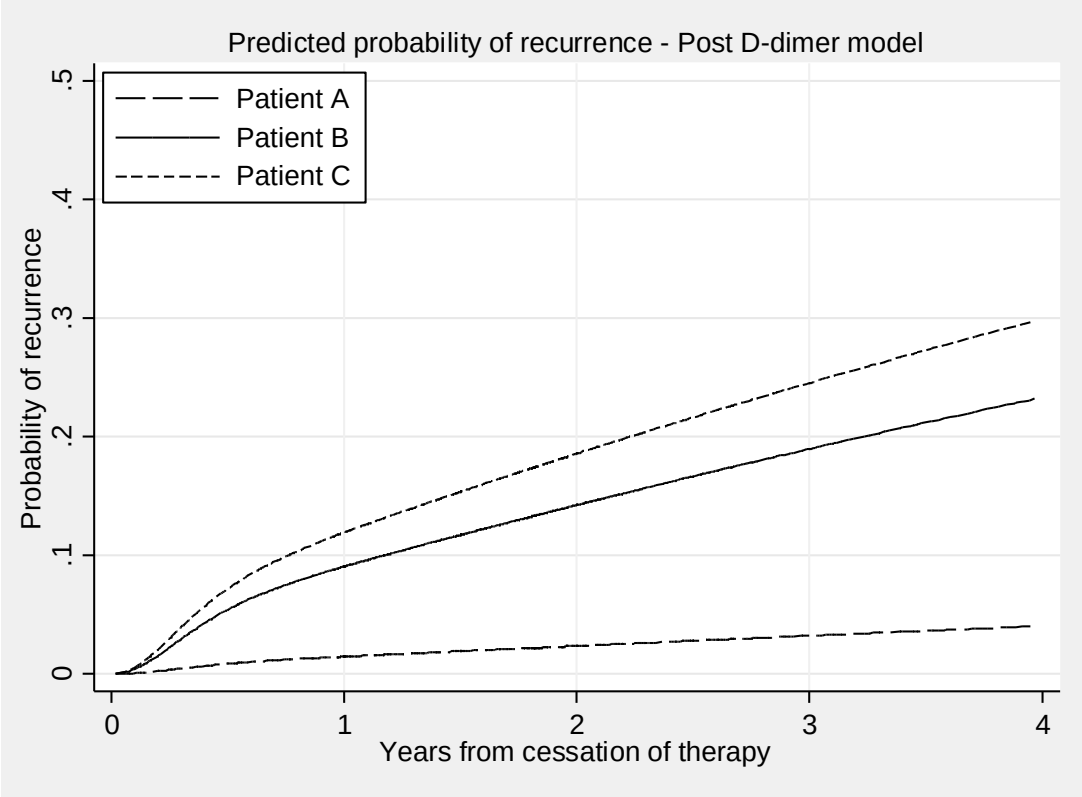


Figure 3.14 - Predicted probability of recurrence for three example patients using the post D-dimer model

3.4 Discussion

As the systematic review in Chapter 2 identified that existing prognostic models were inconsistent in their definition of an unprovoked VTE, and were at a moderate to high risk of bias due to a lack of external validation, it was important to address this in new research. This project therefore used the IPD meta-analysis database supplied by the Recurrent VTE collaborative group (RVTE) to develop and externally validate two new models: a pre D-dimer model and a post D-dimer model. Ideally external validation of the existing models identified by the review would also have been undertaken, but the lack of permission and unavailable predictors in the database prevented this. This work has been published (98), and now the key findings, recommendations and limitations are summarised.

Development and validation of a new prognostic model

The RVTE database contained seven trials (198, 203, 204, 214, 216-218), and therefore allowed the novel framework of Debray et al (20) to be utilised for model development and external validation. This approach adapts the Internal-External Cross-Validation (IECV) procedure first described by Royston et al (66), whereby N-1 trials are iteratively selected from the N total trials in the IPD meta-analysis, and the prognostic model is developed within this subset of trials, leaving the remaining trial for validation of the model. In this manner, it was possible to investigate (across all permutations of the excluded trial) whether model performance remained consistent when applied in another trial's population that was not included during model development. In other words, external validation was possible on multiple occasions. Although a complete-case analysis was primarily performed, a sensitivity

analysis using multiple imputation under a missing at random assumption led to the same set of predictors being included, and gave similar parameter estimates in the models.

In all models, the Royston-Parmar approach was used to flexibly model the baseline hazard using restricted cubic splines. The baseline hazard is essential for individualised predictions from a survival model, and the use of splines allowed the shape to be modelled flexibly, without forcing a particular parametric form. This is likely to improve the performance and generalisability of the developed prognostic model, especially as the shape of the baseline hazard was observed to be very similar across the set of trials in the RVTE database.

The development of the pre D-dimer model was considered as, in contrast to the post D-dimer model, it allows individual risk predictions at the exact time therapy might cease. However, it contained only two predictors: gender and site of index event. Upon external validation (in the IECV approach using the RVTE database) the model had poor discrimination with an average C-statistic of 0.58 across the trials. Although calibration appeared good on average, there was heterogeneity in calibration performance across trials, and in some it was rather poor. The pre D-dimer model is therefore clearly inadequate, which is not surprising given only two predictors were identified as important. Further research is needed to extend the set of included predictors. In particular, it may be that D-dimer measured without a lag-time is also an important predictor, and therefore datasets need to be collected to examine whether this adds prognostic value in terms of calibration and discrimination. At the time of this work there were currently three studies known to be on-going in UK (Exact), Holland (VISTA) and Canada (DODS) which may provide further information regarding D-dimer testing whilst still on oral anticoagulation therapy.

The development of the post D-dimer model allowed the inclusion of D-dimer measured at a particular lag-time, along with age, gender, site of index. This model had substantially improved discrimination performance compared to the pre D-dimer model. C statistics ranged from 0.64 to 0.81 in the IECV cycles, with the average C-statistic of 0.69; other published clinical prediction models have similar discriminatory ability (24). Calibration of the model appeared excellent on average across the external validation trials, especially up to 2 years, although there remained heterogeneity. Ideally, further external validation studies would be helpful to examine this heterogeneity further, as it was estimated with large uncertainty in the IECV approach (due to only 5 external validation trials). Furthermore, the patients included in the RVTE database were enrolled in clinical trials, and thus may not be representative of all populations of interest.

Heterogeneity of performance across studies

Examining heterogeneity of prognostic model performance across studies and populations is a novel idea: most prognostic models are just considered in one external validation study, or just report the average performance across multiple clusters (e.g. practices, studies) (17). Ideally there would be no heterogeneity, but this is a very high standard to attain, and this issue is rarely evaluated for other well-used prediction models (21), and was never considered for the DASH, Vienna or HERDOO2 models for VTE recurrence. The IECV showed that calibration of the post D-dimer model was excellent up to about two years, on average across the trials. This means that across applicable populations, it is expected the final post D-dimer model would perform well on average up to two years. Heterogeneity in calibration performance at the individual population-level would be reduced if a population-specific

baseline hazard were used, rather than our average baseline hazard (or equivalently a population-specific $S_0(t)$ rather than our average $S_o(t)$ currently in the model) (19, 20). Identifying population-specific baseline survival functions would facilitate conditional predictions, where interest lies in predictions at the specific level, and this may also improve performance of the model (17). Heterogeneity might also be reduced by including additional predictors. These areas could be the subject of further work.

One particular area for potential heterogeneity is in the use of various D-dimer assays across the studies in the RVTE database. There is inherent variability in the different D-dimer assays used, particularly in the recommended cut-offs used to decide on a normal or abnormal D-dimer result. In the RVTE database there were five D-dimer assays used, with each study using one assay exclusively (see Table 3.9). This is a potential limitation of the post D-dimer model in that the model was built on data using these five assays to measure patient D-dimer, and therefore predictions from the model in practice may only be valid in cases where one of these assays was used.

Table 3.9 - Different D-dimer assays used within the RVTE database

Trial	Palareti 03	Palareti 06	Poli	Tait	Eichinger	Baglin	Shrivastava
D-dimer assay	VIDAS (ELISA*)	VIDAS (ELISA*)	IL-Test (LIA^)	VIDAS (ELISA*)	Asserachrom (ELISA*)	MDA (LIA^)	STA Liatest (ELISA*)

** Enzyme-linked immunosorbent assay;*
^ Latex immunoassay

However, it may also be considered a strength to have used data based on multiple assays, as this enhances the generalisability of the model, making applicable to a wider population. Previous research has investigated the link between variability in D-dimer assays and

recurrent VTE, and found that various assays do not differ in ability to predict recurrence (15). It is also not possible to differentiate the study-level assay effect from other study-level covariates, such as location of study or year of study. It is therefore difficult to discern if any assay effect is genuine, as it may be confounded by other study-level covariates. Further, if one included assay in the model, then external validation of the model in the excluded trials would not be possible as most trials used a unique assay. The discrimination of the post D-dimer model was shown to be reasonably consistent, with moderate to good discrimination regardless of the D-dimer assay used in the validation study. Similarly, the calibration performance up to 2 years appears very good in all trials, with generally very small miscalibration on average. Therefore, by developing a model using all assays, the hazard ratio obtained for D-dimer appears to provide reasonably robust predictions in external validation on average, regardless of the assay available. Finally, a small sensitivity analysis was conducted to crudely assess the impact of differences in the continuous scale of D-dimer assays on the predicted risk of recurrent VTE from the post D-dimer model, with results showing very little change in the predictions, certainly not enough to alter a clinical decision on choice of therapy (see 0).

An interesting area for discussion is the observed effect of D-dimer and lag time in the final post D-dimer model. The effect of a patients D-dimer level appears to indicate an increase in recurrence rate of around 70% for every 1 unit increase in log D-dimer level, with a hazard ratio of 1.716 (95% CI: 1.43, 2.06). Conversely the lag time between cessation of therapy and measurement of patient's D-dimer appears to decrease recurrence rate by around 20% for every 1-unit increase in log lag time. This effect appears to be counterintuitive as it may be

expected that recurrence rate would increase the longer it takes to measure patients D-dimer and identify those with high D-dimer at greater risk of recurrence. However the observed effect of lag time may be acting as a proxy for time from cessation of therapy itself, in that the more time which elapses from cessation of therapy the greater chance that patients at higher risk of recurrence will have already had a recurrence, leaving a more selected population of healthier patients.

Clinical usefulness

Given the good discrimination and the excellent average calibration performance demonstrated through external validation, the Post D-dimer model would appear suitable for informing patient counselling and clinical decision making at a particular lag-time post cessation of therapy. Section 3.3.5 detailed how to apply the model in practice, to obtain individual risk predictions for new patients.

In terms of usefulness in clinical practice, it should be noted that the post D-dimer model has important limitations. Because anticoagulation significantly lowers D-dimer, measurement of D-dimer in the dataset was always performed after some lag time (or wash-out period), to allow the effects of therapy to subside. Therefore the post D-dimer model is only applicable at a set lag time post cessation of therapy, meaning it can be used only after a delay in making the decision on a patient's therapy. While this is current practice, with D-dimer recommended to be measured around 30 days after cessation of therapy, there has been some evidence toward the predictive ability of D-dimer on therapy (219) and there are several on-going studies investigating the predictive ability of D-dimer on-therapy. Evidence from the RVTE database suggests that approximately 58.7% of recurrent events occurred within the 30 day

lag time before D-dimer measurement (for the pre D-dimer model dataset). Thus, as mentioned above, more clinically useful models might be derived by extending the pre D-dimer model with other predictors measured without any lag time.

It should be discussed that many may consider an initial distal DVT as a low risk group of patients, in whom many would not favour prolonged OAC therapy, and that some would chose not to include such a low risk group within the model development (for example the DASH model did not consider such patients). However, this tendency to cease therapy in patients with initial distal DVT, has in this case been captured within the post D-dimer model through the inclusion of such patients. This means that predictions from the model indicate that in the majority of cases these patients have low predicted risk of recurrence. Subsequently in practice post D-dimer model predictions would lead to the same decision not to prolong OAC therapy.

Limitations

It is important to note that patients who died without any recurrence were censored in the analysis performed in this chapter and therefore predictions from the Post D-dimer relate to a hypothetical world where patients cannot die before a recurrence occurs. The proportion of deaths before a recurrence is likely to be very small (especially up to 2-3 years follow-up where the model calibrates well), and therefore the model predictions would not change importantly if a competing risks model had been used. Further research may look to investigate the performance of a Post D-dimer model developed in a competing risks framework as seen in the updated Vienna model (184).

A potential limitation of the study is the exclusion of BMI as a candidate predictor due to systematic missingness in three of the included studies. While it was the intention of the study to consider BMI as a potential predictor, it was considered inappropriate to impute across studies. Selection of the BMI predictor within the model was also not assumed certain, because while there is evidence to suggest that BMI is an important predictor for a first VTE event, there is conflicting evidence for the effect of BMI on VTE recurrence. The systematic review undertaken in chapter 2, identified three models which assessed the impact of predictors in combination on VTE recurrence, and could therefore be considered the strongest evidence to date of which predictors affect recurrence risk. Of these three, the Vienna model found BMI to be a weak predictor (1.19 HR per 5kg/m^2 change in BMI), and to be non-significant when optimism was adjusted for (40). The DASH model found BMI to be non-significant at univariate analysis, as did the HERDOO2 model (37, 39). The HERDOO2 then went on to split their analysis by gender and only then found BMI to be important in women alone (p-value = 0.02) (39). Heit et al. and Eichinger et al. also provide conflicting evidence suggesting that BMI is a weak risk factor in the order of a HR of around 1.2, with 95% CI's covering values close to 1 (220, 221). This evidence suggests that BMI may not be a strong consistent predictor of VTE recurrence risk when adjusted for other important predictors including site of index event. However, further research is warranted.

Further limitations concern the use of new oral anticoagulants (NOAC's). The studies included in the RVTE database used primarily warfarin to treat patients first VTE, none of the studies used any of the NOAC's. In this regard the model is built on and therefore applicable to patients treated with warfarin. This was a limitation of the available study data, because no

studies used NOAC's, the effect of these drugs could not be accounted for in the modelling process.

It must also be noted that while external validation was possible using the IECV approach, it would also be beneficial to undertake external validation in non-trial datasets, and therefore the post D-dimer model could also be considered at moderate risk of bias (see quality assessment defined within chapter 2), until such external validation is undertaken.

There are also potentially broader uses of the post D-dimer model, as prognostic models are useful at many stages of the translational pathway toward improved patient outcomes (9). For example, it might be used to improve the design and analysis of randomised trials in patients with a first unprovoked VTE, as a stratification factor in the randomisation process (to ensure treatment groups are balanced in the predicted risk of recurrence) or as an adjustment factor to increase statistical power (5, 6). Inclusion criteria for trials may also be restricted to individuals with a high risk of recurrence based on the model. It could also be used to adjust for case mix variation (confounding) in health services research and observational studies.

Recurrent VTE collaborative database

A slight tension in using the RVTE database was that parts of it had already been used to develop the DASH score and Vienna model (37, 40). However, the new research conducted within this chapter can be seen to enhance current research which uses the RVTE database, and research in this field in general, by:

(i) Using the novel IECV approach to externally validate the model multiple times, unlike existing scores which rarely have external validation (see Chapter 2);

(ii) The identification of additional predictors not previously picked up (e.g. age, lag time);

(iii) Directly modelling the baseline hazard, which allows predictions over all follow-up times up to five years or more (rather than at just a few time-points as in previous models);

(iv) Not requiring simplification of the model to make predictions, as the provided equation can be used to predict recurrence using the values at hand (i.e. no need for a simplified score) and the baseline survival;

(v) Identifying the distribution of population characteristics for use in subsequent health economics modelling and impact studies.

Further, the DASH score developed by Tosetto et al. using the RVTE database is fundamentally different to proposed post D-dimer model developed within this chapter. The DASH score is only applicable in a distinctly different population of patients, one which uses a different definition of an unprovoked first VTE (37). Indeed the DASH score includes predictors for hormone intake, where any patients provoked by hormone intake were excluded from the post D-dimer model development as per the pre-defined definition of unprovoked VTE (see section 3.2.2). As such the DASH score could not be compared to the post D-dimer model, as they include different predictors, and are applicable in different populations, despite both being developed within the RVTE database.

Further research recommendations

The systematic review (see chapter 2) and model development within this chapter provide a new prognostic tool to aid clinical decision making for patients who sustain a first unprovoked VTE. The clinical paradigm shifted whilst this project was being undertaken with a view now to identify those patients at sufficiently low risk of recurrence that they can safely stop oral anticoagulation therapy after a relatively short period of time (usually 3-6 months). Currently available models are not routinely used within UK practice and have not been included within NICE guidelines. The post D-dimer model may therefore have an important role in the future. However, a number of further research recommendations arise from this work, which are now outlined.

Develop and externally validate a prognostic model that can be used at the point of considering cessation of therapy. This should build on the pre D-dimer model, and thus include gender and site of index event. Evaluation of the prognostic ability of D-dimer levels measured at the exact time of cessation of therapy is needed (i.e. measured at a lag time of 0).

Further external validation of the post D-dimer model, especially in non-trial populations. Trial populations available within the RVTE database may be a select group of individuals, and thus the post D-dimer model requires validation in other populations, for example from cohort studies or large databases. Such datasets may not currently be available that contain D-dimer values, and so further observational studies are needed that enrol new patients, measure their predictors following cessation of therapy (including D-dimer measurements and lag time), and recording of VTE outcomes.

Further research to examine if between-study heterogeneity in the calibration performance of post D-dimer model can be reduced. Though the post D-dimer model performed excellently on average across all trial populations, there was between-trial heterogeneity in the calibration. Further research should seek to reduce this heterogeneity, by potentially updating the model with additional predictors (requiring further external validation of course) and/or by identifying revised $S_0(t)$ functions for populations that differ importantly from the average $S_o(t)$ currently used in the model (i.e. perform model recalibration).

Further research to develop and validate a prognostic model for bleeding on therapy. There is an immediate need to develop a prognostic model to predict individuals' risk of bleeding whilst on therapy. This would allow the balance between risk of recurrence and risk bleeding to be accounted for in the decision of treatment strategy, and also an effective economic evaluation to be undertaken.

Direction for subsequent chapters of the thesis

The first chapters of this thesis have focused on the development and evaluation of models for VTE recurrence. However, a number of methodological challenges have arisen, which will now become the focus for the remainder of the thesis. In chapter 4, the issue of how to reduce heterogeneity in a model's calibration performance across multiple studies will be considered in detail. In particular, whether recalibration and model updating techniques can be applied to reduce heterogeneity, and thus improve consistency in performance across different populations and settings (21). Unfortunately, due to externally imposed restrictions in using the RVTE database beyond the work presented above, a different clinical setting and IPD meta-

analysis dataset will be used in Chapter 4 to explore the methodology. However, it is hoped that in the future, the holders of the RVTE database will themselves look to examine the issue of heterogeneity using the novel approaches developed. In chapters 5 and 6, the issue of missing predictor values will be revisited, and methodology work undertaken to improve risk predictions in that situation.

CHAPTER 4: INDIVIDUAL PARTICIPANT DATA META-ANALYSIS FOR EXTERNAL VALIDATION AND RECALIBRATION OF A FLEXIBLE PARAMETRIC PROGNOSTIC MODEL

4.1 Introduction

The previous chapter developed and validated a prognostic model for VTE recurrence. This built on a systematic review which identified a number of existing models which unfortunately could not be validated in the data at hand. The new model performed well on average across omitted studies in the internal-external cross-validation (IECV) approach. However, there was heterogeneity in the model's calibration and discrimination performance across studies, especially at later time-points. This suggests that, ideally, approaches to improve the model's performance in particular populations and settings may be important, to reduce or ideally remove this heterogeneity (19, 20, 64, 69, 142-145).

Model recalibration is perhaps the simplest approach to improve model performance in new (external) populations or settings, with model updating further adjusting the existing model (usually the regression coefficients) (34, 64, 69). Methodology in this area is well developed for both binary and time-to-event outcomes (34, 142-145). Royston and Altman also describe potential methods for external validation of an existing Cox model given different levels of reported model statistics, but do not discuss recalibration or updating of an existing model (147). Approaches to model recalibration and updating for time-to-event models have been described by van Houwelingen et al., and are based on parametric models as these allow

estimation (and thus recalibration) of the baseline hazard, an essential part of a survival model (144, 145).

The focus in this chapter is how external validation and model improvement (updating) can be examined when Individual Participant Data (IPD) are available from multiple studies. Where multiple studies are available for external validation, model performance may be assessed on multiple occasions and summarised using evidence synthesis methods. This allows potential comparison of different recalibration methods across a set of validation studies of differing case-mix. Previous work has proposed the use of IPD meta-analysis to compare model implementation strategies focusing on logistic regression models, and multivariate meta-analysis (19, 20, 24). In this chapter, the focus is on external validation of a flexible parametric survival model, and how IPD meta-analysis can help examine model performance, and evaluate the best recalibration strategies in different settings to reduce heterogeneity. FP models improve on standard survival models such as the Cox model, by using splines to parameterise the baseline hazard, enabling external validation and adjustment of the baseline hazard which is of key interest in this chapter.

The remainder of the chapter is structured as follows. Section 4.2 introduces a motivating example and individual participant dataset. Section 4.3 briefly describes the methods and advantages of FP modelling using the Royston-Parmar approach (44, 45), revisits the methods for synthesis of model performance statistics from multiple validation studies, and discusses in detail methods for model recalibration. Section 4.4 presents the results of the recalibration methods using a motivating example in breast cancer, and Section 4.5 concludes with some discussion. Note that, due to restrictions imposed by the owners of the RVTE database utilised

in Chapter 3, it was not permissible to use the VTE IPD meta-analysis any further in this thesis. Hence, the motivating example in this chapter will utilise a new IPD meta-analysis database in breast cancer, kindly provided by Dr Maxime Look (Rotterdam).

4.2 Motivating example

4.2.1 Breast cancer dataset

The IPD meta-analysis dataset used in this chapter contained 5978 breast cancer patients, with follow-up ranging from 1 to 120 months (for more information on the original study see (222)). It was formed by pooling datasets from eight centres (hereafter referred to as 'studies' for simplicity) across six countries, with Rotterdam having the largest patient numbers (see Table 4.1). The focus here is on using the IPD to identify a prognostic model that reliably predicts an individual's probability of recurrence-free survival over time, defined as the time to recurrence or death from any cause. Though there is also methodological interest in how to develop a prognostic model utilising more than one study (20, 66), this chapter only consider how to perform external validation and model updating using the IPD multiple studies, for a previously proposed prognostic model.

To this end, for illustrative purposes and to ensure most studies were available for external validation, we selected just one study (Rotterdam) as a derivation dataset in which to build the prognostic model. The remaining seven studies were thus immediately available for external validation. In the Rotterdam study there were 2627 patients with 1224 events and a median follow-up of 64 months (max 120 months). Previous research has suggested that adequate sample size for model development should have at least 10-20 events per predictor

(EPP), making the Rotterdam study suitable for investigation of up to 60 predictor effects (188, 189, 223, 224). A recent study has questioned the validity of the EPP rules previously proposed, and highlighted that sample size requirements for model development likely vary based on a number of factors and are therefore case-specific, though further research is needed in this area (190).

In the validation studies, the number of events ranged from 80 to 211 per study, with a total of 1019 events across all validation studies. Study characteristics are summarised in Table 4.1. In terms of validation, recent evidence suggests that at least 100-200 events and non-events are required for external validation (76, 191, 215), meaning that the breast cancer data is likely sufficient for validation within each study, perhaps apart from the Nijmegen study which has a relatively small sample size of 80.

Table 4.1 - Summary statistics for Look et al. dataset. NB: RFS – Recurrence free survival; * Median; # Number and percentage.

Model phase	Development	Validation						
Study	**Rotterdam**	**Sweden**	**Lille**	**Nijmegen**	**St Cloud**	**Switzerland**	**Denmark1**	**Denmark2**
Total sample size	2627	621	552	293	499	620	444	322
RFS (Events)	1224	137	150	80	168	150	211	123
% Events	47%	22%	27%	27%	34%	24%	48%	38%
Follow up (Months)*	63.64	106.84	60.71	52.90	97.84	42.71	48.34	55.36
Min	1.18	3.04	1.08	1.28	4.01	1.71	1.02	1.02
Max	120.00	120.00	120.00	120.00	120.00	83.45	120.00	100.17
Age (Years)								
Mean	56.46	58.47	57.00	56.68	58.82	57.94	53.86	56.21
SD	13.28	11.64	11.19	13.01	12.56	11.31	10.85	10.73
Lymph nodes#								
0	1371 (52.19)	226 (36.39)	381 (69.02)	153 (52.22)	233 (46.69)	357 (57.58)	299 (67.34)	152 (47.2)
1 to 3	684 (26.04)	243 (39.13)	96 (17.39)	89 (30.38)	177 (35.47)	165 (26.61)	95 (21.4)	89 (27.64)
4 to 10	422 (16.06)	125 (20.13)	57 (10.33)	33 (11.26)	72 (14.43)	56 (9.03)	46 (10.36)	57 (17.7)
> 10	150 (5.71)	27 (4.35)	18 (3.26)	18 (6.14)	17 (3.41)	42 (6.77)	4 (0.9)	24 (7.45)
Menopausal status#								
pre	1076 (40.96)	191 (30.76)	193 (34.96)	103 (35.15)	146 (29.26)	206 (33.23)	220 (49.55)	104 (32.3)
post	1551 (59.04)	430 (69.24)	359 (65.04)	190 (64.85)	353 (70.74)	414 (66.77)	224 (50.45)	218 (67.7)
Tumour size#								
≤ 20mm	1177 (44.8)	217 (34.94)	302 (54.71)	102 (34.81)	211 (42.28)	298 (48.06)	179 (40.32)	96 (29.81)
>20-50 mm	1296 (49.33)	396 (63.77)	242 (43.84)	165 (56.31)	271 (54.31)	306 (49.35)	232 (52.25)	192 (59.63)
>50 mm	154 (5.86)	8 (1.29)	8 (1.45)	26 (8.87)	17 (3.41)	16 (2.58)	33 (7.43)	34 (10.56)
Adjuvant treatment#								
no	1998 (76.06)	142 (22.87)	278 (50.36)	172 (58.7)	177 (35.47)	125 (20.16)	310 (69.82)	132 (40.99)
yes	629 (23.94)	479 (77.13)	274 (49.64)	121 (41.3)	322 (64.53)	495 (79.84)	134 (30.18)	190 (59.01)

4.3 Methods for examining performance of an FP model using IPD meta-analysis

4.3.1 Flexible parametric models

Chapter 1 introduced the framework for flexible parametric survival models for developing prediction models (43-46). This current chapter assumes a prediction model has already been developed in this framework, and is of the form,

$$\ln H(t) = \ln H_0(t) + \boldsymbol{\beta X} = spline(\ln t) + \boldsymbol{\beta X}$$

Equation 4.1

where the baseline hazard is modelled using splines as discussed in chapter 1 and of the form;

$$spline(\ln t) = \gamma_0 + \gamma_1 \ln t + \gamma_2 z_1(\ln t) + \gamma_3 z_2(\ln t) + \cdots$$

Equation 4.2

In Equation 4.1, $\boldsymbol{X} = (x_1 \dots x_j)^T$ and $\boldsymbol{\beta} = (\beta_1 \dots \beta_j)^T$ define vectors of predictors and their corresponding coefficients.

Advantages of FP models

Flexible parametric (FP) models extend standard parametric survival models, such as the Weibull or exponential model, by modelling the baseline hazard more accurately (43-45). There are several benefits to using FP models over other more traditional parametric or semi-parametric models for this chapter. In particular, semi-parametric models such as the Cox model do not estimate the baseline hazard, providing only relative effects not absolute risks. Being able to parameterise and estimate the baseline hazard is essential for prognostic model research; firstly in order to obtain individualised absolute risk predictions over time and secondly, for out-of-sample prediction enabling external validation. Also, other parametric

models such as the Weibull model cannot capture complex baseline hazard shapes which rise and fall over time. Even parametric models which can encompass turning points in the hazard function (e.g. generalised gamma) will often not be able to fit well to complex functions (46).

Figure 4.1 shows the baseline hazard function as estimated in the Rotterdam study of breast cancer dataset, using varying numbers of degrees of freedom to estimate the baseline spline function. It is clear from the plot that the baseline hazard peaks around 1.5 years from surgery, and then falls steadily up to the end of follow-up time. The dotted line shows the special case of 1 degree of freedom, where the FP model collapses to the Weibull model (see chapter 1). It is clear that the Weibull model cannot capture the peak and then drop in hazard seen in the Rotterdam study. The two solid curves for 2df and 3df best represent the baseline hazard shape; the best fit can be identified by comparison of AIC and BIC values (as in chapter 3). Finally, the generalised gamma distribution is also plotted and again does not capture the true shape of the baseline hazard here, despite being able to handle turning points in the hazard function (see Figure 4.1). Later, we discuss that flexible modelling of the baseline hazard is also important when re-calibrating a model to a new population (see section 4.3.4).

The focus of this chapter is on how to validate an existing FP model using IPD from multiple studies. The following sections re-introduce important performance measures, and explain potential strategies for recalibration of a model to improve model performance in external populations.

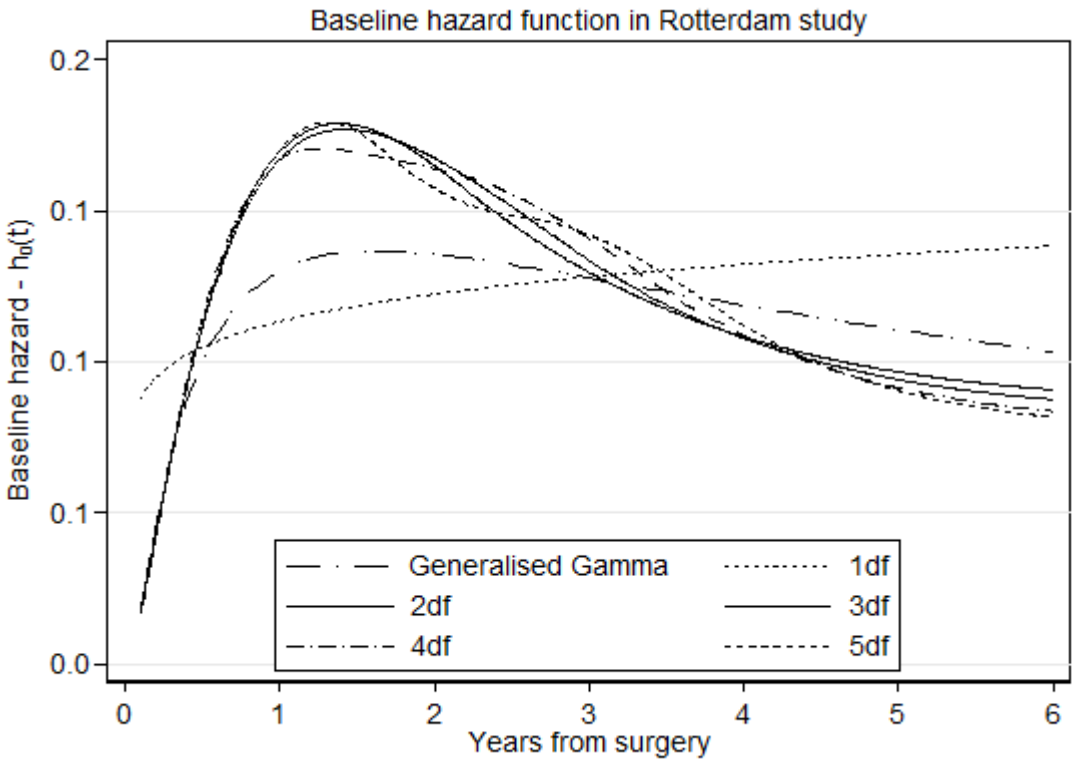


Figure 4.1 - Baseline hazard function in the Rotterdam study estimated using various numbers of knots for the baseline spline in an FP model. The baseline hazard estimated using a generalised gamma distribution is also included.

4.3.2 Performance statistics

The predictive performance of a prognostic model is often described in terms of calibration and discrimination (33), both of which can be measured using various statistics (72). This chapter focuses on expected (E) and observed (O) survival probabilities, allowing calculation of both the ratio (E/O) and difference (E-O) in these quantities at specific time points, as well as calibration plots to summarise calibration performance (147). The discrimination performance of the model was measured using Harrell's C-statistic (212), and Royston's R^2_D and D-statistic (see chapter 1) (75).

Calibration plots are a common method of assessing calibration, in which observed events (typically using a Kaplan-Meier curve), and expected events (as predicted by the prognostic model) are plotted over time, to provide a visual assessment of their agreement.

The E-O statistic can be considered as the difference between the Kaplan-Meier survival estimate and the developed models' predicted survival probability at a specific time point in the validation study (34, 98). If the E-O statistic is >0 or <0 there is indication of either over or under fitting, respectively. That is, predictions are systematically too narrow or too extreme. Perfect agreement between the observed and predicted event rate would give an E-O statistic of zero. For the ratio of expected and observed rates as commonly used in logistic regression models, a value of one defines perfect calibration. However there are some limitations to the E/O statistic in terms of interpretation as highlighted in chapter 1.

Harrell's C-statistic gives the proportion of pairs of patients for whom the model predictions and observed outcomes are similar (212). That is for any two patients randomly selected, one with the outcome and one without, the model predicts a higher survival probability for the patient without the outcome. Royston's D-statistic is a measure of separation in survival curves related to the standard deviation of the linear predictor; it gives the log hazard ratio between two groups defined by dichotomising the linear predictor at the median (75). Royston's R^2_D gives a measure of explained variation based on the D-statistic, to give an interpretation similar to R^2 in linear regression models (75).

4.3.3 External validation in multiple studies with meta-analysis of performance

Given IPD from multiple studies, the predictive performance of an existing model can be externally validated multiple times. This leads to multiple estimates for each validation statistic of interest (e.g. Harrell's C-statistic), and so naturally lends itself to a formal meta-analysis in order to summarise and compare performance across studies (19, 20).

Let Y_i be the estimate of a particular performance statistic of interest where *i* represents the validation study and let S_i^2 be the associated variance of Y_i, then a random-effects meta-analysis could be used as given in Equation 1.19;

$$Y_i \sim N(\theta_i, S_i^2)$$
$$\theta_i \sim N(\theta, \tau^2)$$

Equation 4.3

This model assumes that Y_i follows a normal distribution around the i^{th} study's true performance, θ_i and that θ_i is also normally distributed around an average performance, θ and a between-study variance τ^2. Pooling of different model performance measures may not naturally meet these normality assumptions, and therefore such measures require rescaling to approximate normality before synthesis. Recent studies have suggested appropriate transformations for some common performance statistics as listed in Table 4.2 (24, 71, 111).

Table 4.2 - Transformation of model performance statistics required to approximate between-study normality

Performance statistic	Transformation
C-statistic	Logit
D-statistic	None
R^2_D	Logit
E-O	None
E/O	Ln

Usually of most interest will be the estimated θ and its 95% confidence interval, which is derived by $\theta_i \pm 1.96\, SE(\theta_i)$, where $SE(\theta_i)$ is the standard error of θ_i. Also of interest may be an approximate prediction interval, such as a 95% prediction interval (see Equation 1.21). A t distribution is used in the calculation of the prediction interval, rather than a normal distribution, to allow for the additional uncertainty in τ^2 (112, 113).

$$\theta_i \pm t_{0.025}\sqrt{\tau^2 + SE(\theta_i)^2}$$

Equation 4.4

This infers the potential model performance in a new population similar to those included in the meta-analysis (112, 113). A narrower prediction interval implies more consistent performance in new external populations, and is thus desirable if the model is to be generalizable outside of a few local settings. Stata code to perform the meta-analysis methods described above is provide in the appendix (see APPENDIX C2: Stata code), for Stata 14 (225).

4.3.4 Recalibration strategies in a single validation study

If model performance is not good upon external validation, then recalibration strategies may be considered to improve performance. Here four possible methods are proposed (see Table 4.3) for recalibrating a prognostic model (developed using the FP approach) within a single validation study. The four recalibration methods were chosen to progressively increase the extent to which the model is adjusted, and closely link to previously selected methods for a binary outcome example (226). Stata code is provided in the appendix (see APPENDIX C2: Stata code), for each of the methods.

Table 4.3 - Recalibration methods to be investigated

Method	Recalibration technique
1	Re-estimate intercept within $H_0(t)$
2	Re-estimate entire $H_0(t)$
3	Method 1 + Scale linear predictor (i.e. $\boldsymbol{\beta X}$) by a scaling factor
4	Method 1 + Re-estimate particular predictor coefficients

Recalibration options

Recalibration Method 1: keep the same predictor effects and baseline hazard shape, but change the magnitude of the baseline hazard

For this approach, the developed model is applied to the validation study but the constant term, γ_0 in Equation 4.2 is re-estimated within the validation data, to give γ_{0New} in Equation 4.5. Other terms in the model are not altered. This is the simplest form of recalibration, allowing the baseline hazard in the developed model to be shifted by a constant factor to better represent the validation population's baseline hazard. This kind of recalibration is useful when the baseline hazard rate differs substantially between derivation and validation samples (64). In logistic regression the constant term can be related to prevalence of disease and so can be easily altered in new patient populations (20, 146). In survival data it represents a part of the baseline hazard function and so could be interpreted as a constant shift increasing or decreasing the hazard at any particular time.

$$\ln H(t) = spline(\ln t) = \gamma_{0New} + \gamma_1 \ln t + \gamma_2 z_1(\ln t) + \gamma_3 z_2(\ln t) + \cdots$$

Equation 4.5

Recalibration Method 2: keep the same predictor effects but change the entire baseline hazard (shape and magnitude).

This approach extends method 1 to allow adjustment for any differences between the derivation and validation populations in terms of the overall magnitude (method 1), and

additionally the shape of the baseline hazard. So for example the validation population may have a much earlier and sharper peak in their baseline hazard and a long flattened tail, and so recalibration of the FP model to capture this shape may potentially improve the performance of the model in external populations. To implement this method, whilst keeping predictor effects fixed at their original values, the baseline hazard shape and magnitude are completely re-estimated in the validation sample giving a new baseline hazard term as below;

$$\ln H(t) = \ln H_0(t)_{New} + \boldsymbol{\beta X} = spline(\ln t)_{New} + \boldsymbol{\beta X}$$

Equation 4.6

Recalibration Method 3: keep the same baseline hazard shape, change the magnitude of the baseline hazard and adjust the linear predictor as a whole by a constant (φ).

In this approach, method 1 is extended to additionally adjust the linear predictor by a constant scalar term, φ which is estimated in the validation data, so that Equation 4.1 is adjusted as follows,

$$\ln H(t) = \ln H_0(t) + \varphi(\boldsymbol{X\beta})$$

Equation 4.7

As such method 3 allows for a scaling of the predictor effects as a whole, by some constant scalar. For example, if predictor effects are systematically too large, then φ will be < 1 and predictor effects will be shrunk. This may occur when a model was over-fitted during model development.

Recalibration Method 4: keep the baseline hazard shape and the linear predictor fixed, but re-estimate the effect of particular predictors

Poor performance in a validation study may relate specifically to one or a few predictor effects for example, a single predictor, β_1 corresponding to a particular variable X_1. Method 4 builds

on method 1 by additionally re-estimating β_1 in the validation data to give $\beta_{1\,New}$, keeping other all other terms fixed giving;

$$\ln H(t) = \ln H_0(t) + \boldsymbol{\beta X} = spline(\ln t) + \boldsymbol{\beta X} + \beta_{1New} X_1$$

Equation 4.8

Where $\boldsymbol{\beta X}$ now excludes $\beta_1 X_1$. Method 4 could be extended to allow additional predictor effects to be re-estimated within the validation data. Further all recalibration methods could be combined, though this would be equivalent to developing an entirely new model without using information from the original model.

4.3.5 IPD meta-analysis to compare recalibration strategies

The recalibration methods described above can be subsequently evaluated using IPD meta-analysis to identify the best recalibration strategy to be recommended for implementing the model in general (19). In assessing strategies for model recalibration the aim is to achieve good model performance on average as indicated by the pooled performance statistic $\hat{\theta}_i$, as well as small between-study heterogeneity $\hat{\tau}_i^2$ (see Equation 1.19). Importantly the 95% prediction interval can also be used to assess the generalisability of the model in new populations similar to those included in the meta-analysis (112, 113).

For method 4, to identify a problematic predictor effect across studies (for adjustment) requires the examination of its heterogeneity. The original model could be refitted in each validation study and the parameter estimates for each predictor pooled in a random-effects meta-analysis using Equation 1.19. By synthesising the predictor coefficient from each validation study a pooled predictor effect across studies can be obtained, and importantly an

estimate of between-study heterogeneity, τ^2 as in Equation 1.19. If τ^2 is large, or similarly if the prediction interval for a predictor effect is wide, then this may signal that calibration would be improved if the predictor effect was re-estimated.

4.4 Results: Application to the Breast Cancer Example

4.4.1 Visual comparison of the baseline hazard rate in each study

It is helpful to first plot the baseline hazard rate in each study, and compare it to that from the development study. Figure 4.2 presents the baseline hazard function (when all predictors are fixed at zero) in the validation studies, and shows substantial differences in terms of both the shape and magnitude compared to the baseline hazard in the Rotterdam study used for model development. Differences in the magnitude of the baseline hazard represent differences in the rate of events, while differences in the shape of baseline hazard indicate they are non-proportional. These differences may be expected here as the studies in the IPD dataset are from different countries, which are likely to have different patient case-mix, for example some countries may have better breast cancer screening programmes or differing treatment strategies. Figure 4.2 shows that the baseline hazard rate is highest in the Rotterdam study, and is much smaller and flatter in the other studies. Given these differences calibration performance at external validation is expected to be poor, especially if differences are not explained by the predictor effects captured by the developed model.

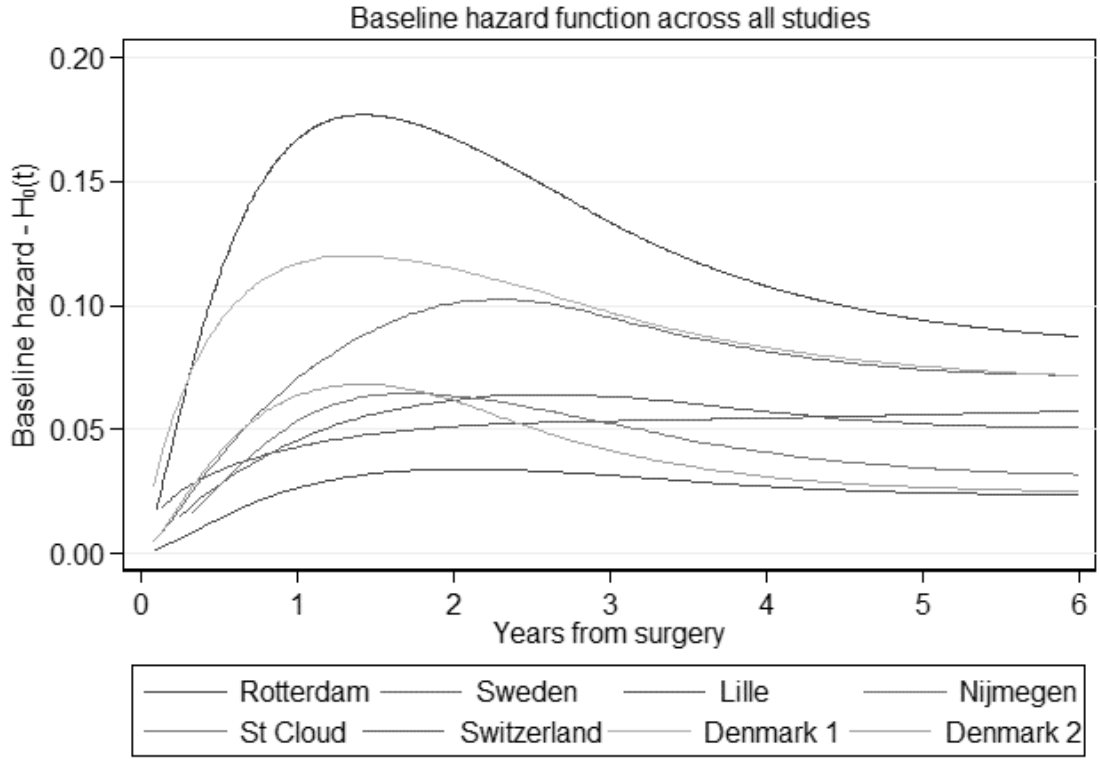


Figure 4.2 - Baseline hazard functions in all 8 studies in the IPD dataset.

4.4.2 Overview of the model development

Using the Rotterdam dataset a prognostic model was developed using a flexible parametric proportional hazards (PH) framework (as described in detail in chapter 1). The included set of predictors was pre-defined, to reduce the potential for optimism in model performance due to predictor selection methods (38). The predictors were: age (in years), tumour size (≤ 20mm, >20-50 mm, >50 mm), number of lymph nodes (0, 1-3, >3-10, >10), menopausal status, and adjuvant treatment (yes or no). Although ideally tumour size and lymph nodes would have been included as continuous, unfortunately their data were provided as categorical. The effect of age was assumed to be linear for simplicity. Given this set-up, including five predictors and some categorical predictors, there were eight predictor effects to be estimated, and thus 153 events for each. Given this large sample size, and because this is merely an illustrative

example, we did not consider it necessary to investigate overfitting (over-optimism) in our developed model; thus, there was no adjustment of estimated predictor effects (i.e. no shrinkage). There was no missing data for the included predictors.

The final model was of the form,

$$S(t) = S_0(t)^{\exp(LP)}$$

where;

$$\begin{gathered} LP = \beta_1 Age + \beta_2 Tumour\ Size(> 20 - 50mm) \\ +\beta_2 Tumour\ Size(> 50mm) + \beta_4 No.\ Lymph\ Nodes(1 - 3) \\ +\beta_5 No.\ Lymph\ Nodes(4 - 10) + \beta_6 No.\ Lymph\ Nodes(> 10) \\ +\beta_7 Menopausal\ status(post) \\ +\beta_8 Adjuvant\ treatment(yes) \end{gathered}$$

Equation 4.9

and S(t) represents the survival probability at time 't', $S_0(t)$ represents the baseline survival probability at time 't', and the β_i represent the regression coefficients of the included predictors. Table 4.4 shows the beta coefficients (log hazard ratios) for the predictors in the developed model. There was strong statistical evidence that all the included predictors were associated with the hazard rate, with increasing tumour size and number of nodes leading to increased hazard rates, and those with adjuvant treatment and post-menopausal women also having increased hazard rates. Age appeared to slightly decrease hazard rates per one year increase in age (HR 0.981, 95% CI 0.973, 0.989). This is consistent with previous analysis of this dataset and with current research, as it is known that young breast cancer patients have a higher mortality rate (222, 227).

Table 4.4 - Predictor effect estimates for the developed model

Predictor	Beta	HR	Lower 95% CI	Upper 95% CI	P-value
Age	-0.019	0.981	0.973	0.989	< 0.0001
Tumour size					
≤ 20mm	0				
>20-50 mm	0.472	1.602	1.399	1.835	< 0.0001
>50 mm	0.833	2.299	1.825	2.897	< 0.0001
Lymph nodes					
0	0				
1-3	0.647	1.909	1.602	2.275	< 0.0001
>3-10	1.260	3.525	2.965	4.191	< 0.0001
>10	1.682	5.376	4.289	6.739	< 0.0001
Menopausal status					
Pre	0				
Post	0.236	1.267	1.036	1.548	0.021
Adjuvant treatment					
No	0				
Yes	-0.454	0.635	0.540	0.747	< 0.0001

The estimated baseline survival function, $S_0(t)$ is shown in Figure 4.3, which shows a gradual decline in survival probability over increasing years from surgery. The curve can be perfectly described by Equation 4.10 below, which is a combination of time and log time variables.

$$S_0(t) = \exp(-\exp(\ln H_0(t)))$$

$$\ln H_0(t) = -1.031 - 1.294(t - 4.265) + 0.349t\,(\ln t - 6.186) + 2.531(X - 3.737) + .044(X \times \ln X - 4.927)$$

Equation 4.10

Where;

$$X = (\ln t + 2.467)$$

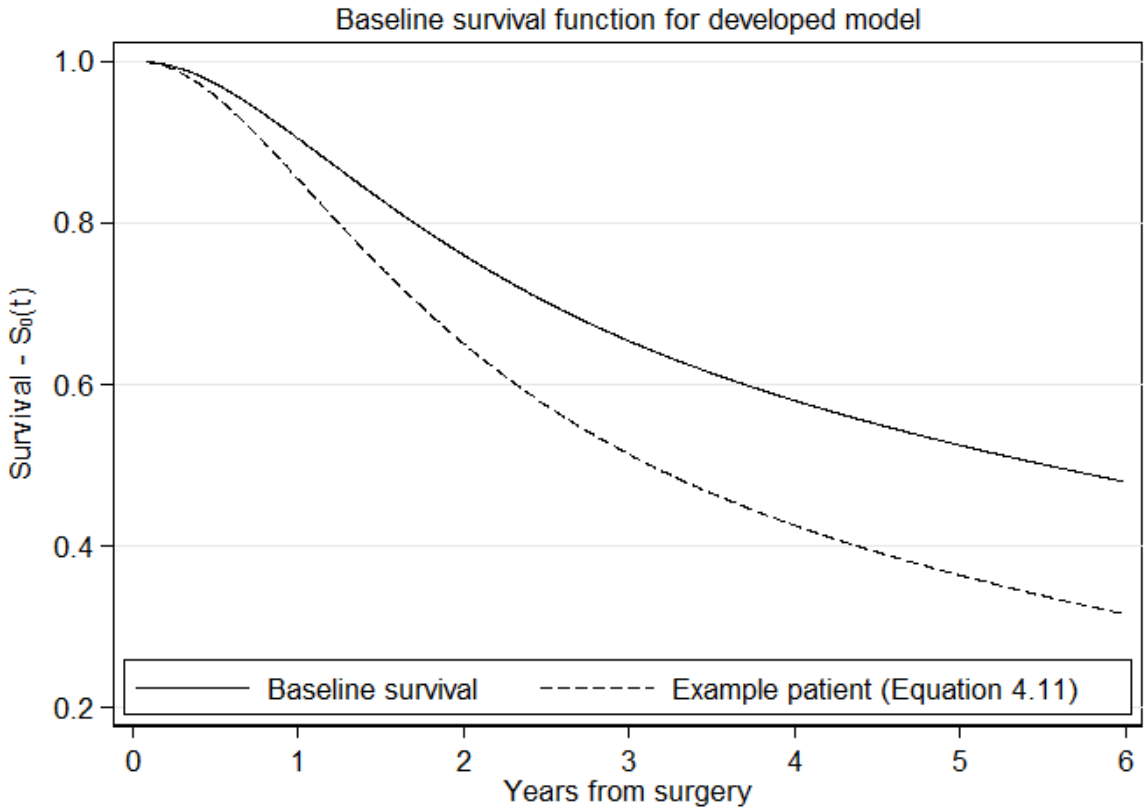


Figure 4.3 - Baseline survival function for developed model (solid line) and predicted survival probability for example individual described in equation 4.11 (dashed line).

Such presentation of baseline survival is rarely reported in practice but is important as it allows researchers to compute the baseline survival at any chosen time point; this can then be combined with predictor effects as in Equation 4.9, to calculate the survival probability for any individual (147). For a new individual, to obtain the probability of recurrence-free survival by time 't', one simply needs to input their predictor values and combine with the $S_0(t)$ value for the time of interest. So, for a 55 year old patient with a 30mm tumour and 5 lymph nodes, who is post-menopausal and has an adjuvant treatment, then their 1 year survival probability can be calculated by,

$$S(1) = S_0(1)^{\exp(-0.019\times Age + 0.472\times Tumour\ Size + 1.26\times No.\ Lymph\ Nodes + 0.236\times Menopausal\ status - 0.454\times Adjuvant\ treatment)}$$

$$S(1) = 0.905^{\exp(0.469)} = 0.852$$

Equation 4.11

The survival function for this example patient is presented in full in Figure 4.3 alongside the baseline survival curve.

Model performance measured in the same data as used for derivation of a model is known as *apparent* performance, and is potentially optimistic, with performance in new patients and settings often substantially lower (see chapter 1) (34). The apparent discrimination performance of the developed model is given in Table 4.5 (with bootstrap derived 95% confidence intervals); it shows promising discrimination in terms of Harrell's C-statistic (C-statistic=0.68), and Royston and Sauerbrei's D statistic (log hazard ratio = 1.108, hazard ratio=3.03) (75, 212). The R^2_D is only 23%, showing that there is a large amount of variation in the outcome which is not explained by the model (see section 4.3.2). Calibration performance is by definition perfect in the derivation dataset (see chapter 1), and can be assessed visually across risk groups of the linear predictor (LP) as in Figure 4.4 (34, 76, 144, 145, 147).

Table 4.5 - Apparent discrimination performance of the developed model

	Estimate	**95% LCI**	**95% UCI**
C-statistic	0.683	0.671	0.696
D statistic	1.108	0.986	1.222
R^2_D *statistic*	0.227	0.188	0.263

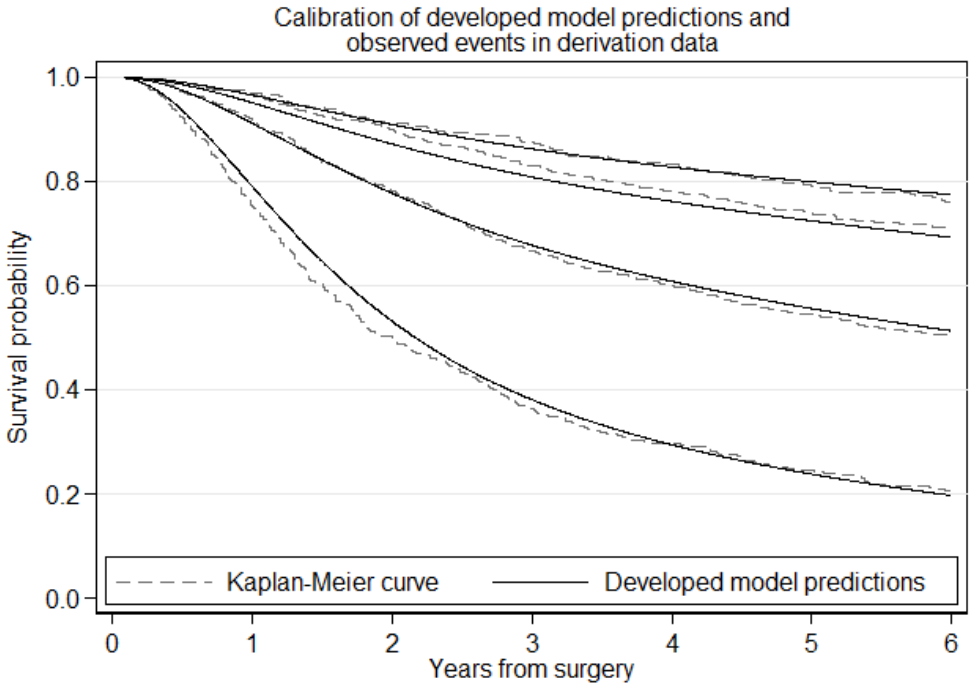


Figure 4.4 - Calibration plot showing apparent performance of the developed model in the Rotterdam derivation data. Dashed lines = KM curve. Solid lines = model predictions.

Ideally the developed model would be used in other patient populations outside of the Rotterdam dataset where applicable, and as such this model requires external validation, which is the focus of the remainder of this chapter.

4.4.3 IPD meta-analysis of external validation performance of original model

The following results present the performance of the developed model in the validation studies (see Table 4.1), using the meta-analysis methods described above with appropriate transformation of performance statistics as required (see Table 4.2). Table 4.6 and Table 4.7 give the summary results from a random effects meta-analysis of the developed model's performance across the seven external validation studies in terms of discrimination and calibration, respectively.

Discrimination performance of the model across the validation studies was moderate, with C-statistics ranging from 0.61 to 0.73 (see Table 4.6). Prediction intervals (PI) from random effects meta-analysis indicate a wide range of possible performance in any new patient populations, showing potentially excellent, but also potentially inadequate discrimination of the model in new patients; for example the 95% PI for the C-statistic ranges from 0.56 to 0.76 (see Table 4.6).

Table 4.6 - Discrimination performance of the developed model when applied to the validation studies. CI – Confidence interval, PI – Prediction interval

Study	C-statistic	R^2_D	D-statistic
Weak performance value	**0.5**	**0**	**0**
Sweden	0.73 (0.68, 0.77)	0.34 (0.23, 0.44)	1.46 (1.12, 1.81)
Lille	0.64 (0.59, 0.69)	0.13 (0.05, 0.23)	0.79 (0.48, 1.11)
Nijmegen	0.66 (0.59, 0.72)	0.18 (0.06, 0.30)	0.94 (0.53, 1.35)
St Cloud	0.65 (0.60, 0.70)	0.15 (0.07, 0.24)	0.86 (0.58, 1.16)
Switzerland	0.72 (0.67, 0.76)	0.30 (0.21, 0.40)	1.36 (1.06, 1.66)
Denmark 1	0.61 (0.57, 0.65)	0.11 (0.04, 0.19)	0.71 (0.42, 1.00)
Denmark 2	0.67 (0.62, 0.71)	0.21 (0.12, 0.32)	1.06 (0.75, 1.40)
Meta-analysis (scale)	**Logit**	**Logit**	**Normal**
Pooled (95% CI)	0.67 (0.63, 0.70)	0.20 (0.14, 0.28)	1.02 (0.80, 1.24)
95% PI	(0.56, 0.76)	(0.06, 0.49)	(0.34, 1.70)
τ^2	0.03	0.22	0.06
I^2 *(%)*	67.75	67.2	68.28

In terms of calibration of the model, Figure 4.5 shows very weak performance of the model in all validation studies, with systematic under prediction of survival probabilities, except in the Denmark studies where the model over predicts survival. This is indicative of heterogeneity in the calibration performance of the model, as seen in Table 4.7, heterogeneity is increasing over time from surgery.

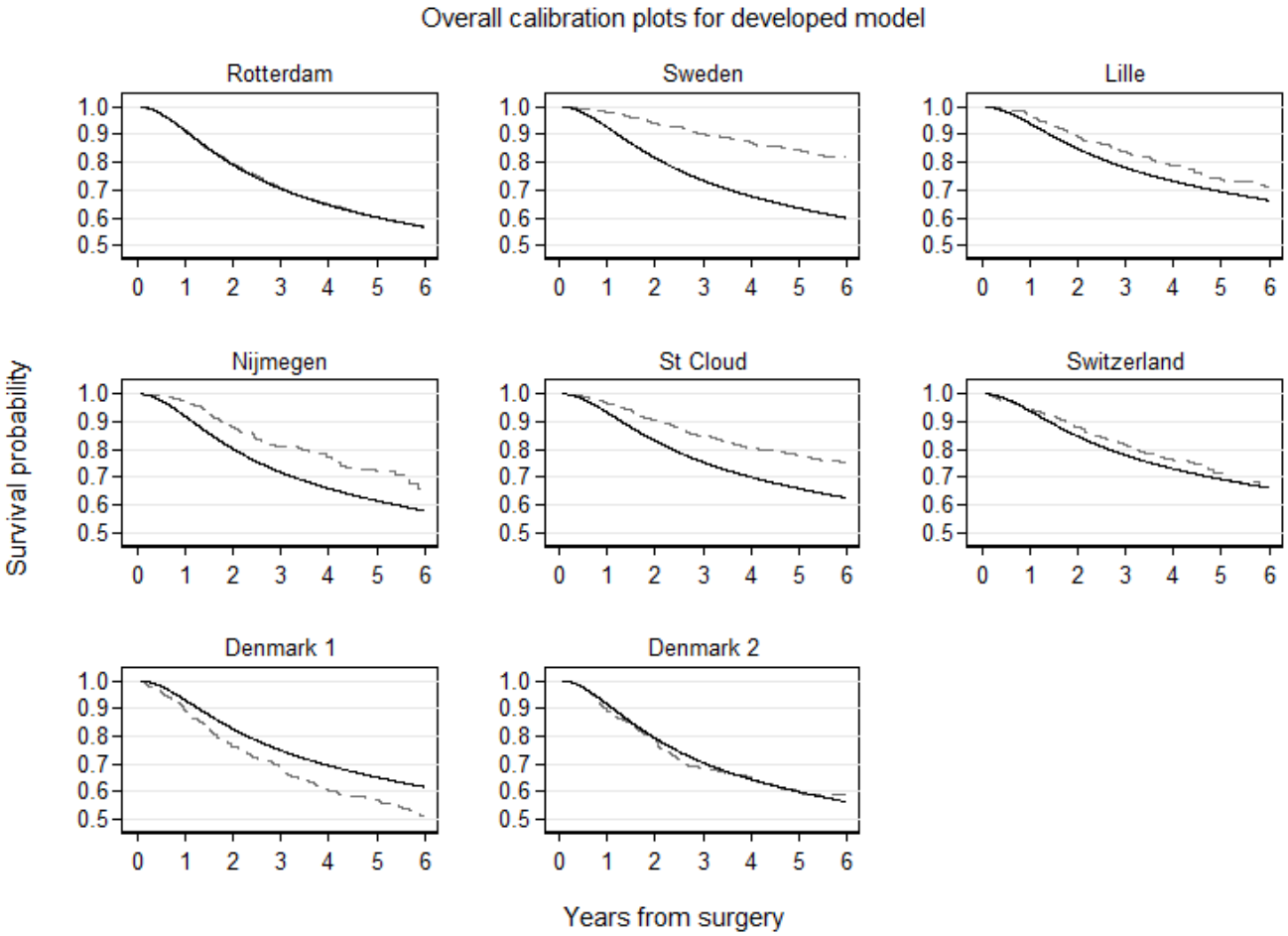


Figure 4.5 - Calibration plot showing performance of the developed model in the seven validation studies. Dashed lines = KM curve of observed survival. Solid lines = model predictions.

Table 4.7 - Calibration performance (E-O & E/O) statistics for the developed model fitted in the validation studies and meta-analysis results (Null value = 0 & 1 respectively).

Study	E-O 6 months	E-O 1 year	E-O 2 year	E-O 3 year	E-O 4 year	E-O 5 year	E-O 6 year
Sweden	-0.01 (-0.02, -0.01)	-0.05 (-0.07, -0.04)	-0.12 (-0.14, -0.11)	-0.17 (-0.19, -0.15)	-0.2 (-0.22, -0.17)	-0.21 (-0.24, -0.18)	-0.22 (-0.25, -0.19)
Lille	-0.01 (-0.02, 0)	-0.03 (-0.04, -0.02)	-0.05 (-0.07, -0.02)	-0.06 (-0.09, -0.02)	-0.06 (-0.09, -0.02)	-0.04 (-0.08, 0)	-0.05 (-0.09, -0.01)
Nijmegen	-0.02 (-0.02, -0.01)	-0.06 (-0.07, -0.04)	-0.08 (-0.11, -0.04)	-0.09 (-0.13, -0.05)	-0.12 (-0.16, -0.07)	-0.11 (-0.16, -0.05)	-0.08 (-0.15, -0.01)
St Cloud	-0.01 (-0.02, 0)	-0.03 (-0.05, -0.02)	-0.08 (-0.1, -0.05)	-0.1 (-0.13, -0.07)	-0.11 (-0.14, -0.07)	-0.12 (-0.15, -0.09)	-0.13 (-0.16, -0.09)
Switzerland	0.01 (0, 0.02)	-0.01 (-0.03, 0.01)	-0.04 (-0.06, -0.01)	-0.04 (-0.07, -0.01)	-0.03 (-0.07, 0.01)	-0.02 (-0.07, 0.02)	-0.01 (-0.06, 0.05)
Denmark 1	0.02 (0, 0.04)	0.03 (0.01, 0.06)	0.06 (0.02, 0.1)	0.06 (0.02, 0.11)	0.09 (0.05, 0.14)	0.08 (0.04, 0.13)	0.11 (0.06, 0.16)
Denmark 2	0 (-0.01, 0.02)	0.02 (-0.01, 0.06)	0 (-0.04, 0.05)	0.02 (-0.03, 0.07)	0 (-0.05, 0.05)	0.01 (-0.04, 0.06)	-0.02 (-0.08, 0.03)
Meta-analysis (Normal scale)							
Pooled	0 (-0.01, 0)	-0.02 (-0.04, 0)	-0.04 (-0.09, 0)	-0.06 (-0.11, 0)	-0.06 (-0.13, 0.01)	-0.06 (-0.14, 0.02)	-0.06 (-0.14, 0.03)
PI	(-0.03, 0.02)	(-0.09, 0.05)	(-0.20, 0.11)	(-0.26, 0.15)	(-0.31, 0.19)	(-0.34, 0.22)	(-0.37, 0.25)
Tau²	0	0	0	0.01	0.01	0.01	0.01
I² (%)	72.58	89.88	93.79	95.22	95.93	96.12	96.16
Study	**E/O 6 months**	**E/O 1 year**	**E/O 2 year**	**E/O 3 year**	**E/O 4 year**	**E/O 5 year**	**E/O 6 year**
Sweden	0.99 (0.98, 0.99)	0.94 (0.93, 0.95)	0.87 (0.85, 0.89)	0.81 (0.79, 0.83)	0.78 (0.75, 0.8)	0.75 (0.73, 0.78)	0.73 (0.71, 0.76)
Lille	0.99 (0.98, 1)	0.97 (0.95, 0.98)	0.95 (0.92, 0.98)	0.93 (0.9, 0.97)	0.93 (0.89, 0.97)	0.94 (0.9, 1)	0.93 (0.88, 0.99)
Nijmegen	0.98 (0.98, 0.99)	0.94 (0.93, 0.96)	0.91 (0.88, 0.95)	0.88 (0.84, 0.93)	0.85 (0.8, 0.9)	0.85 (0.79, 0.92)	0.88 (0.8, 0.98)
St Cloud	0.99 (0.98, 1)	0.97 (0.95, 0.98)	0.92 (0.89, 0.94)	0.89 (0.86, 0.92)	0.87 (0.84, 0.9)	0.84 (0.81, 0.88)	0.83 (0.79, 0.88)
Switzerland	1.01 (1, 1.02)	0.99 (0.97, 1.01)	0.96 (0.93, 0.99)	0.95 (0.92, 0.99)	0.96 (0.92, 1.01)	0.97 (0.91, 1.03)	0.99 (0.92, 1.09)
Denmark 1	1.02 (1, 1.04)	1.04 (1.01, 1.07)	1.08 (1.03, 1.14)	1.09 (1.03, 1.17)	1.15 (1.07, 1.25)	1.14 (1.06, 1.24)	1.21 (1.1, 1.34)
Denmark 2	1 (0.99, 1.03)	1.03 (0.99, 1.07)	1 (0.95, 1.06)	1.03 (0.96, 1.11)	1 (0.93, 1.08)	1.01 (0.93, 1.11)	0.96 (0.88, 1.05)
Meta-analysis (Log scale)							
Pooled	1 (0.99, 1)	0.98 (0.96, 1)	0.95 (0.91, 1)	0.93 (0.87, 1)	0.92 (0.84, 1.01)	0.92 (0.83, 1.02)	0.92 (0.82, 1.04)
PI	(0.97, 1.02)	(0.91, 1.05)	(0.80, 1.13)	(0.73, 1.20)	(0.67, 1.28)	(0.63, 1.35)	(0.60, 1.41)
Tau²	0	0	0	0.01	0.01	0.02	0.02
I² (%)	72.21	89.7	93.81	95.14	95.89	95.94	95.82

4.4.4 IPD meta-analysis of external validation performance after recalibration

The previous section showed that the breast cancer model had large variability in performance across the seven studies used for external validation. Although average performance was good, it would be better to reduce the inconsistency (heterogeneity) in performance and ensure it works well in all settings. The aim of this section is to consider recalibration of the developed model to potentially improve the consistency in calibration performance of the model, and identify the best recalibration strategy using meta-analysis methods.

Recall that recalibration methods evaluated are: re-estimating the magnitude of the baseline hazard (method 1), additionally re-estimating the shape of the baseline hazard (method 2), or additionally estimating a scalar to adjust the linear predictor as a whole in addition to method 1 (method 3). The final method re-estimates the magnitude of the baseline hazard (method 1) and additionally re-estimates the regression coefficient of a single predictor (method 4). Random-effects meta-analysis was used to identify the most heterogeneous predictor effect for re-estimation using method 4. First the developed model was fitted in each validation study separately and all regression coefficients were stored. Next a random-effects meta-analysis was performed on each predictor in turn to quantify the heterogeneity in the predictor effect across the validation studies using Equation 1.19. This process identified the tumour size predictor as having the most heterogeneous predictor effect, with a between-study variance of $\hat{\tau}^2 = 0.242$ and $I^2 = 62.5\%$ (see Figure 4.6).

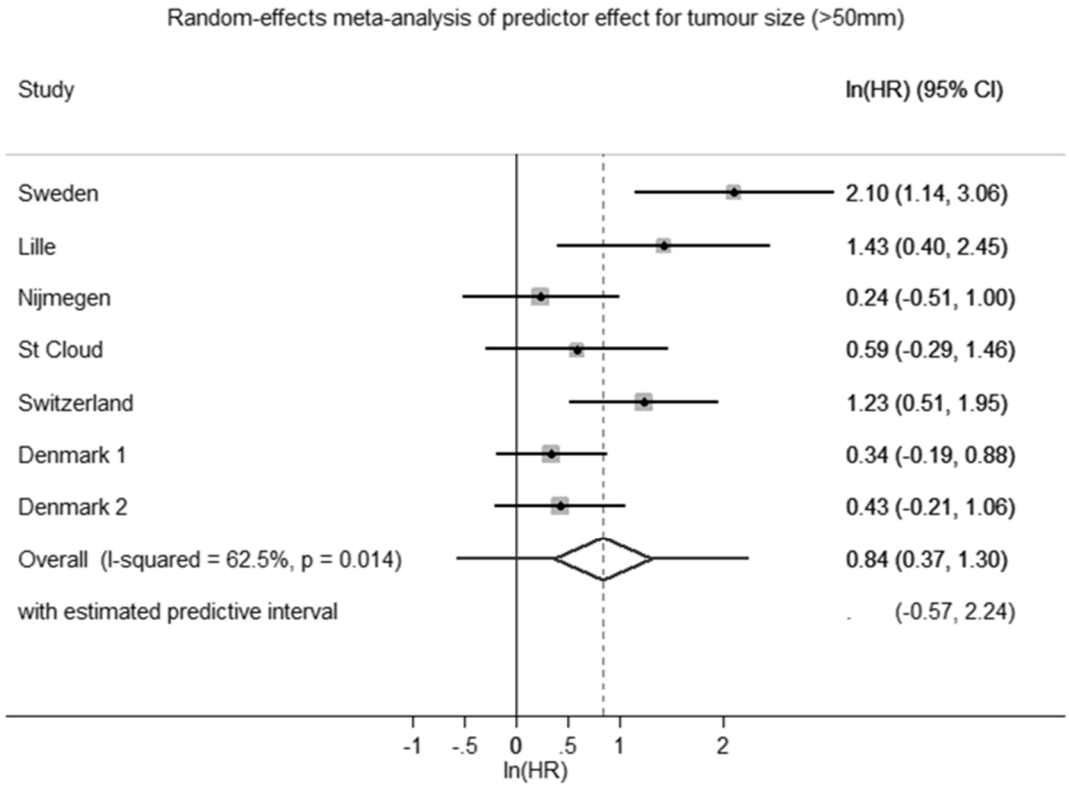


Figure 4.6 - Random-effects meta-analysis of tumour size regression coefficient [ln(HR)] from each validation study.

Impact of recalibration methods on discrimination performance

Table 4.8 shows the performance of the model after using each of the recalibration methods, and additionally repeats the previous summary results relating to the original performance of the developed model for comparison. It is clear that the discrimination performance of the model was unaffected by recalibration methods 1 to 3, as expected given that these did not substantially alter the relative ranking of the linear predictor (69). Discrimination is strongly influenced by the case-mix of the population in which the model is applied; however only method 4 makes any adjustment to the effect of individual predictors, and even then only to one predictor (Tumour size). Harrell's C-statistic varied only by 0.01 in the lower confidence and prediction limit, but did not alter in terms of the pooled performance even after

recalibration with method 4 (see Table 4.8 and Figure 0.51). The D and R^2_D statistics fluctuated across the methods again by only 0.01, showing no influence of the recalibration methods on the discrimination performance of the model. Further, there was no change in the heterogeneity in discrimination performance across validation studies after any recalibration method.

Table 4.8 - Comparison of random effects meta-analysis results for each recalibration method (including both discrimination and calibration performance). CI – Confidence interval, PI – Prediction interval.

Recalibration method	Developed model	Recalibrated model: Method 1	Recalibrated model: Method 2	Recalibrated model: Method 3	Recalibrated model: Method 4
C-statistic					
Pooled effect (CI)	0.67 (0.63, 0.70)	0.67 (0.63, 0.70)	0.67 (0.63, 0.70)	0.67 (0.63, 0.70)	0.67 (0.64, 0.7)
95% PI	(0.56, 0.76)	(0.56, 0.76)	(0.56, 0.76)	(0.56, 0.76)	(0.56, 0.76)
Tau²	0.03	0.03	0.03	0.03	0.03
I² (%)	67.75	67.75	67.75	67.75	66.92
R^2_D					
Pooled effect (CI)	0.20 (0.14, 0.28)	0.19 (0.13, 0.27)	0.20 (0.14, 0.28)	0.19 (0.13, 0.27)	0.20 (0.14, 0.28)
95% PI	(0.06, 0.49)	(0.05, 0.5)	(0.06, 0.48)	(0.05, 0.5)	(0.06, 0.49)
Tau²	0.22	0.27	0.22	0.27	0.22
I² (%)	67.2	70.99	67.2	70.99	68.35
D-statistic					
Pooled effect (CI)	1.02 (0.80, 1.24)	1.01 (0.78, 1.23)	1.02 (0.81, 1.24)	1.01 (0.78, 1.23)	1.04 (0.83, 1.26)
95% PI	(0.34, 1.70)	(0.29, 1.72)	(0.34, 1.7)	(0.29, 1.72)	(0.36, 1.72)
Tau²	0.06	0.07	0.06	0.07	0.06
I² (%)	68.28	71.67	68.28	71.67	69.48
E-O (1 year)					
Pooled effect (CI)	-0.02 (-0.04, 0)	-0.01 (-0.02, 0)	0 (-0.01, 0)	-0.01 (-0.02, 0)	-0.01 (-0.02, 0)
95% PI	(-0.09, 0.05)	(-0.03, 0.01)	(-0.01, 0.01)	(-0.03, 0.01)	(-0.03, 0.01)
Tau²	<0.0001	<0.0001	0	<0.0001	<0.0001
I² (%)	89.88	46.92	0	45.11	46.17
E-O (3 year)					
Pooled effect (CI)	-0.06 (-0.11, 0)	-0.01 (-0.02, 0)	0 (-0.01, 0.01)	-0.01 (-0.02, 0)	-0.01 (-0.02, 0)

Recalibration method	Developed model	Recalibrated model: Method 1	Recalibrated model: Method 2	Recalibrated model: Method 3	Recalibrated model: Method 4
95% PI	(-0.26, 0.15)	(-0.03, 0)	(-0.02, 0.01)	(-0.03, 0)	(-0.03, 0)
Tau²	0.01	0	0	0	0
I² (%)	95.22	0	0	0	0
E-O (6 year)					
Pooled effect (CI)	-0.06 (-0.14, 0.03)	0 (-0.01, 0.02)	0 (-0.02, 0.01)	0 (-0.01, 0.02)	0 (-0.01, 0.02)
95% PI	(-0.37, 0.25)	(-0.02, 0.02)	(-0.02, 0.02)	(-0.02, 0.03)	(-0.02, 0.02)
Tau²	0.01	0	0	0	0
I² (%)	96.16	0	0	0	0

Impact of recalibration methods on calibration performance

Calibration performance was substantially improved through use of each recalibration method investigated, compared to the developed model without recalibration. Each recalibrated model's predictions were extremely close to the observed Kaplan-Meier curve in each study. For example, Figure 4.7 shows the excellent calibration performance of method 1 (shown by the solid lines), in contrast to the large miscalibration for predictions from the originally developed model (short dashed lines).

Expected (E) minus observed (O) statistics represent the difference between these two curves, and results from random effects meta-analysis are presented in Table 4.8. An example forest plot for the E/O ratio statistic measured at 3 years post-surgery using all methods is given in Figure 4.8, and shows excellent improvement after recalibration. The improvement was also observed in the E-O statistic, which is more intuitive to interpret. For example the E-O statistic for the original model in the Sweden validation study is -0.17, which represents a serious under estimation of the risk. This is equivalent to a patients predicted probability of survival being 17% lower than the truth, which could be the difference between a patient being given a particular treatment or not. However after recalibration of the model using method 1, the magnitude of the baseline hazard is increased and the under prediction is reduced to just 2%. Furthermore after recalibration of the whole baseline hazard using method 2, the models' predictions are perfect (E-O=0).

The prediction intervals calculated from the random effects meta-analysis were much narrower after recalibration using methods 1 and 2. A narrower prediction interval indicates that the calibration performance is expected to be consistent across settings or populations

similar to those included in the meta-analysis (see Table 4.8). This reflects very little estimated heterogeneity in calibration performance, with the $\tau^2 = 0.01$ for E-O for the developed model reduced further to zero heterogeneity after use of recalibration methods; similarly the I^2 statistics was reduced from 95% to 0% at 3 years (see Table 4.8).

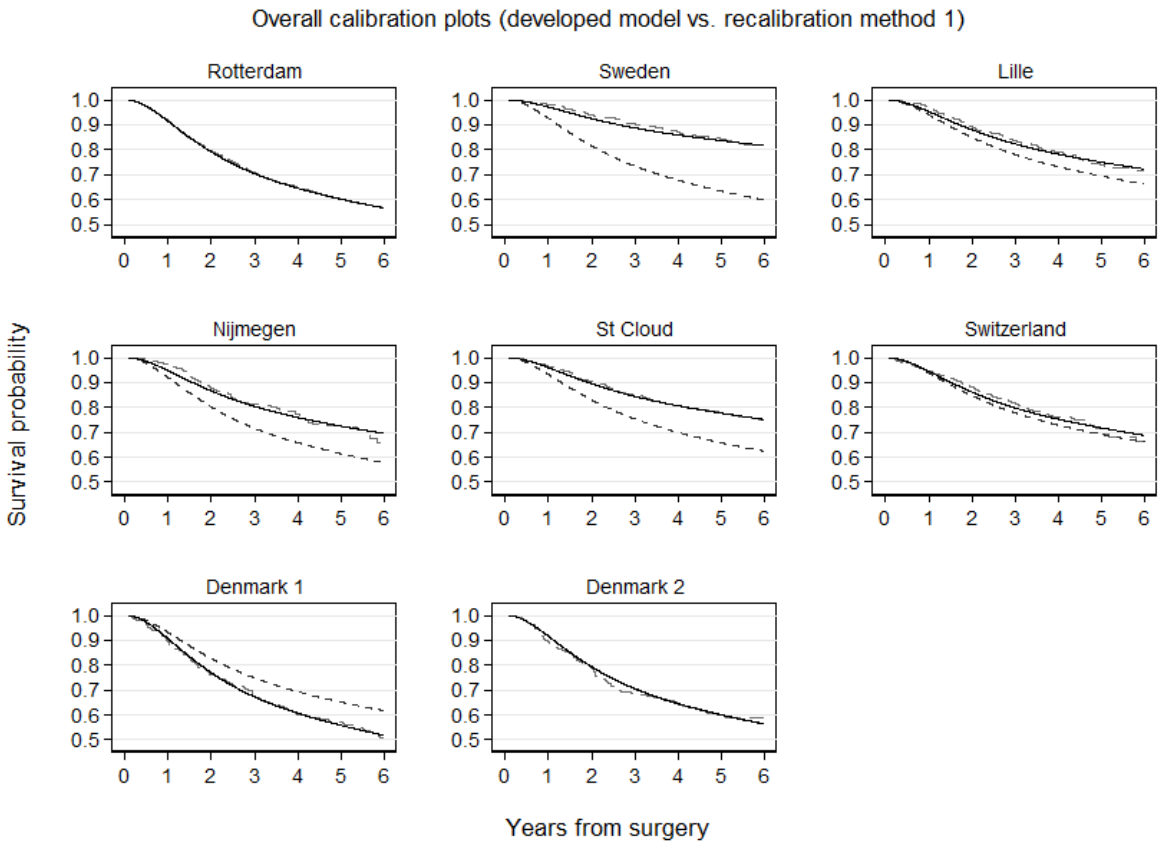


Figure 4.7 - Calibration plot showing performance of the model after recalibration via method 1 compared to the developed model in the seven validation studies. Long dashed lines = KM curve. Solid lines = method 1 model predictions. Short dashed lines = developed model predictions.

Which recalibration method is best?

While all recalibration methods showed distinct improvement compared to the developed model, there was little difference in performance across the methods. For example the E-O at three years was -0.06 for the developed model and, -0.01, 0, -0.01 and -0.01 for methods 1 to 4 respectively (see Table 4.8). The corresponding 95% prediction interval for the developed

model was -0.26 to 0.15, indicating potentially poor performance of the model in new settings. After recalibration using methods 1 to 4, the 95% prediction intervals were narrow ranging from -0.03 to 0.01, showing much greater consistency in calibration performance (see Table 4.8). Given the limited improvements in calibration performance and consistency across validation studies, simple recalibration of the magnitude of the baseline hazard (method 1) appears to be most useful in this applied example. Other examples in different settings would need to be examined to assess which method performs better in general.

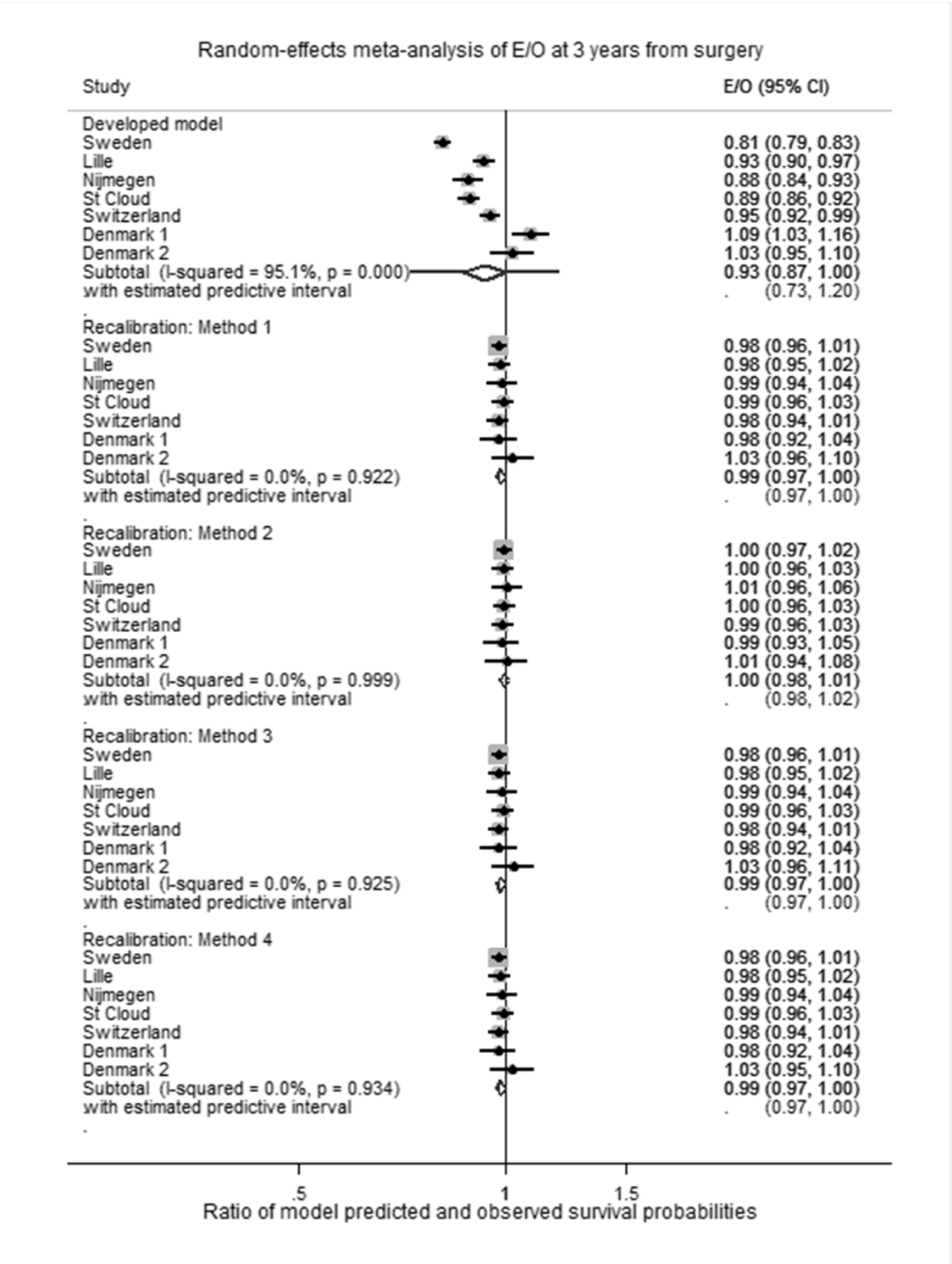


Figure 4.8 - Random effects meta-analysis of calibration performance (E/O at 3 years post-surgery) of the model in all validation studies split by recalibration method. Top panel shows performance of the original model in the validation studies.

4.5 Discussion

Evidence suggests that there is currently much waste in the literature where prognostic models are concerned, with many newly published models but relatively few validations of existing models (9, 82). Development of a new model solely due to poor external performance is perhaps counterintuitive to evidence based medicine, as it is throwing away useful information gleaned from previous patient populations (64, 69). It has been suggested that in this age of large datasets more could be done to perform external validations, and head to head comparisons of existing models, as well as recalibrating existing models to new patient populations (64, 82). Therefore recalibration offers a potential solution to poor performance of existing models in new patients; instead of developing a new model researchers can recalibrate or update existing models, using information from both previous and new patients (64, 69).

This chapter examined the use of recalibration methods for improving the performance of an existing prediction model at external validation, when multiple validation studies are available and the developed model is based on Royston and Parmar's flexible parametric approach. In the applied example, when making predictions using the existing model directly, discrimination performance was moderate, but calibration was poor with substantial over and under prediction of observed event rates as is common in new patient populations, due to changes in case-mix variation and baseline hazard rate (26, 28, 69, 79). Therefore, various recalibration strategies were considered and evaluated using IPD meta-analysis. The findings showed that recalibration of the magnitude of the baseline hazard gave large improvements in the calibration performance of the model. The method is simple, as it only requires re-

estimation of the intercept in the validation study. Similar improvements have been shown through recalibration of the intercept in logistic prediction model examples (69, 142, 146). The other methods investigated (recalibration of the baseline hazard shape, linear predictor, or individual predictors), showed little additional improvement beyond just adjustment of the intercept alone. Between-study heterogeneity in calibration performance was substantial for the original model, but reduced to almost zero after recalibration. Consequently, the 95% prediction intervals were much narrower after recalibration, which highlights that consistency in calibration performance is expected in new populations similar to those in the meta-analysis.

While calibration performance was almost perfect after recalibration of the intercept, discrimination performance was unaltered by such methods. It is well known that adjustment of the baseline risk or hazard does not materially alter the relative rankings of the linear predictor and as such does not impact on model discrimination (69). Only more invasive recalibration or updating methods altering the effect of the predictors in the model, may impact the discriminative performance. As such only method 4, which allowed recalibration of individual heterogeneous predictor effects, altered the discrimination performance of the model at validation in this case study (though not importantly).

One limitation of recalibration methods is the level of information required about the existing model for recalibration to be possible. Here FP models were used because they naturally allow flexibility to investigate the recalibration of various parts of the model; however in practice many published prognostic models are Cox models. This hinders the use of even simple recalibration methods because it is very rare that an estimate of the baseline hazard (or

baseline survival) is published, and this is essential to individualised prediction and recalibration (46, 147). Despite this Royston and Altman have laid out detailed methods for external validation of a Cox model elsewhere (147). For example, even if authors have only presented $S_0(t)$ at a few specific time points, it is possible to make predictions and gain some assessment of calibration performance (though this is rarely reported) (147). In terms of recalibration difficulties remain, though a combination of methods 1 to 3 may be possible provided the linear predictor is adequately reported, though this level of recalibration is essentially a new model development (147).

In the motivating example, IPD was available from seven validation studies, which is perhaps a rare situation. IPD from multiple studies is difficult to obtain, clean and analyse, and the difficulties associated with IPD in general have been widely discussed (ref). However, the availability of IPD is increasing, and similar opportunities arise with big datasets from e-health databases (21). Given large datasets, one could instead simply develop a new model, including study-specific intercepts and, if necessary, study-specific predictor effects to allow for between-study heterogeneity. Previous evidence in logistic prediction models suggests that where the development dataset is relatively large in comparison to the validation dataset, that recalibration of the original model is preferable to new model development (143). While this was not assessed here, further research may look to weigh the benefits of new model development versus recalibration of an existing model. However, it is worth noting that such an approach would neglect the originally developed model, which in the example was itself based on a large dataset. Therefore, it seems sensible to start from the original model, using

recalibration methods as necessary, and then, perhaps as a last resort, develop an entirely new model if recalibration fails.

Further practicalities relate to the use of the model in practice. In this case study recalibration of the intercept provided the best model performance, but this means that for every new population a new intercept must be calculated. This leads to a kind of stratified model, which may be difficult to use in practice if the intended population was different from those used for validation. In that situation one may be forced to use a pooled intercept (e.g. from a random effects meta-analysis of the study-specific intercepts) (206). However such an approach may substantially weaken performance compared to when study-specific intercepts are available.

Importantly any recalibration, in any form could strictly be viewed as a new model, which itself requires external validation in new data, in potentially different patient populations, new settings, different geographical locations or at different times (28, 33, 68, 147). This could mean a never ending cycle of recalibration followed by external validation, with then further recalibration, and so on. Further research of this issue is needed. Ideally this process of external validation and potential recalibration should be performed in the future for the VTE model which was developed in chapter 3.

Previous work has proposed the use of a closed testing procedure to identify the best recalibration method to improve performance while avoiding overfitting for logistic prediction models (226); such an approach could also be evaluated for time-to-event models in further work. Wynants et al. recently investigated the performance of prediction models in clustered data, and recommend the use of conditional model predictions as they perform well at both

the population and centre level in terms of both calibration and discrimination (17). This implies that models using study-specific intercepts may give greatest performance (where interest lies in good study-level performance), as was observed in this case study and previous research (19, 20).

Finally, it is important to note that the conclusions of this work are based on one case study, in one particular cancer, and therefore that results may vary for other examples. To this end, future research could look to extend this work to a simulation study investigating the effect of recalibration methods on external model performance in varying settings.

The next chapter continues to focus on meta-analysis for summarising prognostic performance, but now considers a single predictor and the issue of using aggregate data (rather than IPD).

Key Findings

- Where model performance is weak or heterogeneous at external validation, recalibration methods may substantially improve performance
- A variety of recalibration methods are possible, ranging from changing the intercept to modifying the shape of the baseline hazard and magnitude of predictor effects
- In the case study presented, all recalibration methods showed improvement in the calibration performance of an FP model, but discrimination performance was unaffected in the case study
- Simple recalibration of the intercept alone showed dramatic improvement in the calibration performance of an FP model, and was the method identified as best for this particular case study
- Between-study heterogeneity in model performance was removed after model recalibration using all methods, leading to narrow prediction intervals that indicate consistency in calibration performance was achieved

CHAPTER 5: DEVELOPMENT OF A MULTIPLE IMPUTATION METHOD FOR HANDLING MISSING THRESHOLD RESULTS IN TEST ACCURACY META-ANALYSIS

5.1 Introduction

Previous chapters in this thesis have focused on multivariable prediction models, which contain two or more predictors (prognostic factors) to inform individualised risk prediction. In situations where the outcome to be predicted is short-term (e.g. hypocalcaemia within 48 hours of a thyroidectomy), a single predictor may be sufficient for robust prediction. In particular, if a predictor is associated with something in the body that means the outcome will inevitably become apparent in the near future. Such predictors are sometimes referred to as prognostic tests, and have much similarity to diagnostic tests. Essentially, the test result for each patient is labelled as either positive or negative, with those that test positive considered at high risk of the outcome. The aim is for a prognostic test to have both high sensitivity (probability of being test positive for those that will develop the outcome) and high specificity (probability of being test negative for those that will not develop the outcome). Having fore-knowledge of the likely outcome is important to manage individuals in terms of their care and treatment. For example, for the thyroidectomy patients, a negative test result may allow them to leave the hospital earlier, and thus save beds, whereas those who test positive may be monitored closely or given treatment to manage their calcium levels.

Primary studies that evaluate the accuracy of a continuous test often report sensitivity and specificity values at multiple thresholds which define test positive and test negative patients. However, the choice of threshold(s) used often varies across different primary studies. Therefore, systematic reviews aiming to summarise a continuous test's accuracy across multiple studies are usually limited by heterogeneity in the reported thresholds across studies. This inconsistent presentation of results for multiple thresholds creates a problem for researchers aiming to meta-analyse test accuracy results across multiple studies, for example to establish the best threshold for using the test in practice. Each threshold may have a different number of studies available, and there can be an abundance of information for some thresholds but a scarcity for others. In such cases it is common to meta-analyse the results for each threshold separately, utilising the subset of two by two tables of test accuracy results available for each. Due to the lack of an established and validated alternative method, The Cochrane Handbook for Diagnostic Test Accuracy Reviews (Chapter 10) suggests "Estimating summary sensitivity and specificity of the test for a common threshold, or at each of several different common thresholds" and notes that "Each study can contribute to one or more analyses depending on what thresholds it reports. Studies which do not report at any of the selected thresholds are excluded" (228). However, this approach excludes studies from a meta-analysis if they do not report the threshold of interest, even if they did report results for other (often similar) threshold values. A further concern is selective reporting of threshold results in primary studies, where thresholds are more likely to be reported when they give large sensitivity and specificity results (229). This potentially leads to over-optimistic meta-analysis results, biased toward larger summary sensitivity and specificity results than the truth.

In this chapter a novel statistical method is proposed to deal with missing (partially reported) threshold information in test accuracy meta-analysis, where a single test (predictor) is used to inform the prognosis of an outcome (or condition) for an individual patient. Although the focus of this thesis is on prognosis (future outcomes), the issues and proposed method apply equally to diagnostic tests. Therefore, examples will be given for both prognostic tests (for predicting short-term outcomes) and diagnostic tests (for predicting presence of an existing disease).

Before introducing the method, it is important to highlight existing methods and approaches for dealing with multiple thresholds in test accuracy meta-analysis. Firstly, the multiple thresholds issue could be avoided if individual participant data (IPD) were available, as the continuous test could then be analysed at a consistent threshold in all studies and therefore selective reporting could be avoided. However, in this chapter the focus is on meta-analysis of reported results, that is, where IPD are not available. In this situation, several methods have already been proposed to synthesise results from multiple thresholds simultaneously, but most require a complete set of threshold results (230-232), or require an approximate within-study normality assumption on logit-sensitivity and logit-specificity estimates (with known within-study variances), rather than modelling the two by two data directly (233, 234). Hamza et al. (231) proposed a multivariate random-effect meta-analysis approach which models the within-study relationship between threshold value and test accuracy. An alternative survival model framework for meta-analysing multiple thresholds was later proposed by Putter et al. (232), to counter convergence problems with the Hamza method. However, both these methods require all studies to report all thresholds of interest, and therefore have limited applicability as most applications will encounter missing thresholds in some studies. Dukic and

Gatsonis (230) proposed a method for multiple and missing thresholds, but it only allows a summary ROC (SROC) curve to be derived and does not give summary estimates at each individual threshold of interest. Steinhauser et al. recently proposed a novel method assuming either a normal or logistic distribution for the underlying test or biomarker, and used a linear mixed model to allow estimation of an SROC curve and any desired threshold value (e.g. Youden's index) (235). However, the method requires an assumption of the distribution of the test, which might not be known and may only be estimable if IPD are available.

Riley et al. also proposed a single imputation (SI) method to impute a two by two table for any missing threshold in a study that is bounded between two other available thresholds (236). A meta-analysis can then be done at each threshold, with studies with imputed two by two tables synthesised with studies with known two by two tables. This was proposed as an exploratory method (sensitivity analysis) to examine the potential impact of the missing threshold results on meta-analysis conclusions. The use of imputed results provides more information at each threshold allowing meta-analysis results to be produced with more precision and, in the situation of selectively reported thresholds, potentially less bias. An empirical evaluation was previously undertaken to assess the performance of the SI approach, which showed promising results though no simulations were conducted to compare results to a known truth (236). Further, a concern is that the SI approach provides conservative standard errors for estimates of sensitivity and specificity, as it only includes a single imputed value, which ignores the uncertainty associated with it. In particular, the distance between a missing threshold and its nearest neighbour is ignored; in other words, the smaller the distance

between two known thresholds, the more certain we should be about the imputed results for intermediate thresholds, but this is ignored in the SI approach.

Given these shortcomings, in this chapter a multiple imputation method based on discrete combinations of missing values (MIDC approach) is proposed, to address the potential disadvantages of the SI approach. The proposed method imputes missing two by two tables between two known threshold results similar to the SI approach, but *repeats* this process on multiple occasions, each time using a randomly selected two by two table from the set of all possible discrete combinations of possible missing values. On each occasion, the imputed results are added to the meta-analysis and summary results are estimated using standard methods for each threshold. These multiple sets of summary estimates are then combined using Rubin's rules (56), as in standard multiple imputation applications (56, 57), to give overall summary estimates of sensitivity and specificity for each threshold. In this way the MIDC approach allows for the uncertainty in imputed threshold results and the distance between missing and known results.

The remainder of the chapter is structured as follows. Section 5.2 describes the methods of the SI and MIDC approaches in detail. Stata code is developed and provided in the Appendix, which allows users to quickly apply the SI and MIDC methods to real applications, and thus is potentially useful for Cochrane and researchers conducting test accuracy meta-analyses. Section 5.3 illustrates the approaches using two real examples, and highlight the benefits of the MI and SI approaches over the standard approach of not allowing for missing threshold results. Finally, section 5.4 concludes with some discussion.

This work has been submitted for publication in *Research Synthesis Methods*, and part of the Stata software developed was included within a publication in *Systematic Reviews* (236). The work has also been presented at the statistical conferences including; the Methods for evaluating medical tests and biomarkers symposium (MEMTAB), Birmingham, UK, in July 2016 (237), and the 37th International Society of Clinical Biostatistics Conference (ISCB), Birmingham, UK, in August 2016.

5.2 Methods for single and multiple imputation of missing thresholds

In this section the SI method proposed by Riley et al. is introduced, and then the proposed new MIDC method is outlined in detail, with explanation of accompanying Stata code.

5.2.1 Single Imputation (SI) of missing threshold results

The SI method follows a simple piece-wise linear approach within each study separately, imputing a single value for any missing threshold results that are bounded between two known thresholds in logit receiver operating characteristic (ROC) space (236). It assumes a unit increase in threshold value corresponds to a constant increase (or decrease) in logit-specificity and logit-sensitivity; such a linear relationship is often assumed in meta-analyses that model the relationship across multiple thresholds (231, 234, 235), especially when the number of available thresholds is too small to examine non-linearity.

Considering missing sensitivity results as an example, the following formula is used to impute the missing logit-sensitivity threshold result (y_{ti}) for threshold *t* in study *i*;

$$y_{ti} = y_{i(t-1)} + \left(\left(y_{i(t+1)} - y_{i(t-1)} \right) \times \left(\frac{\left(w_{it} - w_{i(t-1)} \right)}{\left(w_{i(t+1)} - w_{i(t-1)} \right)} \right) \right)$$

Equation 5.1

where $y_{i(t-1)}$ is the observed logit-sensitivity estimate for the nearest reported threshold below t, and $y_{i(t+1)}$ is the observed logit-sensitivity estimate for the nearest reported threshold above t. Weightings, w_{it} are given to each of the observed estimates, depending on how close their respective threshold values are to that of the missing threshold. If the missing threshold value is exactly in the middle of the two bounding thresholds, then the imputed y_{it} is simply the average of $y_{i(t-1)}$ and $y_{i(t+1)}$. Each pair of imputed logit-sensitivity and logit-specificity estimates is then converted back to a two by two table, utilising the known true number of those diseased and non-diseased in the study. It should be noted that the converted two by two table numbers can be non-integer values, and that as such they often need to be rounded to produce a whole number, as otherwise statistical software (usually) will not recognise the data as binomial; to be conservative when using the SI method two by two table numbers are rounded down to the nearest integer.

Meta-analysis is performed for each threshold separately, with any studies giving imputed two by two tables, synthesised along with studies providing an observed two by two table. A standard meta-analysis model is used at each threshold separately, such as the bivariate random-effects approach (238, 239) that follows:

$$TP_{ti} \sim Binomial(D_i, sensitivity_{ti})$$
$$logit(\, sensitivity_{ti}) = \beta_{t1} + u_{ti1}$$
$$TN_{ti} \sim Binomial(ND_i, specificity_{ti})$$
$$logit(\, specificity_{ti}) = \beta_{t2} + u_{ti2}$$
$$\begin{pmatrix} u_{ti1} \\ u_{ti2} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sum_t\right], \sum_t = \begin{pmatrix} \tau_{t1}^2 & \tau_{t1}\tau_{t2}\rho_{t12} \\ \tau_{t1}\tau_{t2}\rho_{t12} & \tau_{t2}^2 \end{pmatrix}$$

Equation 5.2

Here, D_i and ND_i give the number of diseased and non-diseased in study *i*, β_{t1} gives the summary (average) logit-sensitivity and β_{t2} gives the summary logit-specificity at threshold *t*, u_{ti1} and u_{ti2} give the random-effects for logit-sensitivity and logit-specificity, τ_{t1} and τ_{t2} give the between study standard deviations in the true logit-sensitivity and logit-specificity across studies for threshold *t* , and the off-diagonals in $\sum_t$ represent the between-study covariance between the true logit-sensitivity and logit-specificity for threshold *t* . Equation 5.2 collapses to a univariate meta-analysis for each of sensitivity and specificity separately when the between study correlation ρ_{12} is zero. Indeed, as this correlation is typically due to differences across studies in the threshold value (239), when analysing each threshold separately a correlation of zero is quite plausible.

Riley et al. (236) suggested the SI approach as a simple exploratory method, to allow meta-analysis results after imputation to be compared with those from a standard analysis that analyses each threshold separately with no imputation (NI) (i.e. Equation 5.2 applied to each threshold separately, and thus studies with missing results for a particular threshold are excluded for that threshold's meta-analysis).

5.2.2 Multiple imputation of missing threshold results based on discrete combinations (MIDC)

The new MIDC method is now introduced, and a schematic of the approach is provided in Figure 5.2. It follows four key steps, as now described.

Step 1: Identification and random selection of discrete combinations for imputing values for missing thresholds in each study separately

For each missing threshold bounded between two known thresholds, the MIDC method recognises that the missing sensitivity and specificity must lie within a rectangle formed between the two nearest known threshold results (see Figure 5.1). Further, the missing number of true positives (TPs), false negatives (FNs), true negatives (TNs) and false positives (FPs) at the missing threshold must be whole numbers of patients within this quadrilateral between those known for the neighbouring thresholds.

Table 5.1 shows an illustrative example of a study reporting test accuracy results for a continuous test where there are 39 true diseased and 146 true non-diseased patients. Assume that the thresholds of interest for meta-analysis are 1, 2, 3, 4 and 5, but that this particular study only reports two by two tables for thresholds 1 and 5. Therefore, there are three missing thresholds (r=3) to be imputed by the MIDC method, and it is clear that the missing TP must be ≤ 35 and ≥ 30, the missing FN must be ≤ 9 and ≥ 4, the missing TN must be ≤ 104 and ≥ 95, and finally the missing FP must be ≤ 51 and ≥ 42. Therefore, there are six potential values (n=6) for the TP in the missing thresholds (i.e. 35, 34, 33, 32, 31, 30), and ten potential values for the TN in the missing thresholds (i.e. 104, 103 … 96, 95). It follows naturally that imputing TP also defines FN (or vice-versa) given the total diseased, and imputing TN also defines FP (or

vice-versa), and therefore we only need to consider imputing two cells in the missing two by two table; for example, TP and TN.

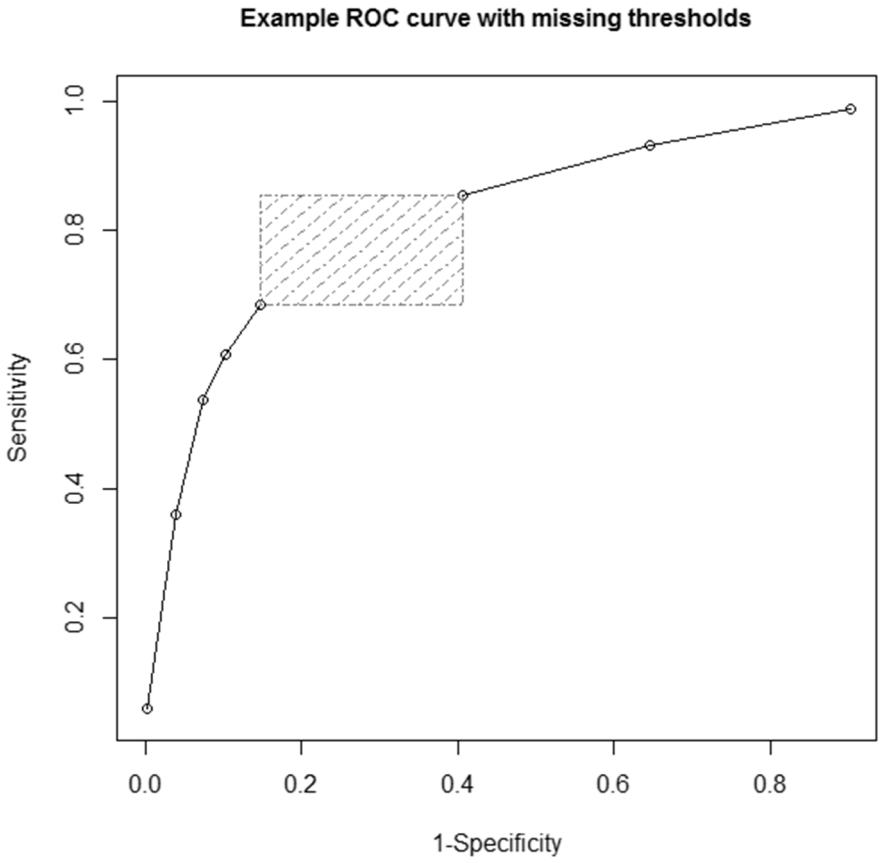


Figure 5.1 - Illustrative ROC curve with missing threshold results bounded within the rectangle

Table 5.1 – Example data for a single study reporting a continuous test measured at a partial set of multiple thresholds of interest for meta-analysis

Threshold	Missing	TP	FN	TN	FP
1	No	35	4	95	51
2	Yes	?	?	?	?
3	Yes	?	?	?	?
4	Yes	?	?	?	?
5	No	30	9	104	42

First focus on imputing TP (and thus FN) for the three missing thresholds. There are 56 possible discrete combinations (of three values) for the TP to be imputed for the three missing thresholds, taking into account that the missing TP is bounded between 35 and 30 and that as threshold value increases the number of TPs must be equal to or less than the number of TPs at the previous threshold (see Table 5.2 for illustration). The MIDC method randomly selects one of these possible combinations, from the list of all combinations, assuming that all possible combinations are equally likely (see Figure 5.1). In the same manner, a value for the missing TN (and FP) can be imputed for the same missing thresholds.

Table 5.2 – First and last five of the 56 possible combinations of the imputed TP values for thresholds 2, 3 and 4 in Table 5.1

First and last five of 56 combinations with repetition for imputed TP values (n=6, r=3)			
Discrete combination number	**Threshold 2**	**Threshold 3**	**Threshold 4**
1	35	35	35
2	35	35	34
3	35	35	33
4	35	35	32
5	35	35	31
….	….	….	….
52	32	30	30
53	31	31	31
54	31	31	30
55	31	30	30
56	30	30	30

The above describes a single imputation using the MIDC method in a single study for a particular set of missing thresholds. In each study separately, this approach can be used to impute TP, FN, TN and FP for all missing but bounded thresholds. Note that, as in the SI method, in a particular study there is no imputation for those thresholds that are not bounded

(i.e. those missing thresholds that fall above the largest reported threshold or below the smallest reported threshold).

Step 2: Meta-analysis of imputed and observed two by two tables

In step 2 a meta-analysis is now applied to each threshold separately, including the imputed and observed two by two tables available from the studies. For example, Equation 5.2 can be applied to produce summary logit-sensitivity and logit-specificity estimates, and between-study heterogeneity estimates.

Step 3: Generate multiply imputed datasets and multiple meta-analysis results

Steps 1 and 2 are then repeated '*M*' times, leading to '*M*' meta-analysis results (one for each cycle). As the imputation procedure generates new imputations in each cycle, the subsequent meta-analysis results may also be different for each cycle.

Step 4: Combine across the multiple meta-analysis results using Rubin's rules

The meta-analysis estimates obtained from each of the '*M*' cycles are then combined using Rubin's rules, as commonly applied in traditional multiple imputation methods and detailed in full elsewhere (56). Briefly, Rubin's rules produce an overall meta-analysis estimate by averaging the `M' summary estimates, and calculates an appropriate standard error by accounting for both within and between-imputation variance of the estimates (56, 57). This provides final meta-analysis estimates for each parameter of interest and its associated standard error (and thus 95% CI).

Multiple Imputation using Discrete Combinations (MIDC) approach schematic

STEP 1: In each study separately;

- Identify all missing thresholds of interest that are bounded by two thresholds for which two by two tables are available
- For each set of missing thresholds contained within a bound, derive the set of all discrete combinations for the missing two by two tables
- Randomly select a discrete combination from the set of all combinations, thereby imputing a single two by two table for all missing but bounded thresholds

STEP 2: For each threshold separately, apply a meta-analysis (e.g. model (2)) to combine the imputed and observed two by two tables from all available studies, to produce one set of meta-analysis results for each threshold

STEP 3: Repeat steps 1 and 2 a total of *M* times, to obtain *M* sets of meta-analysis estimates for each threshold

STEP 4: Use Rubin's rules to combine the *M* meta-analysis results for each threshold separately, to produce a final estimate and standard error for each parameter in the meta-analysis model.

Figure 5.2 - Schematic of the multiple imputation using discrete combinations (MIDC) method

5.2.3 Potential advantages of the MIDC method over the SI method

By repeated random sampling from the set of all possible combinations of the missing two by two tables, the MIDC approach has three potential advantages over the SI method. Firstly, by considering multiply imputed datasets, it accounts for the uncertainty of the imputed two by two tables. Secondly, as it imputes the two by two tables directly, this ensures that all imputed values are whole numbers, which is not necessarily the case with the SI method. Thirdly, the method allows for the distance between known and missing threshold results, such that it is more likely that a missing threshold close to the known threshold will take a TP (or TN) value similar to the observed TP (or TN) value for the closest known threshold. This is illustrated for

the current example in Table 3, where the probability of threshold 2 taking each of the six possible numbers of TPs is given; it is clear that it is most likely that a TP of 35 will be imputed for threshold 2, which is the observed TP value in the closest neighbouring threshold 1. This is simply because there are more combinations that include a TP of 35 than any other value.

Table 5.3 – Probability of each TP value being imputed for missing threshold 2, which is bounded between 35 from threshold 1 and 30 from threshold 5

Possible TP value	Probability of TP value being imputed for threshold 2
35	0.375 (=21/56)
34	0.268 (=15/56)
33	0.179 (=10/56)
32	0.107 (=6/56)
31	0.054 (=3/56)
30	0.018 (=1/56)

5.3 Software to implement the methods

Software was developed (by the PhD candidate, Joie Ensor) to implement the MIDC and SI methods within Stata. The complete code for the MIDC approach is presented in the appendix (see APPENDIX D: Chapter 5 Appendices), and as part of a submitted article to *Research Synthesis Methods*. The software to implement the SI approach has been published as part of the publication proposing the SI method (236). The program for the MIDC approach follows the schematic laid out in Figure 5.2. In particular, each possible discrete combination has a unique combination number, which follows the pattern shown in Table 5.2. For each imputation the discrete combination number is chosen randomly from a uniform distribution ranging from 1 to the total number of possible combinations. The corresponding combination is then identified, by exploiting the relationship between the unique combination number and

cumulative sums of squares, which can be used to calculate the value at each missing threshold separately.

5.3.1 MIDC Stata module

The following sections briefly describe the user written Stata module `midc`, which applies the MIDC method as described above (see section 5.2.2).

Command line syntax

`midc, studyvar(varname) thresholdvar(varname) imputations(int) sensitivity(string) specificity(string) covariance(string)`

Where the data are arranged with one line per threshold for each study, where thresholds are identified by a variable named in `thresholdvar()`, and studies are identified by a variable named in `studyvar()`.

Options

`studyvar(varname)` specifies the name of a variable identifying data relating to each study (Note this may be in any format including numeric, string or factor variables).

`thresholdvar(varname)` specifies the name of a variable identifying data relating to each threshold (Note this may be in any format including numeric, string or factor variables).

`imputations(int)` specifies the number of imputation datasets to be generated, meta-analysed, and combined using Rubin's rules.

`sensitivity(string)` either specifies the name of a variable containing the sensitivity if this data is available, or specifies the name of a new variable to contain the computed sensitivity.

`specificity(string)` either specifies the name of a variable containing the specificity if this data is available, or specifies the name of a new variable to contain the computed specificity.

`covariance(string)` variance-covariance structure of the random effects, where options include;

- `independent` one variance parameter per random effect, all covariances zero
- `exchangeable` equal variances for random effects, and one common pairwise covariance
- `identity` equal variances for random effects, all covariances zero
- `unstructured` all variances-covariances distinctly estimated

Data format

For each threshold value within each study, data for `midc` must consist of true positive (tp), false positive (fp), true negative (tn) and false negative (fn) values which form the traditional 2x2 table format commonly seen in diagnostic test accuracy studies. The program can accept datasets with or without missing rows of threshold information; `midc` prepares a dataset for

imputation by creating the missing threshold rows using the user input study and threshold identifying variables. The following shows a printout of the first five rows of a typical example dataset accepted by the `midc` command.

```
list study threshold tp fp fn tn in 1/5, noo clean
```

study	thresh~d	tp	fp	fn	tn
Apgar	0	19	4	292	2107
Apgar	1	112	83	199	2028
Apgar	2	167	157	144	1954
Apgar	3	189	218	122	1893
Apgar	4	213	311	98	1800

Model estimation

For the MIDC method any number of imputation datasets can be specified within the command line (`imputations` option), though it should be noted that computation time will increase with increasing numbers of imputation datasets. Equation 5.2 is then used to synthesise the available results for each threshold separately, allowing for estimation of the between-study correlation using the previously described options (`covariance` option). It should be noted that fixing the between-study correlation to zero (`independent` option) avoids common computation issues associated with this parameter (240, 241). The model is fitted via maximum likelihood estimation using Gauss-Hermite quadrature with quadrature points equal to the number of studies in the meta-analysis (up to a maximum of 5 quadrature points). Following meta-analysis, Rubin's rules are applied to combine the meta-analysis results from each imputation dataset, giving one meta-analysis result for each threshold. All

95% confidence intervals are derived on the logit scale using the summary estimates combined from all imputation datasets using Rubin's rules, and then back-transformed.

Example Stata output

The following shows an example of the command line and output for the `midc` program as used in the first applied example described later.

```
midc, st(studyname) th(thresholdid) sens(sens) spec(spec)
cov(independent) imp(5)

Note: Number of imputation datasets = 5

Note: Bivariate random-effects model fitted with covariance
structure = independent

Imputing dataset 1
..................................................

Total thresholds imputed = 54

Imputing dataset 2
..................................................

Total thresholds imputed = 54
```

(Output omitted)

```
Imputing dataset 5
..................................................

Total thresholds imputed = 54


Program took 6.061 seconds, for each imputed dataset

Results are stored in memory
```

5.4 Applied examples

In order to illustrate the use of the imputation methods in practice applied examples are now presented, one for a diagnostic test and one for a prognostic test.

5.4.1 Protein/Creatinine ratio (PCR) for the detection of significant proteinuria in patients with suspected pre-eclampsia

The first example is based on a systematic review and meta-analysis investigating the performance of Protein/Creatinine ratio (PCR) as a diagnostic test for the detection of significant proteinuria in patients with suspected pre-eclampsia (30). This review is an ideal situation in which to apply the imputation methodology proposed, as it found 13 studies which reported various possible thresholds for PCR, with each study reporting on a different set of thresholds. This made the original meta-analysis problematic, due to small numbers of studies reporting on any one threshold. In total there were 23 thresholds considered across all 13 studies, with five studies reporting only one threshold, and the largest meta-analysis possible containing only seven studies. The studies and two by two tables for each threshold are presented in Table 5.4 (236).

Table 5.4 - PCR data at each threshold for the 13 studies identified in Morris et al.

First Author	Threshold ID	Threshold value	TP	FP	FN	TN	Total	High proteinuria	Normal proteinuria
Al Ragip	1	0.13	35	51	4	95	185	39	146
	6	0.18	33	42	6	104			
	7	0.19	33	39	6	107			
	8	0.2	31	38	8	108			
	22	0.49	29	23	10	123			
Durnwald	3	0.15	156	35	12	17	220	168	52
	8	0.2	152	27	16	25			
	15	0.3	136	23	32	29			
	19	0.39	123	14	45	38			
	20	0.4	120	12	48	40			
	23	0.5	106	9	62	43			
Dwyer	3	0.15	54	28	2	32	116	56	60
	5	0.17	51	25	5	35			
	7	0.19	50	18	6	42			
	12	0.24	41	8	15	52			
	14	0.28	37	3	19	57			
	19	0.39	31	0	25	60			
Leonas	15	0.3	277	7	5	638	927	282	645
Ramos	23	0.5	25	1	1	20	47	26	21
Robert	15	0.3	27	4	2	38	71	29	42
Rodriguez	2	0.14	69	34	0	35	138	69	69
	3	0.15	68	34	1	35			
	4	0.16	68	26	1	43			
	5	0.17	65	25	4	44			
	6	0.18	62	24	7	45			
	7	0.19	62	21	7	48			

First Author	Threshold ID	Threshold value	TP	FP	FN	TN	Total	High proteinuria	Normal proteinuria
	8	0.2	60	19	9	50			
	9	0.21	60	17	9	52			
Saudan	8	0.2	14	27	0	59	100	14	86
	13	0.25	13	14	1	72			
	15	0.3	13	7	1	79			
	18	0.35	12	4	2	82			
	20	0.4	11	3	3	83			
	21	0.45	10	0	4	86			
Schubert	3	0.15	9	3	0	3	15	9	6
	4	0.16	9	2	0	4			
Shahbazian	8	0.2	35	2	3	41	81	38	43
Taherian	2	0.14	67	7	6	20	100	73	27
	3	0.15	67	3	6	24			
	4	0.16	65	1	8	26			
	5	0.17	64	1	9	26			
	6	0.18	63	0	10	27			
	8	0.2	59	0	14	27			
Wheeler	9	0.21	59	13	9	45	126	68	58
Yamasmit	7	0.19	29	6	0	7	42	29	13
	9	0.21	29	5	0	8			
	10	0.22	29	4	0	9			
	11	0.23	28	3	1	10			
	12	0.24	28	2	1	11			
	13	0.25	28	1	1	12			
	14	0.28	27	1	2	12			
	16	0.31	26	1	3	12			
	17	0.32	25	1	4	12			

The NI method was applied to each threshold using Equation 5.2 with maximum likelihood estimation. This reflects what happens with current meta-analyses of test accuracy studies: each threshold is analysed separately with no imputation, and thus only uses studies which report the results for that threshold. The results are shown in Table 5.5 and Table 5.6, for sensitivity and specificity respectively.

The MIDC and SI methods were also applied using Equation 5.2 with parameter estimation via maximum likelihood, but with between-study correlation set to zero to avoid the common computation issues associated with this parameter (15, 16). For the MIDC method five imputation datasets were performed before Rubin's rules was applied (giving one meta-analysis results dataset), five was selected to reduce computation time. The imputation methods summarise not only the published evidence available for each threshold but also utilise imputed data for the missing, unpublished evidence (for bounded thresholds). As such the methods can be used to assess if the conclusions of the original meta-analysis are robust (e.g. to potential selective threshold reporting bias) for each threshold. The following presents the command line for the `midc` program and output from Stata for this example.

Table 5.5 - PCR example sensitivity results for all methods, including the summary sensitivity, its standard error and the number of studies reporting the threshold

	Without imputed data (NI)			With single imputed data (SI)			With multiple imputed data (MIDC)		
Threshold	No. studies	Summary Sensitivity	Standard Error (logit scale)	No. studies	Summary Sensitivity	Standard Error (logit scale)	No. studies	Summary Sensitivity	Standard Error (logit scale)
1	1	0.897	0.528	1	0.897	0.528	1	0.897	0.528
2	2	0.982	1.684	3	0.957	0.759	3	0.956	0.784
3	5	0.944	0.316	6	0.939	0.206	6	0.938	0.243
4	3	0.960	0.808	6	0.925	0.214	6	0.933	0.276
5	3	0.909	0.247	5	0.911	0.175	5	0.910	0.174
6	3	0.873	0.223	5	0.894	0.161	5	0.894	0.162
7	4	0.902	0.242	6	0.889	0.160	6	0.895	0.175
8	6	0.875	0.188	8	0.886	0.206	8	0.884	0.218
9	3	0.892	0.250	7	0.880	0.146	7	0.881	0.178
10	1	0.983	1.451	5	0.901	0.562	5	0.891	0.523
11	1	0.966	1.018	5	0.868	0.335	5	0.855	0.304
12	2	0.877	0.844	5	0.844	0.295	5	0.849	0.344
13	2	0.953	0.724	5	0.838	0.336	5	0.836	0.339
14	2	0.818	0.701	5	0.816	0.296	5	0.813	0.296
15	4	0.938	0.583	7	0.887	0.451	7	0.889	0.459
16	1	0.897	0.610	5	0.782	0.287	5	0.782	0.281
17	1	0.862	0.539	5	0.761	0.261	5	0.759	0.239
18	1	0.857	0.764	4	0.715	0.214	4	0.725	0.217
19	2	0.662	0.281	4	0.695	0.201	4	0.704	0.222
20	2	0.720	0.165	3	0.724	0.150	3	0.727	0.151
21	1	0.700	0.583	3	0.697	0.146	3	0.705	0.148
22	1	0.744	0.367	2	0.675	0.149	2	0.667	0.147
23	2	0.844	0.996	2	0.844	0.996	2	0.844	0.996

Table 5.6 - PCR example specificity results for all methods, including the summary specificity, its standard error and the number of studies reporting the threshold

	Without imputed data (NI)			With single imputed data (SI)			With multiple imputed data (MIDC)		
Threshold	No. studies	Summary Specificity	Standard error (logit scale)	No. studies	Summary Specificity	Standard error (logit scale)	No. studies	Summary Specificity	Standard error (logit scale)
1	1	0.651	0.174	1	0.651	0.174	1	0.651	0.174
2	2	0.604	0.363	3	0.624	0.208	3	0.631	0.224
3	5	0.562	0.407	6	0.583	0.332	6	0.586	0.336
4	3	0.803	0.719	6	0.660	0.420	6	0.664	0.416
5	3	0.765	0.678	5	0.677	0.448	5	0.678	0.441
6	3	0.856	1.043	5	0.728	0.583	5	0.730	0.554
7	4	0.708	0.130	6	0.720	0.437	6	0.722	0.424
8	6	0.818	0.569	8	0.775	0.405	8	0.777	0.398
9	3	0.750	0.195	7	0.710	0.170	7	0.709	0.166
10	1	0.679	0.594	5	0.712	0.231	5	0.717	0.239
11	1	0.769	0.658	5	0.744	0.277	5	0.744	0.239
12	2	0.863	0.340	5	0.768	0.296	5	0.770	0.286
13	2	0.848	0.280	5	0.811	0.397	5	0.821	0.406
14	2	0.945	0.514	5	0.849	0.474	5	0.845	0.453
15	4	0.917	0.786	7	0.914	0.515	7	0.914	0.520
16	1	0.923	1.041	5	0.890	0.579	5	0.874	0.509
17	1	0.923	1.041	5	0.897	0.568	5	0.892	0.556
18	1	0.953	0.512	4	0.903	0.643	4	0.917	0.771
19	2	0.976	2.598	4	0.935	0.875	4	0.932	0.852
20	2	0.903	0.766	3	0.877	0.491	3	0.878	0.490
21	1	0.994	1.427	3	0.943	1.265	3	0.943	1.263
22	1	0.842	0.227	2	0.833	0.191	2	0.832	0.190
23	2	0.863	0.340	2	0.863	0.340	2	0.863	0.340

Figure 5.3 shows a dramatic change in summary sensitivity and specificity (in ROC space) when using the MIDC method compared to the original NI approach. For example for threshold 13, the summary sensitivity was 0.953 and 0.836 for the NI and MIDC methods, respectively, with associated standard errors of 0.724 and 0.339. After imputation it is clear that there is generally a shift in summary estimates, both downward and to the right in ROC space. This indicates a decrease in sensitivity (downward shift) and specificity (shift right), as may be expected in the presence of publication bias or selective reporting, where weaker performing threshold results may not be reported. Therefore, the application of the MIDC method reveals that the original conclusions from the NI method are not robust, with summary sensitivity and specificity estimates lower than estimated when ignoring missing data.

There are comparatively small differences between the MIDC and SI methods, with both indicating that summary estimates shift in the same direction. The MIDC method shifts the summary estimates by a slightly greater magnitude compared to the SI method, and this difference is likely caused by the MIDC method additionally accounting for the uncertainty in imputed threshold results and/or providing two by two tables that do not require rounding, unlike the SI approach. Taking threshold 13 as an example again, the summary sensitivity for the SI and MIDC methods was 0.838 and 0.836, with standard errors of 0.336 and 0.339, respectively.

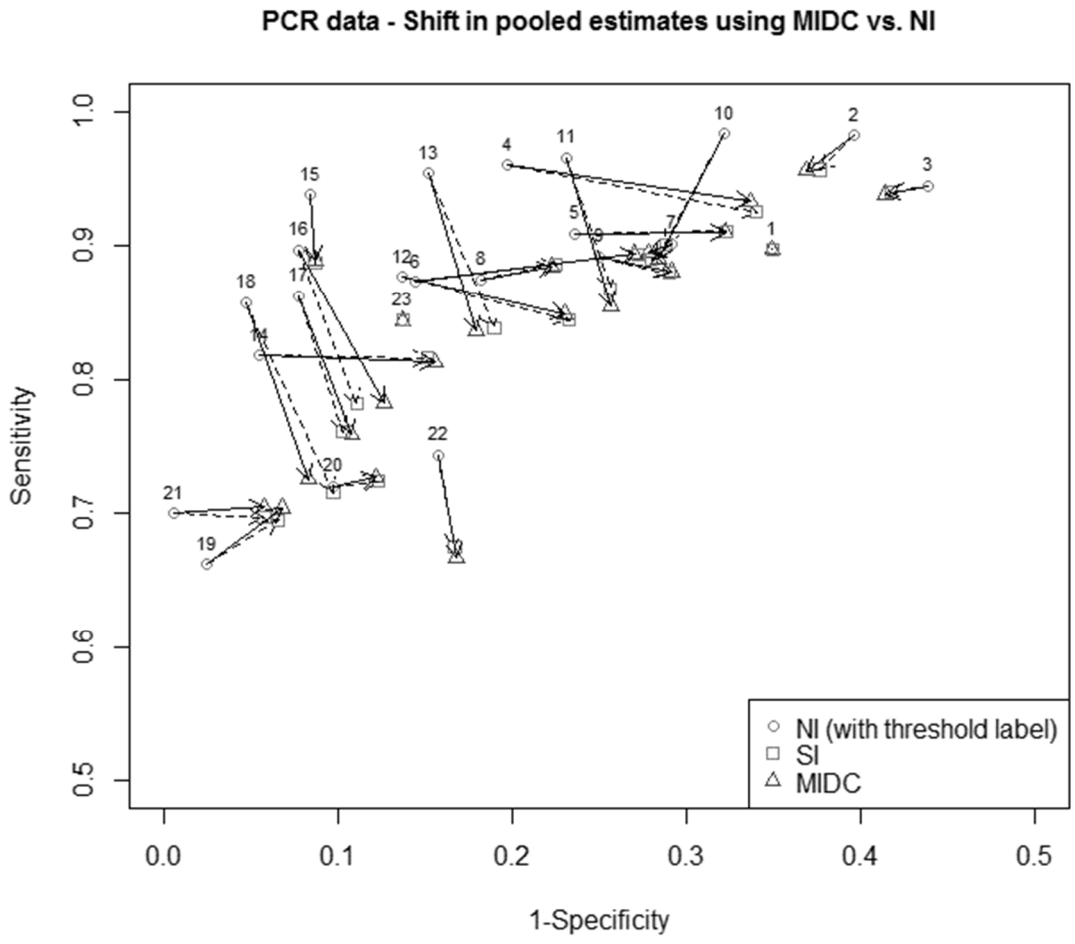


Figure 5.3 - Summary estimates of sensitivity and specificity in ROC space for all methods. NB: Arrows represent change from NI summary estimates

Figure 5.4 presents the standard errors of sensitivity and specificity at each threshold, showing the shift in standard error from the NI to MIDC method. Using the imputation approaches 54 additional threshold results were gained for meta-analysis, which reduced the standard errors of the summary sensitivity and specificity at many thresholds (by as much as 70% in some cases, see Table 5.5 and Table 5.6), also leading to substantially narrower confidence intervals.

The SI approach performed similarly to the MIDC method, with substantially smaller standard errors that the NI approach (see Figure 5.4).

Thus, in summary, the use of more evidence through imputation via either the SI or MIDC methods had a substantial impact on the summary results for the PCR tests at each threshold. Test accuracy results are not as promising for the PCR test after imputation, which was missed when using the NI approach. Further, in this particular example, there appears to be very little difference between the MIDC and SI approaches.

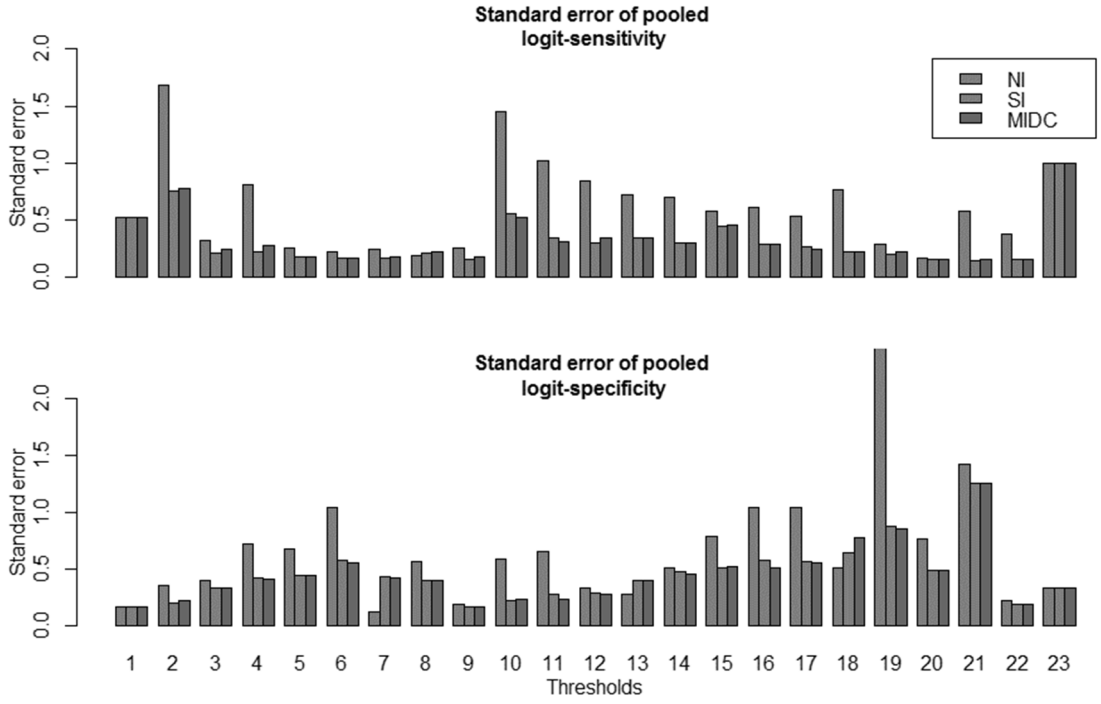


Figure 5.4 – Standard errors of sensitivity and specificity for the PCR dataset using NI and MIDC methods

5.4.2 Apgar score to assess the health of newborn children

The second example is based on a systematic review and meta-analysis investigating the performance of the Apgar score as a prognostic test for neonatal mortality in pre-term (<37 weeks gestation) babies with a low birth weight (<2.5kg) (242). The Apgar score is essentially a prognostic model, as it is formed by a combination of prognostic factors used to measure a newborn baby's health (243); it takes values from 0 to 10 and is usually measured at one, five and ten minutes after birth (244). Babies with low values have a worse prognosis. The review identified 11 studies which examined the prognostic ability of the Apgar score in terms of its sensitivity and specificity after using a threshold to dichotomise the score into high and low risk groups. However, the studies used various possible thresholds for dichotomising the Apgar score. In total across the 11 studies there were five studies reporting only one threshold, and the largest number of studies for a particular threshold was nine. The studies and two by two tables for each threshold are presented in Table 5.7.

Table 5.7 - Apgar data for all thresholds for the 11 studies identified in Malin et al. (242).

First Author	Threshold value	TP	FP	FN	TN	Total	TD	TND
Apgar	0	19	4	292	2107	2422	311	2111
	1	112	83	199	2028	2422	311	2111
	2	167	157	144	1954	2422	311	2111
	3	189	218	122	1893	2422	311	2111
	4	213	311	98	1800	2422	311	2111
	5	237	439	74	1672	2422	311	2111
	6	253	606	58	1505	2422	311	2111
	7	266	859	45	1252	2422	311	2111
	8	290	1365	21	746	2422	311	2111
	9	307	1906	4	205	2422	311	2111
Beeby	3	49	135	39	400	623	88	535
Behnke	3	113	144	48	443	748	161	587
	6	148	326	13	261	748	161	587
Drage	3	101	144	50	1322	1617	151	1466

First Author	Threshold value	TP	FP	FN	TN	Total	TD	TND
	6	128	427	23	1039	1617	151	1466
Heller	3	97	1267	91	31106	32561	188	32373
	6	147	4916	41	27457	32561	188	32373
Ikonen	3	35	25	65	443	568	100	468
	6	55	70	45	398	568	100	468
Issel	3	20	45	52	585	702	72	630
	7	52	169	20	461	702	72	630
Kato	4	6	75	0	147	228	6	222
Luthy	3	26	47	9	164	246	35	211
Serenius	3	28	30	45	108	211	73	138
Tejani	6	47	142	6	197	392	53	339

The NI, SI and MIDC methods were applied as before for the PCR example, where the meta-analysis model given in Equation 5.2 was estimated using maximum likelihood for each threshold of interest, with between-study correlation again set to zero to avoid computational issues (15, 16). For the MIDC method five imputation datasets were selected as for the PCR example, with Rubin's rules used to combine the five imputed datasets into one final meta-analysis results dataset. The results are shown in Table 5.8, for sensitivity and specificity respectively.

Results showed that imputation was only possible in a few instances in the central thresholds; in particular, many studies reported on thresholds 3 and 6, but not 4 and 5. In total 11 additional thresholds were imputed using the SI and MIDC methods. Figure 5.5 shows a shift in the summary sensitivity and specificity (in ROC space) when using the MIDC method compared to the original NI approach, for thresholds 4, 5 and 6 where imputation was possible. For example, for threshold 5, the summary sensitivity was 0.762, 0.707 and 0.685 for the NI, SI and MIDC methods, respectively, with associated standard errors of 0.133, 0.292

and 0.282. After imputation there is generally a shift in summary estimates, downward and to the left in ROC space, similar to what was seen in the PCR example. This indicates a decrease in sensitivity (downward shift) and an increase in specificity (shift left). Applying the SI and MIDC methods therefore reveals that the original conclusions from the NI method for thresholds 4 to 6 are not robust, with summary sensitivity and specificity estimates differing from those estimated when ignoring missing data.

It is of particular interest that in this example the standard errors increase in comparison to the NI approach, where imputation is possible. This is likely due to the large between-study heterogeneity which is better estimated with the increased data available after imputation (see Table 5.8). For example take threshold 5 again, for the NI method only 1 study reported sensitivity for threshold 5, meaning heterogeneity could not be estimated and only this studies estimates could be used for the summary sensitivity. Whereas for the SI and MIDC methods at threshold 5 the estimated heterogeneity was 0.69 and 0.66 respectively, which represents very large between-study variability in comparison to the summary estimate of sensitivity itself.

Thus, as seen in the PCR example, the use of the SI or MIDC methods made a substantial difference to the summary results for the Apgar score at some thresholds, suggesting that the methods are important to use in practice. It was also observed again that there was little difference between the SI and MIDC methods, in terms of summary estimates, standard errors and between-study heterogeneity estimates. This example also illustrates the potential benefit of the imputation methods in improving the estimation of between-study heterogeneity.

Table 5.8 - Summary results for the Apgar example, for summary sensitivity and specificity using NI, SI and MIDC methods.

	Without imputed data (NI)				With single imputed data (SI)				With multiple imputed data (MIDC)			
Threshold	No. studies	Summary Sensitivity	St.Err (logit scale)	Tau	No. studies	Summary Sensitivity	St.Err (logit scale)	Tau	No. studies	Summary Sensitivity	St.Err (logit scale)	Tau
0	1	0.061	0.237		1	0.061	0.237		1	0.061	0.237	
1	1	0.360	0.118		1	0.360	0.118		1	0.360	0.118	
2	1	0.537	0.114		1	0.537	0.114		1	0.537	0.114	
3	9	0.535	0.213	0.60	9	0.535	0.213	0.60	9	0.535	0.213	0.60
4	2	0.691	0.122	0.00	7	0.645	0.297	0.72	7	0.646	0.198	0.70
5	1	0.762	0.133		6	0.707	0.292	0.69	6	0.685	0.282	0.66
6	6	0.819	0.291	0.67	7	0.796	0.284	0.71	7	0.788	0.316	0.80
7	2	0.807	0.292	0.34	2	0.807	0.292	0.34	2	0.807	0.292	0.34
8	1	0.932	0.226		1	0.932	0.226		1	0.932	0.226	
9	1	0.987	0.503		1	0.987	0.503		1	0.987	0.503	
Threshold	No. studies	Summary Specificity	Standard error (logit scale)	Tau	No. studies	Summary Specificity	Standard error (logit scale)	Tau	No. studies	Summary Specificity	Standard error (logit scale)	Tau
0	1	0.998	0.500		1	0.998	0.500		1	0.998	0.500	
1	1	0.961	0.112		1	0.961	0.112		1	0.961	0.112	
2	1	0.926	0.083		1	0.926	0.083		1	0.926	0.083	
3	9	0.879	0.256	0.76	9	0.879	0.256	0.76	9	0.879	0.256	0.76
4	2	0.774	0.382	0.53	7	0.852	0.284	0.74	7	0.879	0.181	0.71
5	1	0.792	0.054		6	0.817	0.272	0.66	6	0.806	0.270	0.65
6	6	0.710	0.287	0.70	7	0.724	0.254	0.67	7	0.722	0.261	0.69
7	2	0.664	0.222	0.31	2	0.664	0.222	0.31	2	0.664	0.222	0.31
8	1	0.353	0.046		1	0.353	0.046		1	0.353	0.046	
9	1	0.097	0.074		1	0.097	0.074		1	0.097	0.074	

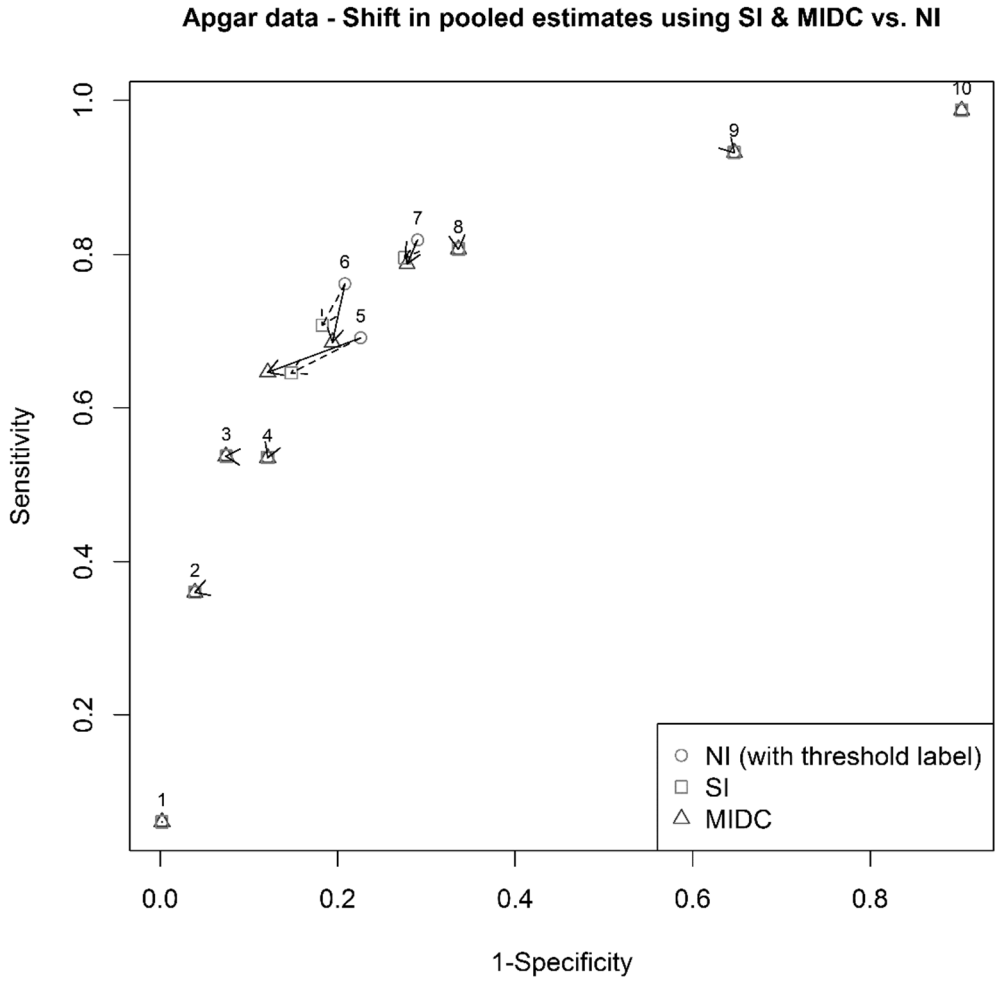


Figure 5.5 - Summary estimates of sensitivity and specificity in ROC space for all methods, for the Apgar example. NB: Arrows represent change from NI summary estimates

5.5 Discussion

To inform clinical decision making, meta-analysis results on the accuracy of a test should be accompanied with clear guidance on the threshold to use for defining test positive and test negative patients. However, often test accuracy meta-analyses suffer from having small and/or discrepant numbers of studies for each threshold of interest. In this chapter a new method was proposed to deal with this problem, based on multiple imputation from the set of all available discrete combinations of missing two by two tables. The results of the applied

PCR example showed a shift in the summary sensitivity and specificity estimates using both imputation methods, indicating that ignoring the missing threshold information (NI method) may have led to biased summary results. The example also highlighted a large improvement in precision, with standard errors substantially smaller for both imputation methods due to the addition of 54 threshold results for meta-analysis.

Theoretically, the MIDC method was developed to improve on the current SI method by allowing for uncertainty associated with the imputed threshold results, and also uncertainty associated with the distance between the missing threshold and the nearest known threshold results. However, in the two examples, there was very little difference between the SI and MIDC results. The MIDC method is more complex as it imputes missing threshold results multiple times and combines these results to give one summary performance at each threshold; this inevitably increases computation time, but only slightly (in the applied example the SI and MIDC methods took 9.55 and 30.3 seconds to run, with MIDC performing five imputations). While this work focuses on the application to tests reporting continuous results future research could extend these methods for use with tests reporting ordinal results. The Software to implement the MIDC procedure in Stata is presented in the appendix.

There has been debate concerning the order in which meta-analysis and Rubin's rules should be performed. The MIDC approach performs meta-analysis on the imputed dataset and then applies Rubin's rules after meta-analysis, as Rubin's theory and recent evidence suggests (56, 130).

One limitation of the imputation methods is that neither the MIDC or SI approaches constrain the ordering of threshold results, meaning that in real world examples the imputed results may lead to imperfect SROC curves (as seen in the PCR example Figure 5.3) (234). The methods of Hamza et al. and Riley et al. constrain the threshold results to be ordered, however these methods are limited by requiring either complete data or a normality assumption on the logit estimates within-studies (231, 234). Further work may look to address this problem.

The multiple imputation approach for the MIDC method used discrete combinations, but an alternative approach may be to assume some distribution for the true underlying diagnostic test distribution, which would allow potential imputations to be drawn from some posterior distribution as in standard multiple imputation approaches; however this would require the distribution of the test be known, and this will often be unlikely and impractical for those researchers (often non-statisticians) who are undertaking such meta-analyses. Steinhauser et al. propose a method based on distributional assumptions for the underlying test distribution (235). In their examples either a normal or logistic distribution is assumed for the diagnostic test, though in practice more complex distributions may better fit particular examples. Indeed the distributions of diseased and non-diseased patients within individual studies may differ, meaning multiple distributional assumptions are needed. In practice the true underlying test distribution is difficult to ascertain without IPD, though in some cases such distributions may be available in published literature.

Finally, there are now numerous published methods to address the synthesis of multiple and missing thresholds in test accuracy meta-analysis (230-236, 238, 239), and so it is pertinent

that future work should aim to investigate a head-to-head comparison of the available methods through simulation (245).

5.5.1 Motivation for subsequent chapter

Given that the MIDC and SI approaches appeared to give similar results in terms of both summary estimates of sensitivity and specificity, as well as their associated standard errors, the next chapter will examine whether these findings generalise using simulations. A simulation study enables the performance of the new MIDC method to be compared against the SI and NI methods in relation to a known truth. This is not possible in a single example (such as the PCR example above), as the true summary estimates are unknown and so no assessment of the methods statistical properties is possible. In contrast, the simulation study in the next chapters allows the statistical properties of the methods to be compared, relative to known true values of sensitivity and specificity, to identify situations in which the imputation methods perform better or worse than the NI approach. In particular, methods are compared in terms of properties including; the bias and precision of summary estimates, the coverage of estimated 95% confidence intervals, and the bias and precision of between-study heterogeneity estimates.

CHAPTER 6: A SIMULATION STUDY TO EVALUATE THE PERFORMANCE OF IMPUTATION METHODS FOR MISSING THRESHOLD RESULTS IN TEST ACCURACY META-ANALYSIS

6.1 Introduction

Where multiple primary studies report on the performance of a continuous test, it is important to synthesise all the evidence at multiple thresholds to inform how best to use the test in clinical practice. In the previous chapter, it was highlighted that a common problem in a test accuracy meta-analysis is missing threshold data, due to reporting of different thresholds across the primary studies. To address this, a new Multiple Imputation by Discrete Combinations (MIDC) method was proposed. The MIDC method imputes a two by two table for any missing threshold in a study that is bounded between two other available thresholds, and repeats this multiple times to account for uncertainty. The MIDC method theoretically improves on the previously proposed single imputation (SI) method of Riley et al. in a number of ways (see Chapter 5) (236). In particular, it accounts for the uncertainty of imputations, and always imputes whole numbers.

In chapter 5 the MIDC and SI methods were applied to two real examples, which illustrated how they could be used in practice to investigate the robustness of results from a standard meta-analysis approach of only analysing observed data with no imputation (NI) (228). In the PCR example there was evidence of a large improvement in the precision of the summary estimates of sensitivity and specificity when using either the SI or MIDC imputation methods,

as expected given the increase in available data after imputation. In terms of summary estimates there was a shift in their values compared to the NI method, potentially indicating that there was bias in the original results using the NI method. The second example additionally indicated that the imputation methods could be beneficial in improving estimation of between-study heterogeneity. Thus there were promising signs of benefit for both imputation methods over the NI method. However, the applied examples also indicated that the SI and MIDC methods may be similar, as they produced very similar summary sensitivity and specificity results and associated standard errors.

These are only two possible examples, and so it is now important to investigate whether the observed performance of the methods generalises to other situations. The methods may perform better or worse under different scenarios, and so a simulation study is warranted to evaluate the MIDC, SI and NI approaches across a range of settings and under different data generating mechanisms (245). An empirical evaluation of the SI method has been published elsewhere, but no simulations comparing to a known truth were conducted (236).

In this chapter a simulation study is therefore designed and implemented to evaluate the performance of the MIDC and SI approaches, in comparison with each other and the standard NI approach. Several scenarios are considered including varying the proportion of missing information, the missingness mechanism, the relationship between threshold value and test accuracy, and levels of heterogeneity between studies. The work has been submitted to the journal *Research Synthesis Methods*, and is currently under review.

The outline of the chapter is as follows. Section 6.2 describes the simulation study design and methods in detail. Section 6.3 presents the results for all the simulation study scenarios, and Section 6.4 concludes with some discussion and recommendations.

6.2 Simulation study methods

A simulation study is now described to compare performance of the MIDC and SI methods with each other, and also with the standard approach of ignoring missing thresholds which is referred to throughout as 'NI' (no imputation). The simulation procedure was undertaken for each of a range of different scenarios. This followed a step-by-step process, as now explained in detail.

As described in chapter 5, the Stata code to perform the SI and MIDC methods as well as the simulation study, was developed by the PhD candidate (Joie Ensor) and Stata 14 code is provided in appendix for the MIDC method (see APPENDIX D: Chapter 5 Appendices) (225).

6.2.1 Step 1: Define the scenario

Table 6.1 shows the 15 different scenarios considered, covering different values for the amount of heterogeneity, the amount of missing data, the missingness mechanism, and the assumed threshold spacing. A simulation was carried out for each of the 15 scenarios. All simulations assumed there were 10 studies available for the meta-analysis, which is typical of the number available in practice.

Table 6.1 - Simulation scenarios including base case and sensitivity scenarios

Scenarios	Studies	Prevalence	Tau (τ)	Missing %	Missing mechanism^	Threshold spacing*
Base case						
1	10	10%	0	50	MCAR	Equal
2	10	10%	0.25	50	MCAR	Equal
3	10	10%	0.5	50	MCAR	Equal
Greater chance of missingness						
4	10	10%	0	70	MCAR	Equal
5	10	10%	0.25	70	MCAR	Equal
6	10	10%	0.5	70	MCAR	Equal
Missing not at random						
7	10	10%	0	50	MNAR	Equal
8	10	10%	0.25	50	MNAR	Equal
9	10	10%	0.5	50	MNAR	Equal
Unequal threshold spacing						
10	10	10%	0	50	MCAR	Unequal
11	10	10%	0.25	50	MCAR	Unequal
12	10	10%	0.5	50	MCAR	Unequal
Extreme unequal threshold spacing						
13	10	10%	0	50	MCAR	Extreme Unequal
14	10	10%	0.25	50	MCAR	Extreme Unequal
15	10	10%	0.5	50	MCAR	Extreme Unequal

*** Assumed threshold spacing; ^MCAR = Missing Completely At Random, MNAR = Missing Not At Random**

6.2.2 Step 2: Generate the number of participants per study

For each meta-analysis dataset of 10 studies, the number of patients per study were randomly selected to be between 30 and 200 using a uniform(30,200) distribution; so as to replicate a small to moderate sample size within studies.

6.2.3 Step 3: Generate the true disease status for each patient in each study

Prevalence of disease for all scenarios was set to 10% (see Table 6.1), with disease status in each meta-analytic dataset being sampled from a Bernoulli(0.1) distribution, to reflect a typical disease prevalence of 10%. Later, as an extension to the work, a prevalence of 50% was

also investigated and this is briefly discussed in section 6.3.6 (with results shown in APPENDIX E5: Extensions).

6.2.4 Step 4: Generate the true sensitivity and specificity values for each threshold in each study

The true sensitivity and specificity results at each threshold for each study were calculated using two linear models with either logit-sensitivity or logit-specificity as responses, and threshold value as an independent predictor (see Equation 6.1). This approach naturally induces a linear relationship between threshold value and logit sensitivity/specificity. For the assumed linear relationship, the coefficients for the constant and threshold predictor were calculated based on those from a previous test accuracy meta-analysis (243), as follows,

$$\begin{aligned} &True\ logit\text{-}sensitivity\ in\ study\ i\ at\ threshold(t) \\ &\quad = \alpha_{1i} + (-0.2719091 \times threshold(t)) \\ &True\ logit\text{-}specificity\ in\ study\ i\ at\ threshold(t) \\ &\quad = \alpha_{2i} + (0.2851818 \times threshold(t)) \\ &\alpha_{1i} = N(3.304182, \tau^2) \\ &\alpha_{2i} = N(-0.0129091, \tau^2) \end{aligned}$$

Equation 6.1

Where τ^2 defines the amount of between-study heterogeneity in the true logit-sensitivity and logit-specificity, which was set to either be zero, 0.25 (moderate) or 0.5 (high) depending on the scenarios chosen (see Table 6.1). The mean summary ROC curve that these equations (Equation 6.1) represent is shown in Figure 6.1 and the assumed linear relationship is illustrated in Figure 6.2 for logit-sensitivity over a range of threshold values.

For each study, the true intercepts (α_{1i} and α_{2i}) were drawn randomly from the two normal distributions in Equation 6.1. This then produced two final equations in each study, which could be used to derive the true logit values at each threshold ($t = 1\ to\ T$), from which the true sensitivity ($sens_{ti}$) and true specificity ($spec_{ti}$) could be obtained by back-transforming from the logit scale.

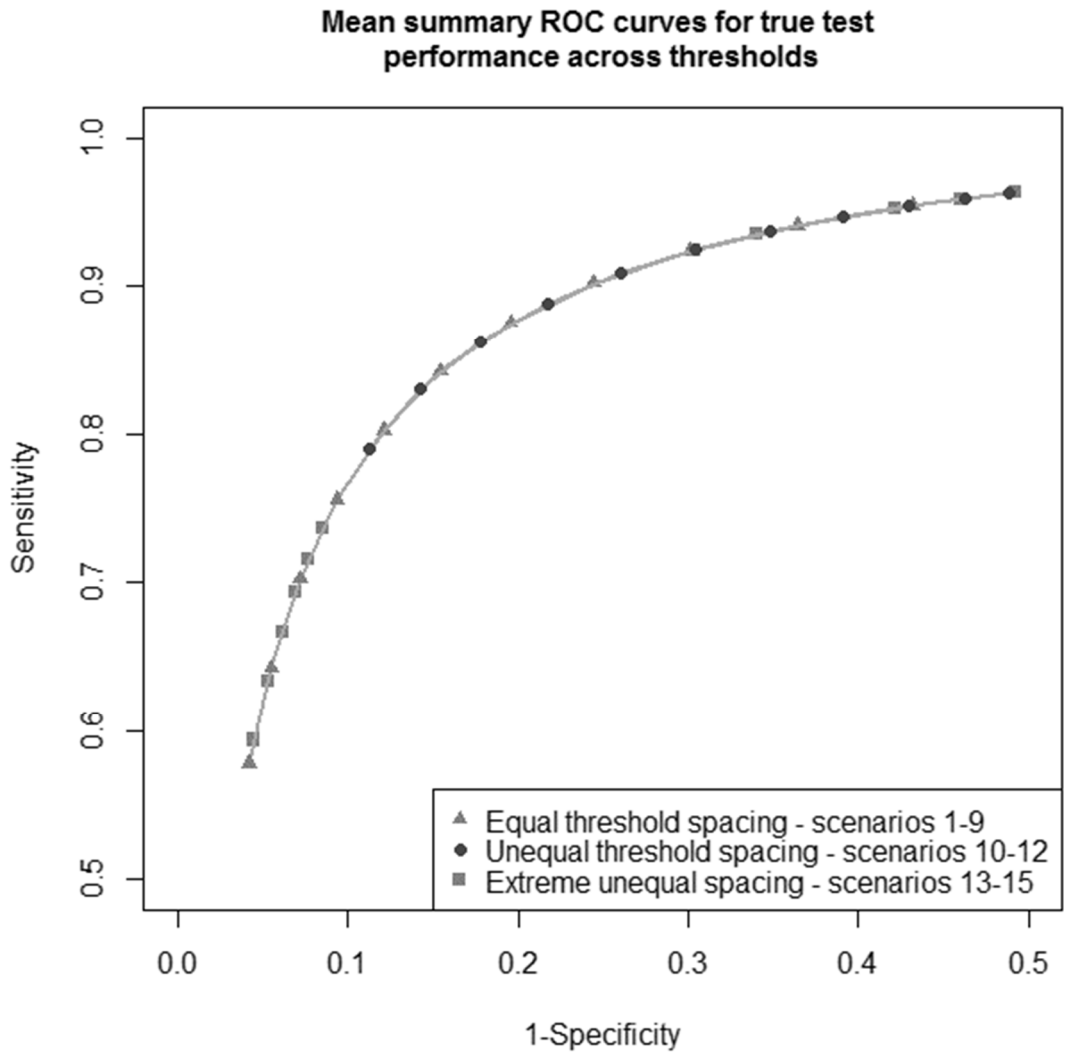


Figure 6.1 - Mean summary ROC curve used for the simulations based on Equation 6.1, illustrating the different threshold spacing as defined by the scenarios in Table 6.1

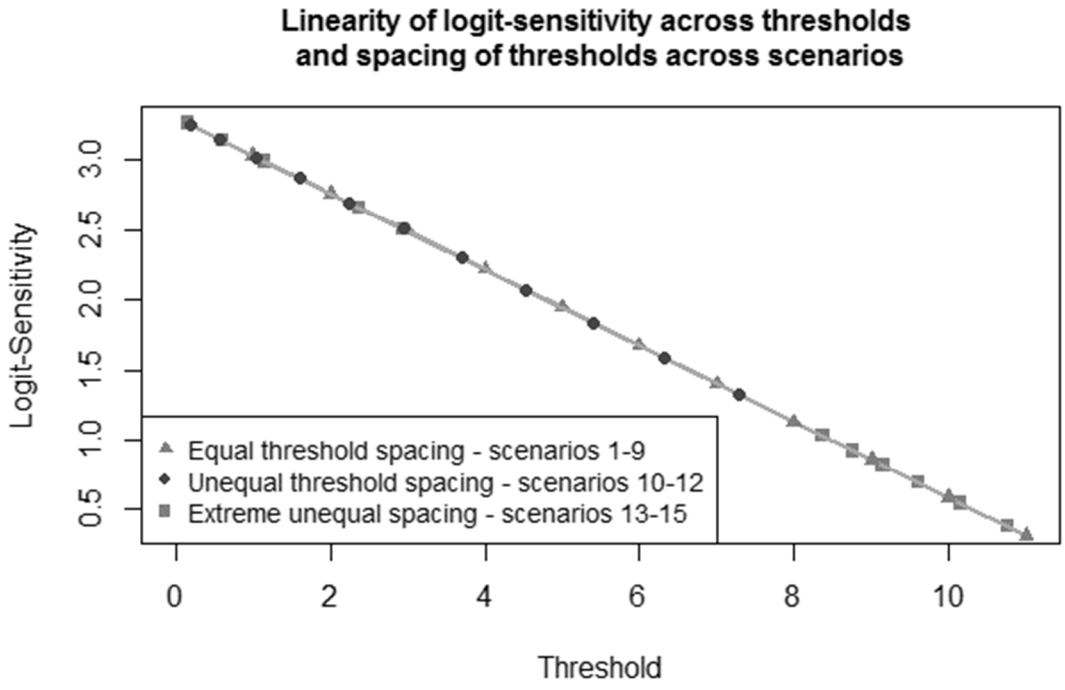


Figure 6.2 - Illustration of the linearity assumption between logit-sensitivity and threshold as defined by Equation 6.1, with threshold spacing defined by the scenarios in Table 6.1

Scenarios 1 to 9 assumed equal threshold spacing, but in scenarios 10 to 15 unequal threshold spacing was assumed (see Figure 6.1 and Figure 6.2), with scenarios 13 to 15 investigating a more extreme unequal threshold spacing.

The eleven thresholds were fixed for all scenarios to be consistent, and this was based on the real example used to inform the parameters in Equation 6.1 (243). The numerical values of the continuous test corresponding to the 11 thresholds were held constant across all studies within each scenario. The chosen threshold values were integer values from 1 to 11 for simulation scenarios 1-9; for scenarios 10-12 a transformation of these threshold values was used to induce unequal spacing, using Equation 6.2 as follows:

$$Threshold\ value\ for\ scenarios\ 10\text{-}12 = \frac{(Threshold\ value\ for\ scenarios\ 1\text{-}9)^{1.5}}{5}$$

Equation 6.2

Scenarios 13-15 were introduced as an extreme scenario to examine the effect of missing thresholds in the centre of the ROC curve. For these scenarios extremely unequal spacing of thresholds was considered as a combination of "bunching" of similar thresholds (at either end of the continuous measurement scale) and of "large gaps" between thresholds in the centre of the continuous scale (see distribution of squares in Figure 6.1 and Figure 6.2). In real applications, bunching of thresholds may be observed, for example, in meta-analyses when a set of studies have used a very similar threshold, but with differences in the last digit or decimal place. Large gaps may occur, for example, between thresholds chosen to optimise a test for rule out purposes (where a low threshold would be evaluated) or rule in purposes (where a high threshold would be evaluated). Threshold values were hand selected to create this effect for scenarios 13-15, with thresholds at 0.15, 0.6, 1.15, 2.37, 2.95, 8.37, 8.76, 9.15, 9.6, 10.15 and 10.76 (see Figure 6.2). It should be noted that scenarios 13-15 examine the effect of missing thresholds in the centre of the ROC curve, though a similar issue could occur elsewhere on the ROC curve.

6.2.5 Step 5: Generate the observed number of TP, TN, FP and FN at each threshold

For each study separately, an observed two by two table was then generated for each threshold. To create this the multinomial distribution was expressed as a series of conditional binomial distributions (231). Firstly, to calculate the number of TPs above threshold 1 (the lowest threshold, t =1), TP_{1i} was randomly sampled from a binomial distribution with D_i=the total number of diseased in study *'i'*, and $sens_{1i}$=estimated sensitivity for threshold 1 in study *'i'*.

$$TP_{1i} \sim Binomial(D_i, sens_{1i})$$

Equation 6.3

And thus,

$$FN_{1i} = D_i - TP_{1i}$$

Equation 6.4

For subsequent thresholds, the TP_{ti} were derived using a conditional binomial distribution. For example, the number of TPs were generated above threshold 2 out of the subset of patients who were positive above threshold 1, by calculating the observed TP_{1i} minus a random sample from a $TP_{2i} \sim Binomial\left(TP_{1i}, \frac{sens_{2i}}{sens_{1i}}\right)$ distribution. In this way, each successive TP_{ti} accounted for the previous $TP_{(t-1)i}$ value.

TN_{ti} and FP_{ti} for the non-diseased population were generated in a similar manner for each threshold.

6.2.6 Step 6: Create missing results for some thresholds

Step 5 produced complete data (i.e. a two by two table) for each threshold in each study. To create missing data, some of the two by two tables in each study were removed using either a missing completely at random (MCAR) or, for scenarios 7 to 9, a missing not at random (MNAR) mechanism (57, 246). For the missing completely at random scenarios, each two by two table was given a percentage probability of being missing according to the percentage missing data dictated by the scenarios setting (see Table 6.1). For the missing not at random scenarios, two by two tables could only be missing where the observed Youden's index was < 0.7 (where Youden's index = sensitivity + specificity – 1, and therefore provides an overall measure of test performance) (247). This reflects a situation where test performance is worse than expected at this threshold, and so it is vulnerable to publication bias and non-reporting in the study publication (248-250). Such thresholds were given a 50% chance of being missing in scenarios 7 to 9. Other thresholds where Youden's index was at least 0.7 were always assumed to be available.

6.2.7 Step 7: Apply meta-analysis to each simulated dataset using NI, SI or MIDC methods

Steps 1 to 6 were applied to obtain 1000 meta-analysis datasets in each scenario. For each dataset, the NI, SI and MIDC methods were applied separately to produce 1000 meta-analysis results for each type of approach. For the MIDC method five imputation datasets were performed before Rubin's rules was applied (giving one set of meta-analysis results per

dataset) (56); five imputations was chosen to reduce the computation time of the overall simulation study. A bivariate random-effects meta-analysis model (see Equation 5.2) as presented in chapter 5 was used to synthesise the available results for each threshold separately (238, 239), but with the between-study correlation set to zero to avoid the common computation issues associated with this parameter (240, 241). The model was fitted via maximum likelihood estimation using Gauss-Hermite quadrature (251), with quadrature points equal to the number of studies in the meta-analysis. However, where more than five studies remained after missingness was generated, the number of quadrature points was restricted to be no greater than five, as additional points over this limit had little impact on summary estimates but increased computation time. Convergence issues with the bivariate model meant that some simulation runs were discarded; only runs which converged for all three methods were retained, so as to maintain a fair comparison.

The performance of the three methods in each scenario was summarised and compared in terms of the bias in the mean of the summary (logit) sensitivity and (logit) specificity estimates; the mean of the standard errors of the summary results; the mean bias in the estimate of τ, and the percentage coverage of the 95% confidence intervals for the summary sensitivity and specificity. The 95% confidence intervals for the summary sensitivity and specificity at each threshold were derived on the logit scale using the summary logit estimate +/- 1.96 *s.e.(summary logit estimate), and then back-transformed.

6.3 Results

6.3.1 Base case settings (Scenarios 1 to 3)

The three base-case settings of scenarios 1 to 3 each involved a 10% prevalence, 50% of thresholds missing completely at random, and equal threshold spacing, but varied according to the magnitude of between-study heterogeneity (see Table 0.14, Table 0.15, Table 0.16, Table 0.17, Table 0.18 and Table 0.19). For each scenario, mean summary estimates from the 1000 simulations from each of the NI, SI and MIDC approaches were plotted at each threshold in ROC space and compared with the true ROC curve (see Figure 6.3). When the heterogeneity was zero or moderate, there was very little bias at all thresholds, for either the NI method or the imputation methods. However, there was slight bias when the heterogeneity was large for all methods (Scenario 3), with summary sensitivity and specificity underestimated across the thresholds. The bias was slightly worst for the SI method, and the MIDC and NI methods performed similarly (see Figure 6.3).

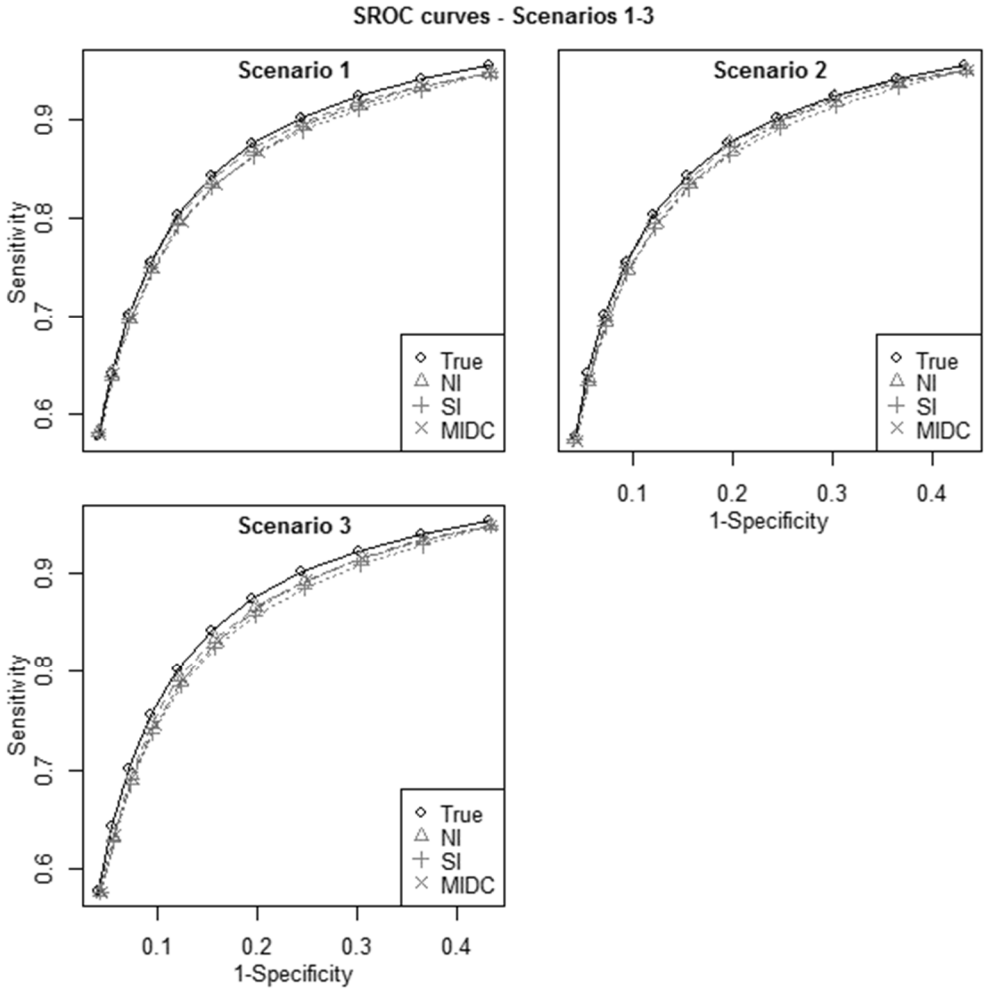


Figure 6.3 – ROC curves compared to true estimates (base case scenarios 1-3)

Coverage of 95% confidence intervals was similar across the methods when there was no between-study heterogeneity in scenario 1 (coverage ranged from 92% to 96%), with the NI method performing at least as well as the SI and MIDC methods (see Figure 6.4). At moderate heterogeneity (scenario 2), the two imputation approaches had improved coverage over the NI approach (i.e. closer to 95%) at most thresholds for specificity, while SI and MIDC performed similarly (see Figure 6.4). For example, at a threshold value of 4 the coverage for specificity was 89.2%, 92.3% and 93.8% for the NI, SI and MIDC approaches, respectively. The

improvement in coverage by using the MIDC or SI approaches rather than NI was even more pronounced at a high level of between study heterogeneity (see Figure 6.4), with MIDC performing best. For example, in scenario 3 with a threshold value of 4 the coverage for specificity was 83.7%, 88.6% and 90.1% for NI, SI and MIDC approaches respectively.

The improvement in coverage by using the imputation methods is most likely due to the estimate of between-study heterogeneity being substantially improved by including more studies at each threshold after imputation (see Figure 6.5). Maximum likelihood estimates of variances are known to be downwardly biased in small samples, and thus increasing the sample size via imputation ensures an improvement. Though downward bias in $\hat{\tau}$ remains for both MIDC and SI, it is far smaller than the bias for NI, and the MIDC method consistently provides the least biased $\hat{\tau}$ values across the thresholds (see Figure 6.5).

Despite the larger estimates of between-study variance ($\hat{\tau}$) after imputation, the incorporation of additional results via imputation substantially improves the precision of summary estimates in the MIDC and SI approaches compared to NI (see Figure 6.6). The gain in precision is largest in the central thresholds, where there is greater opportunity for imputation (i.e. higher chance of a missing threshold falling between two known bounding thresholds); indeed for these thresholds the imputation approaches were usually able to recover data on all ten studies in the meta-analysis (see APPENDIX E1: Base case scenarios). Mean standard errors were almost identical for the SI and MIDC approaches, but with the MIDC method providing slightly inflated standard errors as it accounts for the uncertainty in imputations.

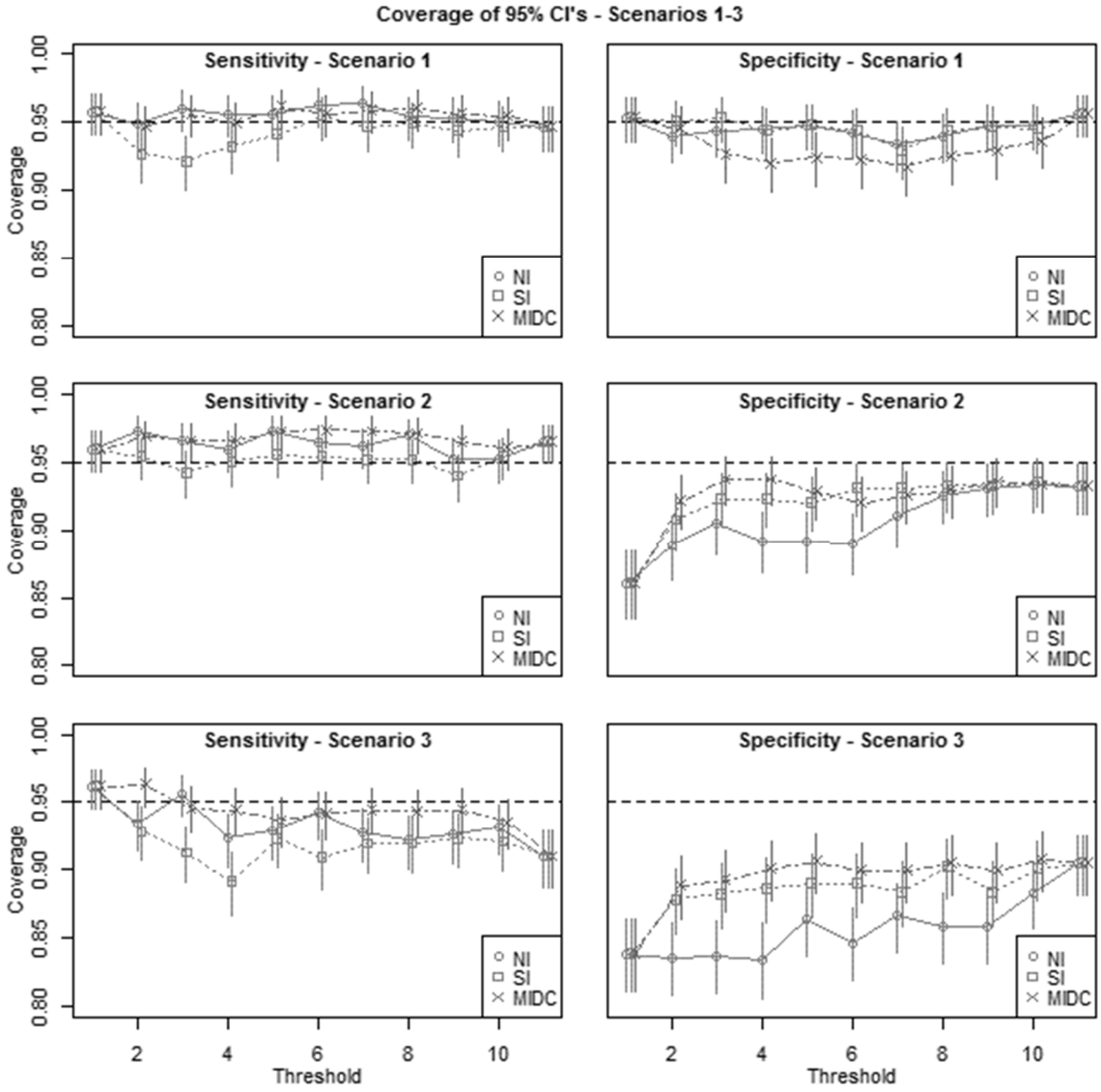


Figure 6.4 – Coverage of 95% confidence intervals (base case scenarios 1-3). Dashed line indicates ideal 95% coverage.

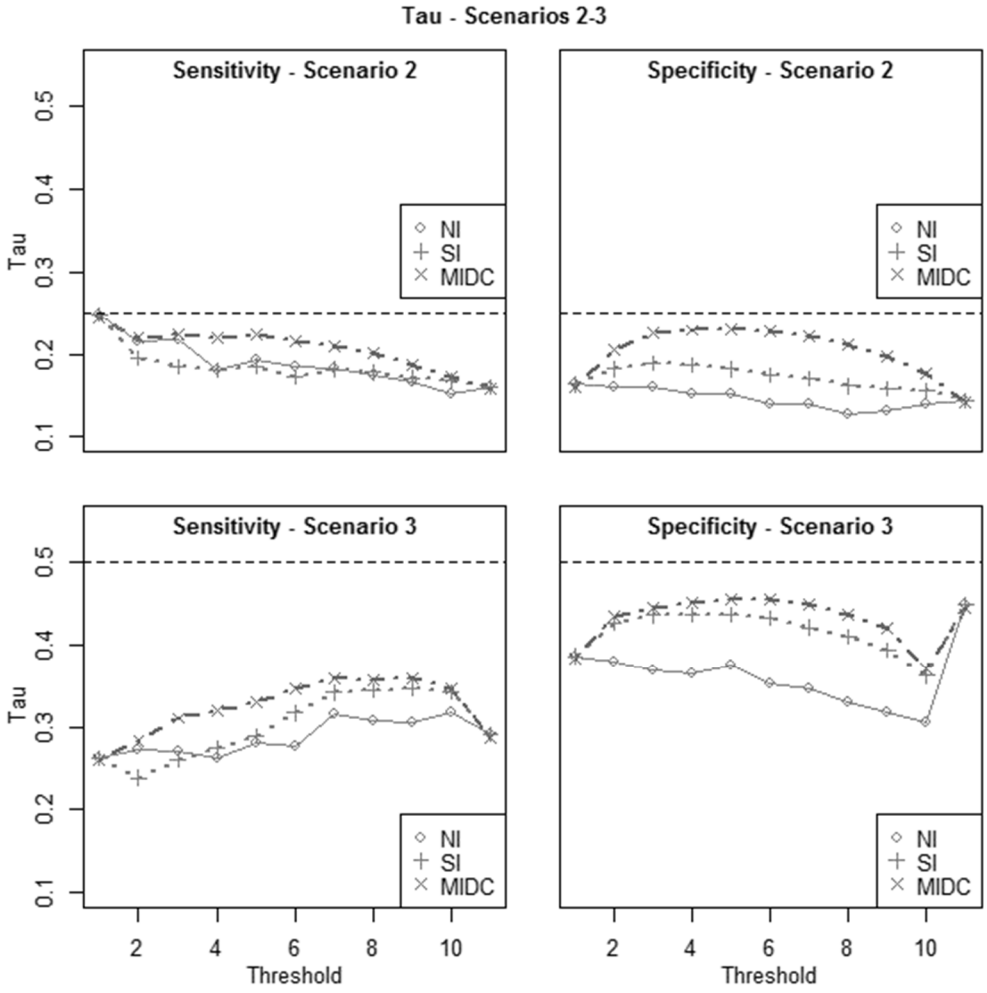


Figure 6.5 – Mean estimate of $\hat{\tau}$ for summary sensitivity for scenario 2 and 3. Dashed line indicates the true simulated $\hat{\tau}$ for scenario 2 ($\hat{\tau}$=0.25) and scenario 3 ($\hat{\tau}$=0.5).

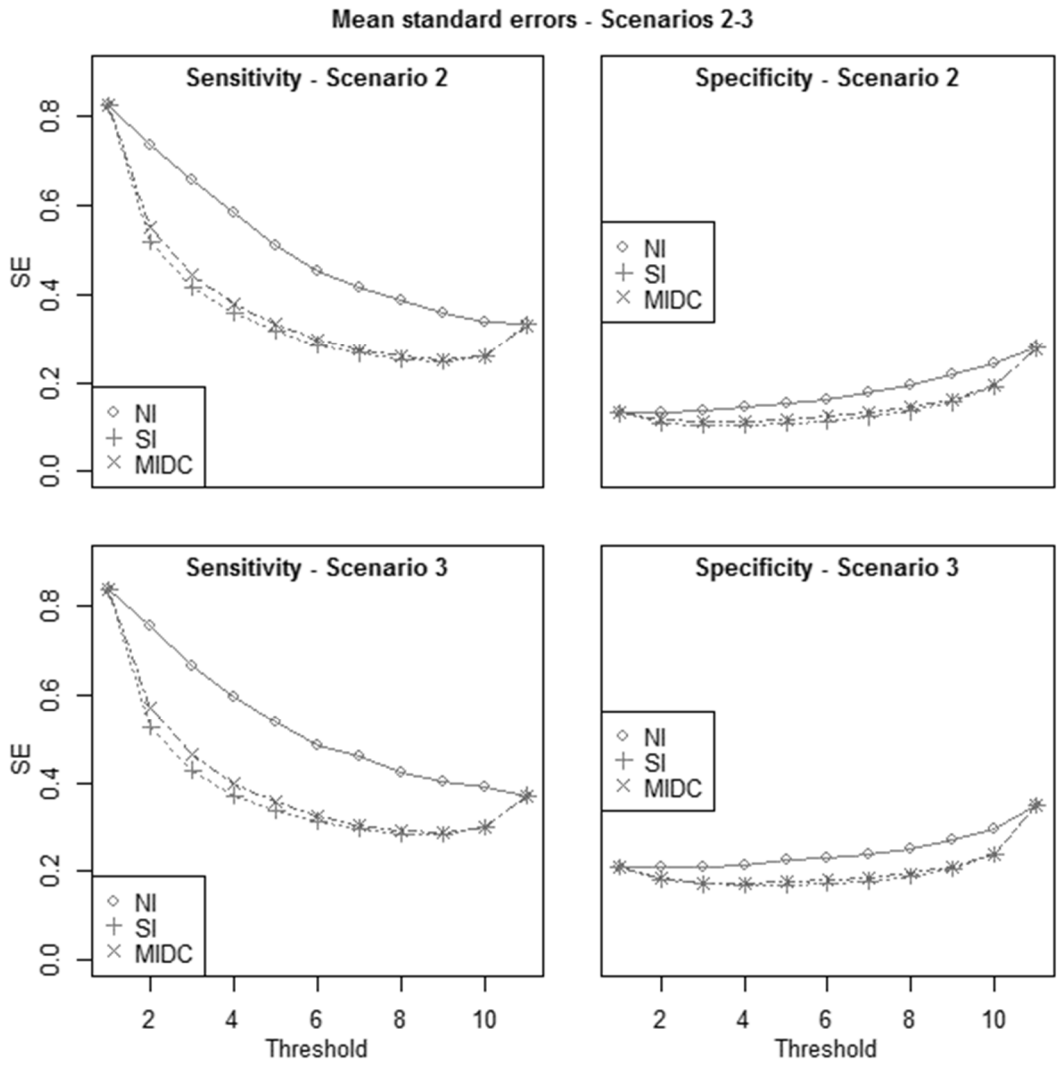


Figure 6.6 – Mean Standard errors (base case scenarios 2-3)

6.3.2 Greater chance of missingness (Scenarios 4 to 6)

In scenarios 4 to 6 the percentage of thresholds missing completely at random was increased to 70%. In terms of bias, the findings were similar to those in scenarios 1 to 3, with the summary ROC curves showing very small bias for all the methods when the between study heterogeneity was large. Coverage and precision was again substantially improved by using the MIDC and SI approaches, even more so than for scenarios 1 to 3 due to the larger

percentage of missing data. For example, at a threshold value of 5 in scenario 6 the coverage was 78.9%, 88.6% and 88.7% for the NI, SI and MIDC methods, respectively. The two imputation methods were similar in terms of mean standard errors, which were reduced by up to 43% compared to the NI method.

6.3.3 Missing not at random (Scenarios 7 to 9)

Under the MNAR assumption in scenarios 7 to 9, a threshold was always present when the observed Youden's index was > 0.7, but otherwise had a 50% chance of being missing akin to selective reporting bias (see Table 0.20, Table 0.21, Table 0.22, Table 0.23, Table 0.24 and Table 0.25). Figure 6.7 shows the summary ROC curves for each method from scenario 9 (high heterogeneity), and for the NI method it reveals an upward bias (overestimation) of summary sensitivity and specificity at each threshold when not accounting for the missing threshold data. There was also a similar upward bias for the NI method in scenarios 7 and 8, where there was lower heterogeneity. In contrast, the SI and MIDC methods reduce this bias through imputation, and produce mean ROC curves that are close to the true summary ROC curve in each of scenarios 7 to 9.

Coverage of 95% confidence intervals was again consistently better when using the imputation approaches than the NI approach. Differences between the SI and MIDC methods were generally small, though slightly better for the MIDC method when heterogeneity was large (see Figure 6.8 for results of scenario 9). Precision was also greatly increased when using either the SI or MIDC methods as previously noted.

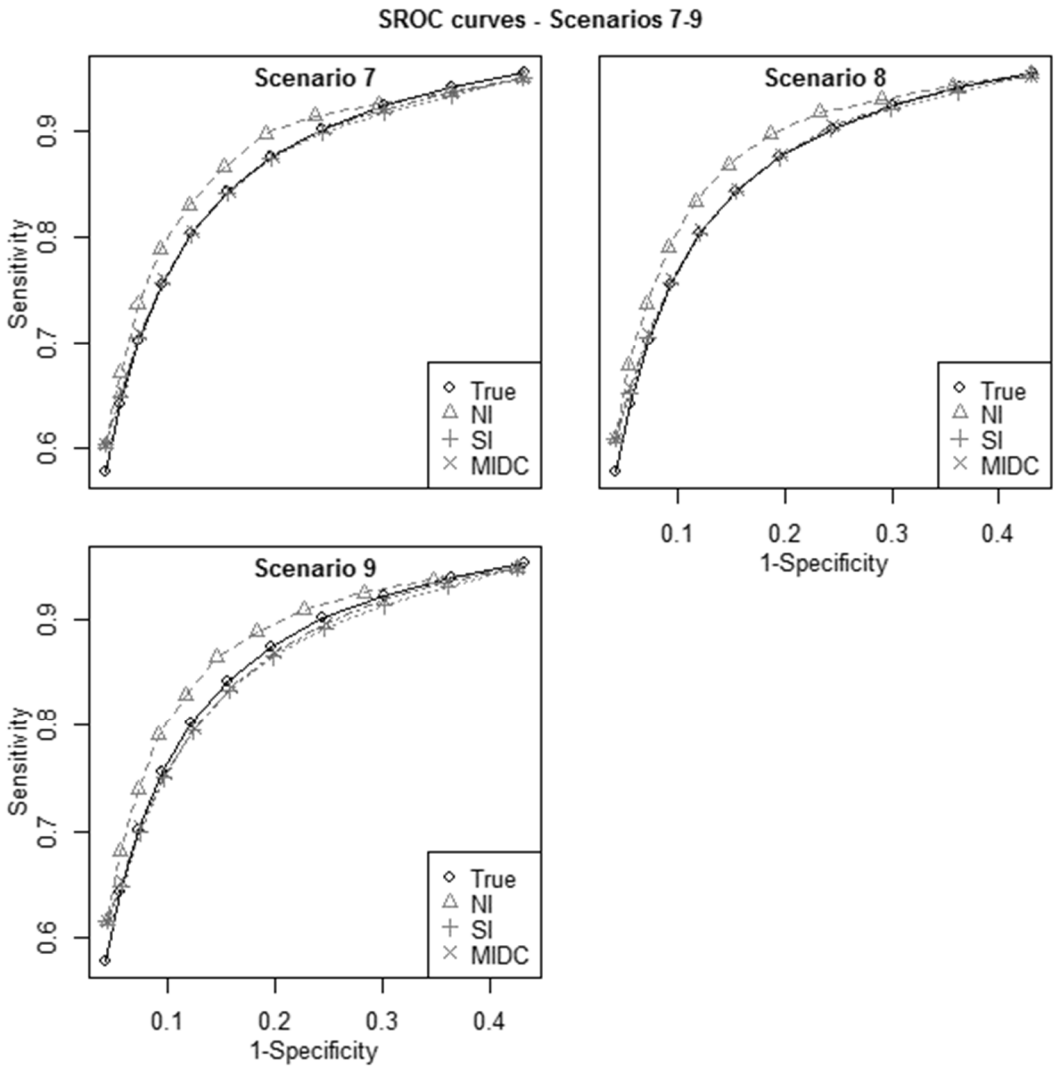


Figure 6.7 - Mean summary ROC curves for all methods. Scenarios 7 to 9.

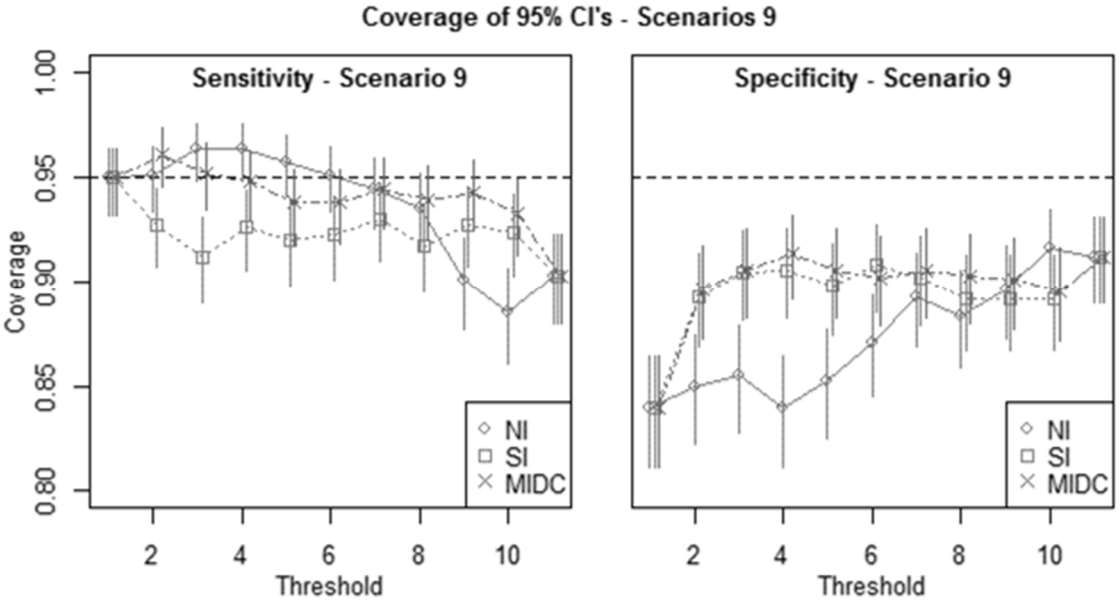


Figure 6.8 - Coverage of 95% confidence intervals for MNAR scenario 9. Dashed line indicates ideal 95% coverage.

6.3.4 Unequal threshold spacing (Scenarios 10 to 12)

All previous scenarios involved equal spacing of logit sensitivity and logit specificity across the range of thresholds. Scenarios 10 to 12 assess how the imputation methods perform given unequal threshold spacing along the ROC curve, as was depicted previously in Figure 6.2. Figure 6.9 presents the mean of the summary ROC curves from the simulation results for scenario 12 (high heterogeneity). All methods showed some downward bias in summary sensitivity and specificity estimates, but the SI method performed worst (see Table 0.26 and Table 0.27). The imputation methods again improved performance over the NI in terms of smaller standard errors and less biased estimates of $\hat{\tau}$, with the MIDC generally performing best in terms of coverage (see Figure 6.10).

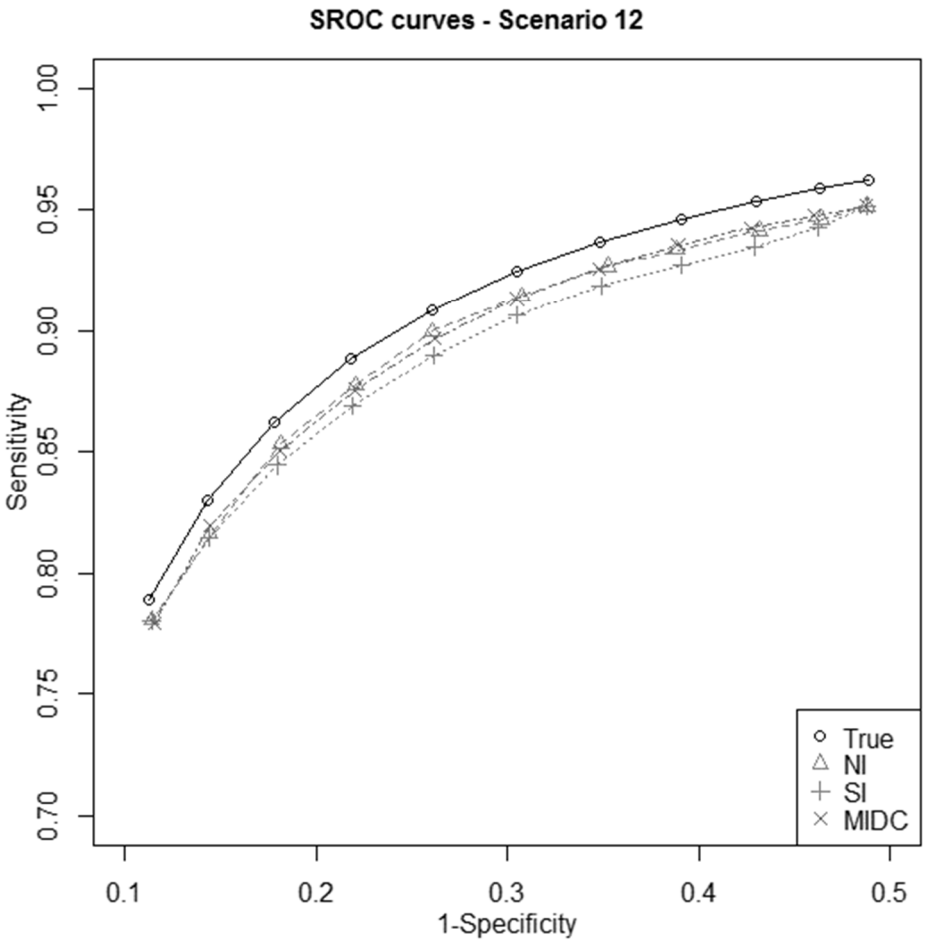


Figure 6.9 - Mean summary ROC curves all methods. Unequal threshold spacing scenario 12

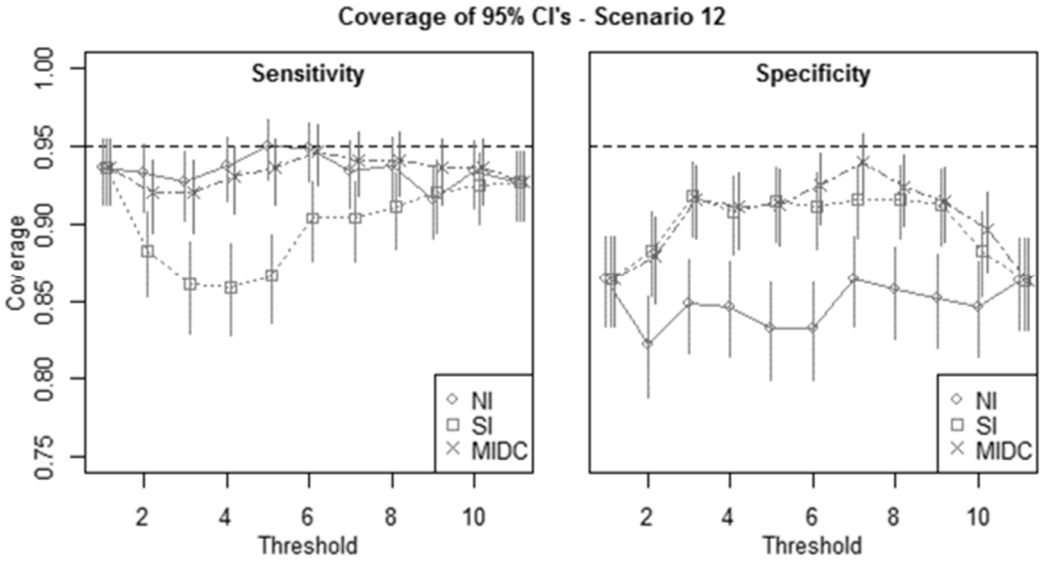


Figure 6.10 - Coverage of 95% confidence intervals for unequal threshold spacing scenario 12. Dashed line indicates ideal 95% coverage.

6.3.5 Extreme unequal threshold spacing (Scenarios 13 to 15)

The impact of unequal threshold spacing was explicitly examined under more extreme, unequal spacing situations (scenarios 13 to 15). These represented scenarios where thresholds were reported at either end of the ROC curve, but not reported in the central part of the ROC. This scenario may occur where studies are considering tests for rule in or rule out purposes as discussed in section 6.2.4. The concept was shown graphically in Figure 6.2.

Figure 6.11 presents the mean of the summary ROC curves for the simulations results of scenario 15 (high heterogeneity). Interestingly, the NI method showed little bias in sensitivity and specificity at the available thresholds; however, the two imputation methods performed poorly in terms of the central thresholds (see Table 0.28 and Table 0.29).

The coverage of 95% confidence intervals appeared better when using either the SI or MIDC imputation approaches in the outer reported thresholds (see Figure 6.12). However, the coverage of the imputation methods dropped substantially in the central thresholds. For example, in scenario 15 at a threshold value of 5 (a central threshold) the coverage for sensitivity was 96.4%, 63% and 73.8% for the NI, SI and MIDC respectively, while at a threshold value of 8 (an outer threshold) the coverage was 90.6%, 92.8% and 93.8% respectively. Precision of the summary results was again improved using SI or MIDC compared to NI, but clearly this is at the expense of poor coverage in the central thresholds in this particular setting.

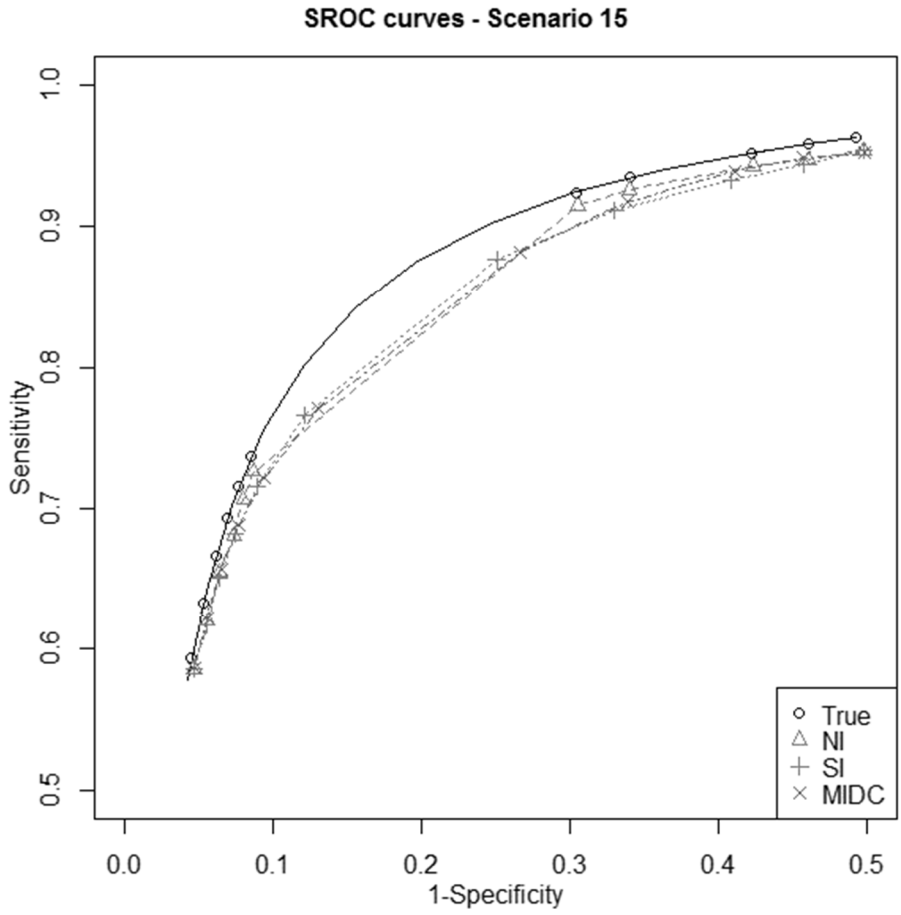


Figure 6.11 - Mean summary ROC curves all methods. Unequal threshold spacing scenario 15

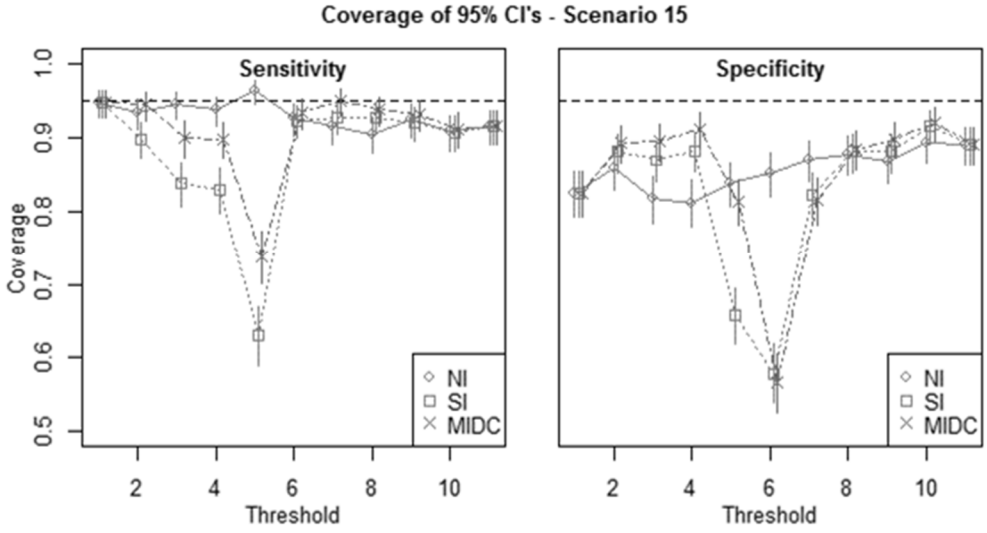


Figure 6.12 - Coverage of 95% confidence intervals for unequal threshold spacing scenario 15. Dashed line indicates ideal 95% coverage.

6.3.6 Extensions

Prevalence of 50%

All simulations were repeated with a prevalence of 50%, rather than 10%. In particular, for scenarios 1 to 9 the NI approach generally performed worst, whilst the MIDC method performed generally best in terms of lower standard errors, coverage closest to 95%, and reduction in bias of $\hat{\tau}$ estimates and, in scenarios 7 to 9, summary estimates. For example, Figure 6.13 shows the coverage results for scenario 9 with a prevalence of 50%, for which the MIDC is generally closest to the 95% level for all thresholds for which there is missing data. Only in extreme scenarios 13 to 15 was the NI method preferable, for example in terms of coverage at central thresholds as seen under the 10% prevalence results.

It was observed that under the assumption of a 50% prevalence of disease, the coverage of 95% confidence intervals for summary sensitivity and specificity were similar (regardless of the method used), unlike at 10% prevalence, where the coverage of 95% confidence intervals for summary sensitivity was higher than those for summary specificity (see Figure 6.13). At 10% prevalence the estimates of specificity will have far greater precision than sensitivity, due to the greater number of non-diseased individuals in the sample.

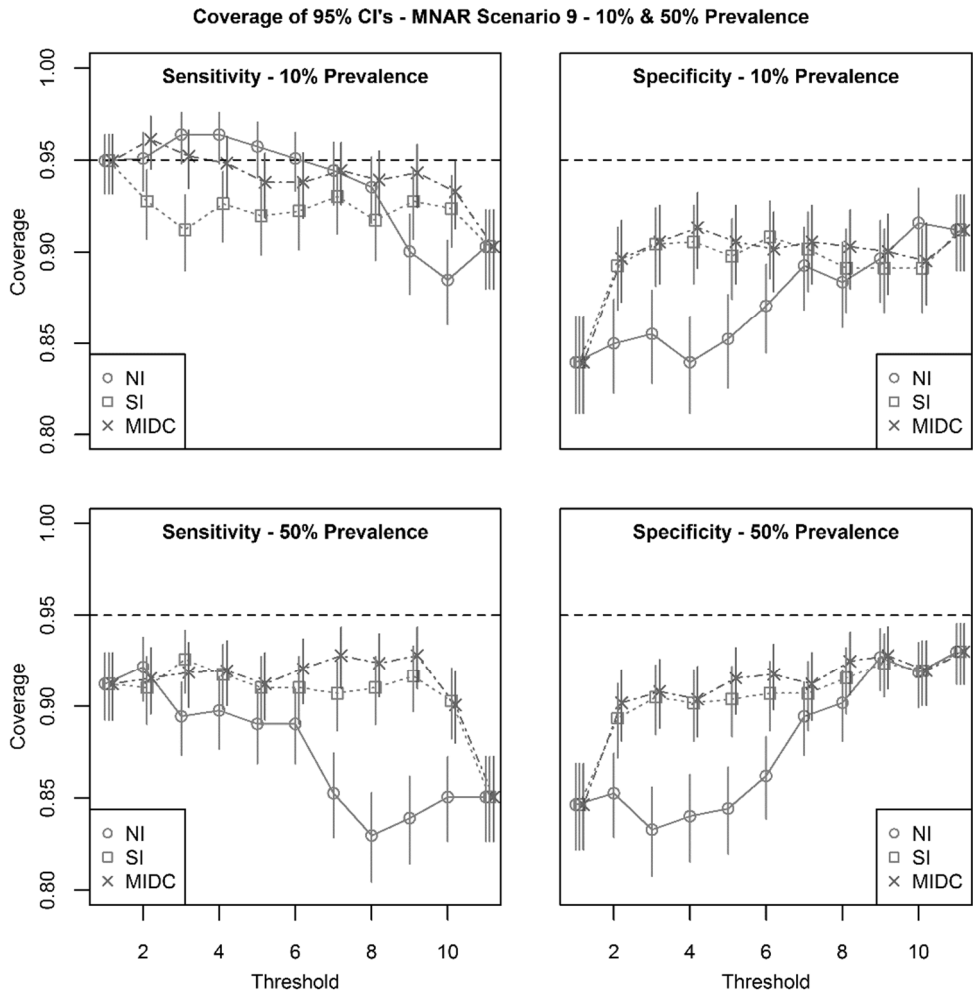


Figure 6.13 - Coverage of 95% confidence intervals for MNAR scenario 9. Comparing simulations results at 10% prevalence (top figures) and 50% prevalence (bottom figures). Dashed line indicates ideal 95% coverage.

Alternative number of studies

Simulations were also conducted with a smaller number of studies; five studies was chosen to represent a smaller meta-analysis dataset as is commonly seen in the literature. Conclusions remained broadly similar, though the methods struggled with convergence issues, likely due to the use of the bivariate random-effects model as discussed earlier. Convergence was good under a no heterogeneity setting (scenario 1), but very poor at moderate and high

heterogeneity leaving only 59/1000 runs for analysis. It is important to note that this is not the number of SI and MI runs which converged, but rather failures to converge for the NI method which meant the corresponding SI and MI results were removed to allow fair comparison. The NI failures are likely due to very small numbers of studies, around 2 studies per meta-analysis as the base case scenarios created 50% missing threshold results; even where the SI and MI could impute the missing thresholds there would be very little data for estimating the model. Results of this extension are presented in APPENDIX E5: Extensions, for the base case scenarios.

Increased number of imputation datasets

Finally, as an extension the number of imputation datasets used in the MIDC method was increased from five imputation datasets to ten to investigate the effect of additional imputations on the summary estimates. Conclusions remained the same with ten imputation datasets, indicating that small numbers of imputations may be sufficient at least in the base case settings. Results of this extension are presented in APPENDIX E5: Extensions, for the base case scenarios.

6.3.7 Summary of simulation findings

In general, across scenarios 1 to 9 where the equal spacing assumption was made, the simulations suggest that both the SI and MIDC methods perform better than the current standard NI method (228), in terms of coverage and precision of summary sensitivity and specificity estimates, either when thresholds are missing completely at random or selectively reported according to Youden's index (247). This held for prevalence's of either 10% or 50%. Improvements are due to the extra information arising from the imputed data, which also

leads to improved estimation of the between-study variances. Further, when there is selective reporting due to Youden's index (247), the findings suggest that the SI or MIDC methods can even reduce bias in the summary ROC curve, as well as improving coverage and precision. There is generally very little difference in the SI and MIDC methods, but the latter was noticeably better in terms of estimating the between-study variances and generally gave better coverage, due to slightly larger standard errors of summary estimates. However, when moderate unequal threshold spacing was assumed (Scenarios 10 to 12) the NI and MIDC methods performed better than the SI in terms of bias, with MIDC giving better estimation of standard errors and between-study variances as before.

A concern, however, for both the imputation methods is that their performance deteriorated under extreme unequal spacing (scenarios 13 to 15), and the NI method performed far better in this situation, for example in terms of coverage and bias.

6.4 Discussion

The results from the simulation study, across a wide-range of scenarios, suggests that the previously proposed SI method and the new MIDC method help regain otherwise lost information, and generally improve performance of meta-analysis results compared to the NI method, unless there is extreme unequal spacing of thresholds. The SI and MIDC methods dramatically increase precision of summary estimates of sensitivity and specificity at each threshold, as more data are added via imputation, which usually leads to a distinct improvement in coverage of 95% confidence intervals compared to the standard NI method. This is especially evident when heterogeneity is large, as the SI and MIDC methods improve

estimates of between-study variance. Further, when thresholds are selectively missing due to a poor Youden's index, the findings suggest that the SI or MIDC methods generally reduce bias in the summary ROC curve compared to the NI method. In the situation of moderate unequal threshold spacing the MIDC and NI methods perform better in terms of bias and coverage; and therefore the threshold spacing assumption is important to the correct performance of the SI method. There is very little difference in most scenarios between the SI and MIDC methods; however, in a few scenarios there was improved estimation of between-study variances with the MIDC method and increased standard errors allowing for uncertainty in the imputation, sometimes leading to slightly better coverage. Therefore, the simulation suggests that generally the MIDC approach should be preferred to the SI method. Results were found to be consistent at lower (10%) and higher (50%) levels of prevalence.

Limitations of the simulation study

Model convergence is known to be an issue with the bivariate model used in the simulation study and there were some problems with convergence observed. However this was not substantial enough to effect the conclusions of the main simulation study. The largest number of simulations failing to converge occurred in the high heterogeneity scenario with 70% missing threshold data, where difficulties with convergence may be expected due to sparse data. Even in this scenario the number of failed convergences only reached around 500, leaving 500 simulations for analysis.

Another limitation of this simulation study was the focus on using 10 studies in most simulations. This was chosen firstly because of the increased computation time required to conduct the simulations for larger numbers of studies, and because it is typical of many meta-

analyses in the test accuracy field. However, the extension to investigate a smaller meta-analysis datasets of five studies showed similar findings, but suffered with convergence issues due to the small number of studies further reduced through missingness. Due to the large computation time of the simulations, only five MIDC datasets were used throughout the simulation study. As with common multiple imputation methods, the number of MIDC datasets (or imputations) can be increased to reduce uncertainty further. However, in reality sensitivity analyses could be conducted in individual cases to decide on an appropriate number of MIDC datasets to reduce uncertainty as required. One possible recommendation could be to ensure *M* is set to be 100*p, where p is the proportion of missing studies for the threshold with the most missing data that can be imputed, to be consistent with guidance elsewhere (57, 58). An extension of the simulation study investigated the use of ten imputation datasets for the base case scenarios, and results showed no difference indicating that a small number of imputations may be sufficient when using the MIDC method.

Limitations of the MIDC method and further work

The simulation study identified some limitations of the imputation methods. In situations of extreme unequal spacing, the MIDC and SI methods perform less well, and so judgement is needed as to whether this is likely to be the case in particular examples. Therefore a key issue is how we detect extreme unequal spacing, and how we draw conclusions from the imputation methods in these cases; further research is needed to investigate these questions. One cause of this problem (illustrated in the extreme unequal spacing scenarios) is “bunching”, where we have a set of thresholds across studies of which some are very close together (for example say that we had thresholds of a biomarker at 10, 20, 20.1, 20.15, 20.2, 20.25, 20.3, 30 40, 50).

Where bunching occurs, two possible solutions could be i) to group thresholds that are very close together, or ii) to select a subset of thresholds for analysis, so that the thresholds are roughly equally spaced. The effects of different subsets of thresholds could then be investigated in sensitivity analyses.

Also, the imputation methods utilise a linearity assumption in the change in logit-sensitivity and logit-specificity as the threshold value increases by one-unit. Steinhauser et al. assume a distributional shape for both the diseased and non-diseased population, but assessing the true distribution of the test would require IPD across the range of the test, which will often not be reported (235). Further work may aim to assess how suitable the linearity assumption is, and to what extent deviation from this distribution matters.

As discussed in chapter 5, there is now a clear need for a large simulation study to compare head-to-head the available methods for handling multiple and partially reported thresholds in diagnostic and prognostic test accuracy meta-analyses (230-232, 234-236). Such a study could help to identify which methods perform well under certain situations and guide use of the methods in practice to improve the reporting test accuracy meta-analyses.

Conclusion

In conclusion, the simulation study extends the work of the previous chapter and supports a recommendation for the use of the MIDC method in practice when threshold spacing is not extreme. If the spacing of results across threshold values is not clear, then the MIDC method could still be used as a sensitivity analysis. For example, researchers who retain the NI approach as their primary analysis method could subsequently use the MIDC method to

investigate the potential impact of missing threshold information on the summary sensitivity and specificity of the test, assuming the underlying spacing is not extreme. This could be particularly important in flagging that summary test accuracy may be weaker than originally thought from the NI results, perhaps due to publication bias related issues. Similarly, if applicable, other methods for dealing with multiple thresholds might be considered (230-232, 234). Indeed, further research is urgently needed to compare the MIDC approach with all the other competing suggestions for dealing with missing thresholds in test accuracy meta-analysis.

CHAPTER 7: DISCUSSION

7.1 Overview of the thesis

Prognosis research forms a crucial part of medical research, informing clinical judgement through prediction of risk of future health outcomes given some baseline condition (3). The work in this thesis has focussed on prognosis research to identify and evaluate single prognostic factors and, in particular, multivariable prediction models to predict individuals' risk of outcome given specific patient characteristics (9). Prediction model research includes several important steps which are needed to ensure a model is useful in practice, including development, validation and impact assessment (1, 3, 8-10, 26-28). Evidence showing poor quality in the reporting and methodology used in the prediction model literature has led to various studies calling for improvements and providing guidelines for future research (33, 68, 77). The increasing availability of large datasets including clustering of patients, in studies, centres, regions or countries, provides a great opportunity for the future of prediction model research (252). By utilising multiple studies and the clustered nature of large datasets the work showed how more robust inferences can be made for prognostic factors and prediction models; in particular, models can be developed, validated and updated in various settings and patient populations immediately (21), whilst investigating heterogeneity between populations of patients and the effect of such heterogeneity on the model and its performance (19, 21, 24, 71).

The overall aims for the thesis were the application and development of novel statistical methods for prognosis research, using evidence from multiple studies. Chapters 2 to 4

focussed on prediction models, and the reporting, development, validation and updating of such models, particularly where data is available from multiple studies. Chapters 5 and 6 focussed on issues with the synthesis of prognostic (or diagnostic) predictors reported at different thresholds in primary studies, and developed a novel statistical method to handle such issues. The following provides a short summary of the chapters of this thesis.

7.1.1 Summary of the chapters

The thesis contained a mix of clinical application and methodological development. A systematic review was undertaken in chapter 2, of studies developing or validating prediction models for individual recurrence risk in patients with a first unprovoked VTE. The aim was to determine if reliable prediction models exist and if not what further research is needed in the field. The review identified a number of existing models with various shortcomings, leading to several recommendations for further research, including the need for external validation across multiple settings. Given the findings of the review in chapter 2, a new prognostic model for VTE recurrence risk was developed in chapter 3. The model development aimed to address issues highlighted through chapter 2 in a number of ways including; the use of previously identified predictors, robust statistical modelling, a novel internal-external validation procedure, and appropriate presentation of the final model for future research. Chapter 3 also raised awareness of heterogeneity in predictive performance across settings, and the need to address this formally in further research. Therefore, Chapter 4 aimed to assess the use of recalibration methods to improve the external performance of existing models, in this way previous research can be combined with newly collected information, reducing research waste. A novel IPD meta-analysis approach was used to compare the performance of

recalibration strategies across multiple studies in an applied example; results showed substantial improvements in calibration performance of the existing model in new patient populations via a simple recalibration of the model's intercept term.

Chapters 5 and 6 focused on meta-analysis methods for synthesising continuous predictor effects where primary studies report on various different thresholds. Chapter 5 proposed a novel method to impute missing threshold information between two known threshold values within primary studies. The method improved on previous work in a number of ways, in particular by allowing multiple (rather than single) imputation, and was shown to indicate potential bias in summary meta-analysis estimates in an applied example. Chapter 6 further assessed the new method proposed in chapter 5 through an extensive simulation study. Various settings were simulated by varying the proportion of missing information, the number of studies in the meta-analysis, the disease prevalence and the assumptions of the method. The new method was compared head-to-head against a previous approach and the current standard analysis method, with results suggesting the new method gives benefits over current approaches in most settings, apart from under extreme situations.

7.2 Publications arising from this thesis

The work in this thesis has led to a number of publications, of which the PhD candidate (Joie Ensor) is either the first author or a co-author. A first-author protocol and final review manuscript have been published in *Systematic Reviews* and *BMJ Open* respectively (95, 158), based on the systematic review of prognostic models for recurrent VTE risk conducted in chapter 2. Further, the work conducted in chapter 2 combined with the development of a new

prognostic model in chapter 3 has been published in full in *Health Technology Assessment* (98), with the PhD candidate as the first author. Another manuscript describing the development of the methods proposed in chapter 5, and assessed through simulation in chapter 6, is currently under review at *Research Synthesis Methods*, again with the PhD candidate as the first author. The Stata code for the single imputation approach was part of a paper published in *Systematic Reviews* (236), and the multiple imputation code will be submitted to the *Stata Journal*. Finally, the work described in chapter 4 for the recalibration of flexible parametric survival models, will be drafted for publication and submitted in the near future, with *Statistics in Medicine* the target journal.

Additional publications which the PhD candidate has contributed to from knowledge gained through this thesis include; a recent guide to systematic review and meta-analysis of prediction model performance (71), and an earlier article on the methodological challenges facing systematic reviews of prognosis studies (253), using knowledge from chapter 2. A number of publications based on the use of meta-analysis methods for the development and validation of prediction models in clustered datasets (19, 21), based on knowledge from chapters 3 and 4. Also the candidate has contributed to publications based on meta-analysis methods for multiple and missing thresholds in diagnostic test accuracy meta-analysis based on knowledge gained through the work included in chapters 5 and 6 (234, 236). Finally, the PhD candidate has also contributed to other work in the general area of individual participant data meta-analysis methods whilst conducting the research included in this thesis (120, 254, 255).

7.3 Contribution to applied and methodological research

This thesis has contributed to both applied and methodological research as outlined in the following by chapter and summarised in Figure 7.1.

Chapter 2

The systematic review conducted in chapter 2 identified published prediction models for recurrence risk in patients with a first unprovoked VTE. As the first systematic review of prediction models for VTE recurrence risk, this work adds importantly to the clinical area by identifying all available models and summarising their strengths and weaknesses to help practitioners decide on which if any models to use in practice (95). Risk of bias was assessed using the PROBAST tool to judge the quality of the developed models (92). The review identified a set of predictors that were found to be important within all models, and there is therefore strong evidence that these predictors are associated with risk of recurrent VTE after adjustment. The published evidence also suggested that there are differences in the working definition of an unprovoked VTE population. The lack of a standardised definition for the population limits the applicability of the identified prediction models. Finally the review identified several methodological and reporting issues in the development and validation of prediction models; such as the handling of missing data, and data-driven categorisation of predictors. This adds further evidence of the shortcomings in prognosis and prediction model research, and lends more weight to the need for initiatives such as the TRIPOD statement, to help improve prediction model research in the future (33, 68).

Chapter 3

Chapter 3 built on the review undertaken in chapter 2, to develop a new prognostic model for the prediction of VTE recurrence risk in patients with initial unprovoked VTE. While external validation and head-to-head comparison of the existing models was not possible in the available dataset, the new model could aim to improve on previous models in some aspects. Notably the use of a flexible parametric framework transfers a number of benefits to the new model (44, 45). Flexible parametric (FP) models flexibly capture the shape of the baseline hazard and fully parameterise it, enabling out-of-sample prediction of absolute risks (46). A fully parameterised model enables prediction in new patients, external validation and updating of the model. Previous models were developed using a Cox model, which does not parameterise the baseline hazard (41, 42). Given this the published models can only be used to predict at specific time points for which absolute risks have been provided, and only for groups of patients defined by certain predictor values (37, 40, 161). Assessment of model calibration is particularly difficult where absolute risks cannot be predicted for individuals, with calibration assessment of a Cox model only possible in risk groups defined by the linear predictor (144, 147). In contrast Chapter 3 provides an equation for the baseline survival as well as the linear predictor for the new model, allowing individual prediction at any time point as recommended by the TRIPOD guidelines (33), and thus preparing the new model for future research.

Chapter 3 also highlighted the use of IPD meta-analysis methods for the development and validation of prediction models using data from multiple studies. The internal-external cross-validation (IECV) approach proposed by Royston et al. and extended by Debray et al. allows

use of all the data for both development and validation (20, 66). The Post D-dimer model in chapter 3 was developed using the IECV method, by developing the model on N-1 studies, validating the model in the N^{th} study, and repeating this process N times each time excluding a different study population for validation. This meant that N external validations were possible, providing N sets of performance statistics which could be synthesised using a traditional random-effects meta-analysis. Meta-analysis allows examination of the heterogeneity in the models external performance across studies, and also the potential performance in a new setting, using a 95% prediction interval (19, 21, 24).

Chapter 4

The work in chapter 4 focusses on reducing heterogeneity of the performance of an existing prediction model by updating it in new settings. Results of the case study in breast cancer used in chapter 4, indicated that recalibration and updating methods could substantially improve the calibration performance of an existing FP model in external populations. Previous work has introduced methods for recalibration and updating of survival models in particular case studies, using a Weibull parametric model (144, 145). Chapter 4 contributes to the methodological literature firstly by illustrating the use of recalibration methods using FP models, which offer benefit over and above the Weibull model which can only capture monotonic baseline hazard shapes (as discussed in chapter 1). Secondly the new work used IPD random-effects meta-analysis techniques to compare recalibration methods across multiple validation studies, which in the case study indicated that a simple recalibration of the intercept substantially improved model calibration performance and reduced between-study heterogeneity.

Chapters 5 and 6

As IPD is not always available, Chapter 5 described the development of new methodology for test accuracy meta-analyses suffering from differing numbers of studies providing information on each threshold of interest. The new Multiple Imputation with Discrete Combinations (MIDC) method aimed to improve on a previously proposed Single Imputation (SI) method, which imputes missing threshold information between bounding known threshold information. The MIDC method developed in chapter 5 is more theoretically sound than the previous SI method, as it accounts for uncertainty in the imputed threshold data, and in the distance between known and missing threshold values. The MIDC method imputes missing threshold data multiple times and combines these using Rubin's rules to give the overall summary performance at each threshold. Software code for the method was developed using Stata, and when applied to two examples this showed evidence of potential bias in summary estimates of sensitivity and specificity.

Chapter 6 evaluated the statistical properties of the MIDC method compared to the SI method and the standard analysis approach in which No Imputation (NI) is performed. Simulation results showed substantial gains in the precision of summary performance at each threshold using the MIDC and SI methods, and improved estimation of between-study heterogeneity estimates likely due to increased threshold data gained by imputation. Bias in the summary performance at each threshold was reduced using the imputation methods, particularly when threshold data were selective reported due to publication bias. The MIDC method improved on the SI method by allowing for the uncertainty in imputations leading to slightly inflated

standard errors. Simulations also indicated that the imputation methods performed poorly under an extreme unequal threshold spacing assumption.

The new MIDC method developed in chapters 5 and 6 provides a fast and useful method to assess the potential impact of missing threshold information in diagnostic test accuracy meta-analysis, particularly in the presence of publication bias. Importantly a Stata program to implement the new MIDC method was developed in chapter 5 enabling researchers to use the method in practice.

Key research contributions of the thesis

- Summarised and appraised available prediction models for recurrent VTE risk in unprovoked patients
    - Identified a lack of a standardised definition of unprovoked VTE
    - Identified a set of potential important predictors associated with VTE recurrence risk after adjustment for confounding
    - Identified several methodological and reporting issues in the existing prediction models
- Developed a new prediction model for VTE recurrence risk in the unprovoked population
    - Used flexible parametric modelling framework allowing individual absolute risk prediction in new patients over time
    - Used a novel internal-external cross-validation procedure to maximise the available data for both development and validation of the model
    - Assessed model performance across different patient case-mix, and found moderate heterogeneity in performance across validation populations
    - Reported the full model including baseline survival, for use in future research
- Illustrated the use of model recalibration and updating methods to improve the external performance of an existing prediction model
    - Used data from multiple studies and random-effects meta-analysis methods to externally validate an existing FP model
    - Used four recalibration methods to improve model performance including; re-estimation of the magnitude or shape of the baseline hazard, or re-estimation of the predictor effects, either as a whole or individually
    - Results showed that recalibration methods substantially improved calibration performance of the existing model in new patients
- Developed and evaluated a new method for imputing missing threshold information in diagnostic test accuracy meta-analysis
    - Developed new Multiple Imputation by Discrete Combinations (MIDC) method, which theoretically improves on current methods
    - Developed Stata code to implement the new method
    - Applied to two real examples, and showed that original results (without imputation) may be considerably different to those from MIDC
- Conducted extensive simulation study to evaluate performance of the new method
    - Investigated various scenarios including, varying the amount of missing threshold information, the prevalence of disease, and the missingness assumption
    - The new method was compared head-to-head with existing methods
    - Results indicated that the new method is a useful technique to assess the potential impact of missing threshold data in meta-analysis, particularly in the presence of publication bias
    - Simulations indicated that under the assumption of extreme unequal threshold spacing the imputation methods performed poorly

Figure 7.1 - Key research contributions of the thesis

7.4 Further research

While the work in this thesis has made some important contributions to current clinical research and methodological knowledge as described above, there are various areas for further research arising from the work presented in this thesis, which are now highlighted below.

Chapter 2

The systematic review in chapter 2 highlighted several areas in which future research could be improved (95). Primarily there is an immediate need for a standardised definition of what constitutes an unprovoked first VTE. Without a standard definition of the population of interest, prediction models cannot be developed on the same population in which they are intended for use. Where a model is applied to patients not meeting the same definition as that used in the model development cohort, model performance is likely to be different, with potential miscalibration of predictions (9, 28, 33, 68, 76).

Secondly any future model development studies should aim to utilise predictors that were found to be consistently important after adjustment across all identified models. This set of predictors should also be collected routinely in clinical studies assessing VTE patients, as they have strong evidence of an association with recurrence risk, and future cohorts in which these factors have been collected could be potentially used as model validation datasets.

The review assessed the quality of identified prediction models and found several limitations, primarily the need for further external validation of the models. Therefore further research is needed to validate the existing models in new patient populations to assess their

generalisability (21, 28, 63, 69, 256). It is also important that new research compares the existing models performance in terms of discrimination and calibration, in the same patients, to assess which models are most useful (257). Finally, future research beyond head-to-head comparison and validation studies, should look to assess the impact of the available models on patient and cost-effectiveness outcomes (26, 64, 69, 79).

Chapter 3

There are several areas for further research arising from the new model development conducted in chapter 3, in particular the new Post D-dimer model requires further external validation. While the model was developed using a novel IECV approach, this technique could be seen as somewhere between internal and external validation (33, 34, 68, 256). As such truly external validation in new patients is needed before the model is assessed for it impact on patient and cost outcomes (26, 28, 64, 69). Particularly chapter 3 used IPD from seven randomised controlled trial datasets, and therefore future validation studies should aim to assess the models performance in non-trial populations. Further research could also look to assess whether model updating methods could reduce the between-study heterogeneity in model performance observed during the IECV procedure in chapter 3.

Chapter 3 also attempted to develop a model for use at the time of cessation of therapy, allowing treatment decisions to be made without putting patients at risk of recurrence through withdrawal of treatment. The Pre D-dimer model showed very poor discriminative performance during validation, making it clinically useless. However, future research may look to improve the performance of this model by incorporating additional important predictors, as prediction at cessation of therapy is of most use clinically. D-dimer appears to have

important prognostic value (15, 198, 199), as identified in the Post D-dimer model in chapter 3, therefore future research should investigate the predictive value of D-dimer measured on therapy.

Finally, it is important that future research aims to develop a prediction model for an individual's bleeding risk in this population. Treatment decisions for unprovoked VTE patients must weigh the balance between off-therapy VTE recurrence risk and on-therapy bleeding risks (156, 258). There is no existing model to predict bleeding risks in this population making it difficult to assess this balance in practice.

Chapter 4

Chapter 4 found that recalibration and updating methods were potentially useful to improve model calibration in new patients, which indicates further the potential power of such methods in the fight against research waste in the prognosis and prediction field. The use of recalibration methods could substantially reduce the trend of increasing numbers of new models in the literature where useful models already exist (82, 84, 133). Future work could attempt to evaluate recalibration methods in a range of other case studies, to identify the most appropriate recalibration methods to use in general in practice. Vergouwe et al. recently proposed a closed testing procedure to select recalibration methods for a logistic prediction model (226), a similar procedure could be evaluated for survival models in the future.

Similarly it is important that future research identifies if and when there is greater benefit in developing a new model, compared to recalibrating an existing model. It may be that where large datasets are available for external validation, new model development could have

greater power than the existing models. However this would mean ignoring the previous evidence captured in the existing model, recalibration methods combine information from previous patient cohorts with that from new patients meaning no information is wasted (64, 69). A previous example in a logistic setting indicated that recalibration is preferred if validation samples are small compared to the original model development sample (143).

Chapters 5 and 6

As with any new methodological development the newly proposed MIDC method requires further research to address limitations, and identify alternative approaches that may be possible. The simulations undertaken in chapter 6 tested the MIDC and SI methods under an assumption of extreme unequal threshold spacing and indicated that in this scenario the methods perform poorly. Therefore it is important for further research to investigate firstly how situations of extreme unequal spacing can be detected, and secondly how the results of the imputation methods can be interpreted in such settings.

Both imputation methods provide estimates at particular thresholds of interest, however neither method constrains the ordering of thresholds, which can lead to imperfect SROC curves in practice. Previous studies have addressed this issue by constraining threshold values to be ordered but also have limitations (231, 234).

The MIDC method assumes a linear relationship between threshold and logit-sensitivity/specificity, and the suitability of this assumption and effect of departures from it could be assessed in further research. Importantly other research has assumed a distributional shape for the underlying continuous measurements in both the diseased and non-diseased

populations (235). However it is likely that in practice the true underlying distribution of the test could not be assessed without IPD over the range of the test, which is unlikely to be reported.

Finally, there are now a number of methods available for handling multiple and missing threshold data in test accuracy meta-analysis (230-236, 238, 239), and it is therefore important that future research aims to compare these methods head-to-head on the same data and in various settings in an extensive simulation study.

7.5 Limitations of IPD meta-analysis

The majority of this thesis used individual participant data (IPD) from multiple studies. IPD provides the raw patient level information from the original studies, as opposed to aggregate data which is summarised patient data, such as mean age or proportion of males in the study. IPD meta-analysis methods were used in this thesis for the development, validation and updating of prognostic models. IPD meta-analyses offer many opportunities for prognosis research, however it is also important to highlight their limitations (259-261).

Firstly, the quality of an IPD meta-analysis is only as good as the quality of the primary studies included (260). As such, when primary studies are of poor quality an IPD meta-analysis is not necessarily more reliable than an aggregate data meta-analysis of the same set of studies. Furthermore, there is no guarantee that studies willing to share their data are the studies of highest quality, and perhaps the only way to guarantee quality is to set up a prospective IPD meta-analysis (262-265). Therefore, IPD meta-analyses should ideally include some form of

quality assessment to highlight differences across studies (that do and do not provide IPD) and consider the potential impact of quality on conclusions.

Secondly, akin to the quality of evidence from primary studies is the potential for availability bias in IPD meta-analyses. Availability bias describes the potential bias caused when the provision of studies IPD is somehow dependent on the study's results (122, 260). For example, smaller studies with non-significant results may be less likely to be published and therefore harder to identify for inclusion (259). Even where these studies are identified, authors may no longer have the IPD, or be unwilling to collaborate. In this way it is common for IPD meta-analyses to obtain IPD from less studies than there are published summary results (266); where this is the case methods to combine IPD and aggregate information may be utilised (267, 268).

Thirdly, it is important to consider before undertaking any IPD meta-analysis, the substantial investment of time and costs involved (122, 260, 266). For example the time taken to contact primary study authors, to obtain and clean the IPD from all studies, and to prepare the data for meta-analysis ensuring consistency across studies.

Many other limitations exist, such as the between-study heterogeneity in follow-up lengths, methods of measurement, outcome definitions, and factors recorded (269). Some of these were prevalent in the IPD meta-analysis of Chapter 4, and warrant further research.

7.6 Opportunities with Big data

The increasing availability and use of 'big' datasets from IPD meta-analyses, multicentre collaborations, and patient Electronic Health Records (EHR) data provides a unique opportunity to improve prediction model research (21). This thesis has focused on the use of IPD from multiple studies in prediction model research, but the methods used here could be applied to other big datasets in which patients are clustered by centre, hospital, region, country or some other clustering factor. Of the phases of prediction model research outlined in chapter 1 (3, 8-10), one area in particular that could potentially gain the most from availability of big data is external validation (21). It has been discussed that there are ever increasing numbers of new model development studies being published without a corresponding increase in external validation studies (9). This leads to research waste, with many models available to predict the same outcomes (82, 84, 133), and little to indicate which models if any to use in practice (78).

One reason for a lack of corresponding external validation studies is the availability of data, with development studies often including only one cohort making it most efficient to retain all patient data for development of the model (62). But with big datasets this is no longer an issue, as such datasets can be split into development and validation samples without reducing the power to detect genuine predictor-outcome associations. And with clustered data it is possible to assess the developed models generalisability in external patient data using non-random splitting by cluster (e.g. by centre and country as in chapter 4) (33). In large clustered datasets model performance can be tested in multiple different external populations, allowing assessment of variation in performance due to differences in patient case-mix (70). Such

variation in performance across validation populations can be examined using meta-analysis methods as shown in this thesis and elsewhere, enabling investigation of the causes of heterogeneity in model performance (19-21). Further, this interrogation of heterogeneity in model performance can help to identify where, and how performance can be improved in particular clusters using recalibration methods as shown in chapter 4 of this thesis.

Big datasets therefore offer a great opportunity to enhance the reliability of prediction models, increase uptake of such models into practice and reduce research waste by getting the most out of existing research and all available patient data (21). However, similar to that raised above for IPD meta-analysis, EHR data may come with many methodological challenges to be addressed (12, 252, 269, 270). In particular the quality of EHR data is often poor in comparison to primary research studies because it is routinely collected data input by general practitioners (GPs) and not originally intended for research (271). This leads to several issues including; non-standard definition of patient outcomes and diagnoses, unrecorded predictors, incomplete follow-up and missing patient level data. Data may be systematically missing, where all information for a predictor is not recorded in a particular cluster of the data, and in this case novel methods are needed to impute such data to enable validation in the cluster (123-130).

There are also substantial time and monetary costs involved in obtaining and checking large datasets of EHR or IPD from multiple studies (21). One of the benefits of big data highlighted above was the ability to split such datasets into development and validation samples, but this also requires care. Where few clusters are available as in IPD meta-analyses, novel methods such as the IECV approach used in this thesis could be employed to maximise the use of the

data for development and validation (20, 66). In large EHR datasets there may be many clusters meaning a large number could be reserved for validation, ensuring testing is possible in populations of differing patient case-mix (21). Where model development is of interest care must be taken with large EHR datasets containing millions of patient records, as spurious predictor-outcome associations are likely to be falsely identified as important. Predictor-outcome associations could be investigated across clusters to identify potentially heterogeneous predictor effects for exclusion, which could reduce heterogeneity in the performance of the final model (20).

7.7 Conclusions

Prognosis research aims to understand, explain and predict future outcome risk for individuals with a particular disease or health condition, in order to inform and improve patient care through better decision making and stratified medicine. The prognosis field is historically plagued by poor reporting and methodology, leading to a concerted effort in recent years to provide guidance on best practice in prognosis research, which should lead to future improvements (1, 3, 8-10, 25-28, 33, 68, 93). This thesis has contributed toward these improvements in terms of both applied and methodological work, by encouraging and facilitating the use of meta-analysis methods to summarise predictive performance, especially when large, combined datasets are available (21, 252). It is hoped that the use of such evidence synthesis methods (19, 20, 24), alongside robust statistical models for absolute risk prediction (27, 44, 45), will help improve the development, validation and updating of prediction models in the coming years, and thereby improve patient outcomes.

APPENDIX

APPENDIX A: Chapter 2 Appendices

Full text inclusion criteria

CRITERIA	Yes	C/T	No	Prognostic model
Reviews & Discussions				Does the study do more than just discuss a model
Population				Are patients at least 18 years old
				Could the population or a defined subpopulation be considered as unprovoked (if not why not)
				Can we identify results specifically for the unprovoked population
				Did patients receive at least 3 months treatment with either a vitamin K antagonist or an oral anticoagulant
Outcome				Does the model predict least one of: recurrence/mortality/bleeding/QoL
Models				Does the model aim to do more than assess a single factor adjusted for other things
				Is the model used to predict risk of one of the above outcomes
DECISION				
Exclude *with reason*				
Does the study include an economic evaluation of a model?				
Comments				
IF INCLUDED:	**Yes**	**C/T**	**No**	**Factors**
If the study is included, what is their definition of unprovoked?				Major surgery
				Lower limb trauma
				Use of oral contraceptive pill or hormone replacement therapy
				Pregnancy
				Significant immobility
				Cancer
				Thrombophillia (e.g. antiphosphlipid syndrome, factor V leiden etc.)

Unprovoked = no history (within 3 months) of major surgery; lower limb trauma e.g. fracture, cast; use of the combined oral contraceptive pill or hormone replacement therapy ; pregnancy; significant immobility e.g. confined to bed for 3 days; cancer

A "yes" in all categories under CRITERIA indicates to include a study, any "no's" indicate exclusion

APPENDIX B: Chapter 3 Appendices

APPENDIX B1: Summary characteristics of the RVTE database

Within this section a description and summary of the individuals and candidate predictors in the RVTE database is presented.

Description of data

Summary statistics for the baseline patient characteristics and available predictors in the RVTE database are described in Table 0.1, and show that across the whole database there were 230 recurrent events out of 1634 patients with a first unprovoked VTE. There were some trials with very small numbers of recurrences, for example Tait (218) and Shrivastava (216), with 17 and 9 recurrent events respectively. Other trials were larger, with the Eichinger trial having the largest number of events and patients, with 69 recurrent events out of 391 patients. The exclusion of hormone related index VTE events, inline with the definition of unprovoked VTE within the study (*see section 3.2.2*), showed that there were 14 hormone related recurrent events excluded.

The median follow-up across all seven trials was 22 months, with the longest follow-up being almost 10 years in the Eichinger (203) trial, giving sufficient follow-up time to yield meaningful conclusions from the prognostic model.

Summary statistics for each of the candidate predictors are also presented in Table 0.1, with continuous predictors described as means and standard deviations, and categorical predictors described as counts and percentages. Across the seven trials, patient age appeared to be similar with an overall average age of 61 years for the whole population. Treatment duration appeared generally similar across trials, with an average across trials of around 12 months, and the greatest average treatment duration seen in the Palareti 2006 (204) trial of 21 months. D-dimer levels appeared to have large variability, with large interquartile ranges, and three trials (Poli (214), Eichinger (203) and Shrivastava (216)) having noticeably lower D-dimer levels. There may be significant outliers causing the large variation seen in D-dimer levels recorded in each trial, and this was investigated in the exploratory analysis (see section 0). The mean BMI across the seven trials was around 28 kg/m^2, however there were a large proportion of missing data across the trials for BMI. There was also missing data from one trial for lag time, but overall across the trials there was an average lag time of around 38 days, with the greatest mean lag time being 143 days in the Shrivastava (216) trial. The percentage of males and females were consistent across the trials (see Table 0.1), as were the proportions of index site, except for distal DVT where the Eichinger (203) trial had a noticeably greater proportion of patients with a first distal DVT, possibly explained by differences in the inclusion and exclusion criteria used in the original studies which collected the data (*see Table 0.2*).

Trial	Palareti 03	Palareti 06	Poli	Tait	Eichinger	Baglin	Shrivastava	Total
Recurrences/Total	31/280	38/438	26/156	17/100	69/391	40/178	9/91	230/1634
Follow-up (months)								
Median	20.8	20	24	22.2	28.6	37.6	26	22.1
Longest	31.4	37.2	96	41.6	119.2	70.9	51.2	119.2
Candidate predictors								
Age (Years)*	70.1 (12.3)	64.5 (13.4)	62.9 (15.2)	60.9 (13.8)	54.1 (15)	64.6 (16.6)	55.4 (12.6)	62.1 (15.2)
BMI (kg/m^2)*	-	-	-	28.9 (6.5)	28 (4.8)	26.9 (6.6)	32.3 (7.2)	28.5 (6)
Treatment duration (Months)*	7.5 (6.2)	21.1 (104.7)	14.9 (11.2)	5.8 (0.9)	8.2 (11.2)	6.3 (0.9)	7.9 (5.2)	11.8 (54.9)
D-dimer$^{\#}$ (ng/mL)	500 (310, 1012.5)	540 (330, 860)	247 (201, 356)	541.5 (306.5, 979.5)	347 (230, 557)	687 (423, 975)	350 (200, 660)	445 (272, 792)
Lag time$^{\#}$ (Days)	28 (24, 33)	31 (29, 35)	30 (30, 30)	26 (22, 35)	21 (16, 27)	-	48 (30, 227)	30 (23, 33)
*Gender^*								
Female	128 (45.7)	179 (40.9)	55 (35.3)	37 (37)	147 (37.6)	65 (36.5)	24 (26.4)	635 (38.9)
Male	152 (54.3)	259 (59.1)	101 (64.7)	63 (63)	244 (62.4)	113 (63.5)	67 (73.6)	999 (61.1)
*Site of index event^*								
Proximal DVT	217 (77.5)	274 (62.6)	96 (61.5)	60 (60)	148 (37.9)	107 (60.1)	60 (65.9)	962 (58.9)
Distal DVT	12 (4.3)	0 (0)	0 (0)	0 (0)	88 (22.5)	0 (0)	12 (13.2)	112 (6.9)
PE	51 (18.2)	164 (37.4)	60 (38.5)	40 (40)	155 (39.6)	71 (39.9)	19 (20.9)	560 (34.3)
Unspecified DVT	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)

Table 0.1 - Summary of baseline characteristics and candidate predictors

NB: *Mean (standard deviation); $^{\#}$Median (LQ, UQ); ^Count (percentage)

Table 0.2 - Inclusion and exclusion criteria of trials within the RVTE database (15)

Trial	Inclusion criteria	Exclusion criteria
Palareti 03	First VTE	Lupus Anticoagulant
Palareti 06	First unprovoked VTE	Recent pregnancy or puerperium, leg fracture, immobilisation for > 3 days, surgery, APS, active cancer, antithrombin deficiency, serious liver or renal disease, other indication or contraindication for anticoagulation, limited life expectancy, geographic inaccessibility
Poli	First unprovoked VTE	APS, Active cancer
Tait	Acute VTE (last 5 weeks)	Life expectancy < 3 months, anticipated duration of OAC > 1 year, unavailable for follow-up
Eichinger	First unprovoked VTE	Surgery, pregnancy, or trauma in previous 3 months, Cancer, APS, Natural coagulation inhibitor deficiency, long-term anticoagulation
Baglin	First VTE	Postoperative or pregnancy associated VTE, APS, Cancer, thrombosis within 6 weeks of surgery, other indication for prolonged anticoagulation
Shrivastava	First unprovoked VTE	Surgery or trauma within 90 days of first VTE, APS, previous or active cancer, life expectancy < 3 years

A summary of the percentage of missing data across the trials, and as a whole, is presented by candidate predictors in Table 0.3. As mentioned previously there was a large amount of missing data across the trials for the candidate factor BMI, with around 57% of BMI data missing over the whole database. This mostly consisted of three trials (Palareti 2003 (217), Palareti 2006 (204) and Poli (214)) where patient BMI data was not originally recorded, but there was also a significant amount of missing BMI data in the Baglin (198) trial (27% missing). Across the trials there was also some missing data on D-dimer values and lag time, with 15% and 11.4% missing respectively. There was a large percentage of missing D-dimer values in the Palareti 2006 (204) (38%) and Poli (214) (48%) trials. Lag time data was not available by individual patient within the Baglin (198) trial, though D-dimer was reported to have been measured between one to two months after cessation of therapy (198). No missing data was present for age, gender, treatment duration or site of index event variables.

Table 0.3 - Percentage of missing data for candidate predictors

Trial	Percentage of missing information (%)							
Candidate factors	**Palareti 03**	**Palareti 06**	**Poli**	**Tait**	**Eichinger**	**Baglin**	**Shrivastava**	**All**
Age	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
BMI	100.0	100.0	100.0	0.0	1.8	27.0	0.0	56.9
Treatment duration	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
D-dimer	0.0	38.4	48.1	0.0	0.0	0.0	2.2	15.0
Lag time	0.0	38.4	48.1	0.0	1.0	100.0	5.5	11.4
Gender	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Site of index event	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Other candidate predictors were known in all studies for at least some patients. As there was a degree of missing data in these candidate predictors (D-dimer, lag time, treatment duration) a sensitivity analysis imputing any missing information was considered. Therefore, as mentioned in the Methods section above (*see section 3.2.6*); prognostic models were developed based firstly on a complete case scenario, and secondly on a scenario using multiple imputation techniques to impute missing patient data.

Distribution of candidate predictors, correlation and outliers

An exploratory analysis was performed on each of the candidate predictors firstly considering their empirical distributions and assessing these for normality using histograms and normal probability plots (*see 0*), with transformations considered as appropriate where there were departures from normality. Possible outliers were inspected, with erroneous patient values leading to removal of patient data, and outliers deemed to be extreme (but plausible) considered for sensitivity analysis to assess their effect on the final model.

Associations between the candidate predictors were investigated using scatter plots (*see 0*) and correlation statistics (*see Table 0.4*) for continuous factors, and box plots (*see 0*) to assess the relationship between categorical and continuous factors.

The candidate predictors of patient age and BMI were found to be approximately normally distributed, with some extreme values identified (patient ages of zero, and BMI values lower than 10) which were removed from the dataset as erroneous data. D-dimer score, lag time and treatment duration were all found to have a strong positively skewed empirical distribution, and a log transformation was therefore considered in order to approximate normality (for histograms and normal plots of transformed factors *see 0*). Patients with treatment durations above 1000 days were removed as this was considered erroneous data based on clinical expertise.

Continuous candidate predictors were examined visually using scatter plots (*see Figure 0.17*) and empirically using correlation coefficients (*see Table 0.4*). It is clear from Table 0.4 that there were low to moderate correlation between the continuous predictors, and visual inspection of the scatter plots confirmed these findings. The strongest correlation was between age and log D-dimer, which was 0.5 (*see Table 0.4*). Investigation of relationships between continuous factors and categorical factors for gender and site of index event was undertaken using box plots (*see Figure 0.18 to Figure 0.29*). Across the five continuous predictors considered for inclusion there appeared to be no distinct differences between males and females, or between proximal DVT, distal DVT and PE based on visual examination of the box plots. There were several outliers observed in the box plots, particularly for treatment duration and lag time, but also in the other candidate factors. When establishing the final prognostic models a sensitivity analysis was performed by excluding any outlying values for any predictor and checking the robustness of the model to these extreme values.

Table 0.4 - Correlation coefficients between continuous candidate predictors

Candidate factors	Age	BMI	Log D-dimer	Log lag time	Log treatment duration
Age	1.00				
BMI	-0.03	1.00			
Log D-dimer	0.50	0.02	1.00		
Log lag time	0.02	0.14	0.08	1.00	
Log treatment duration	-0.13	0.06	-0.06	-0.02	1.00

APPENDIX B2: Exploratory analysis figures

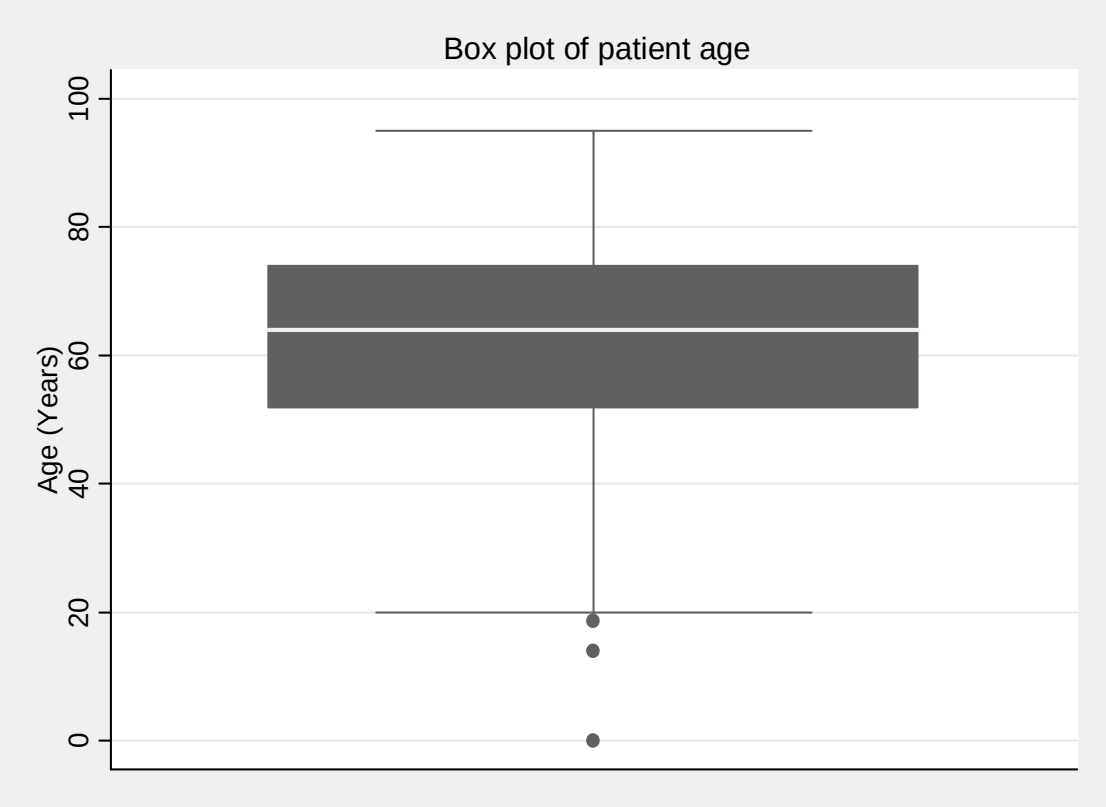


Figure 0.1 - Box plot of patient age (years)

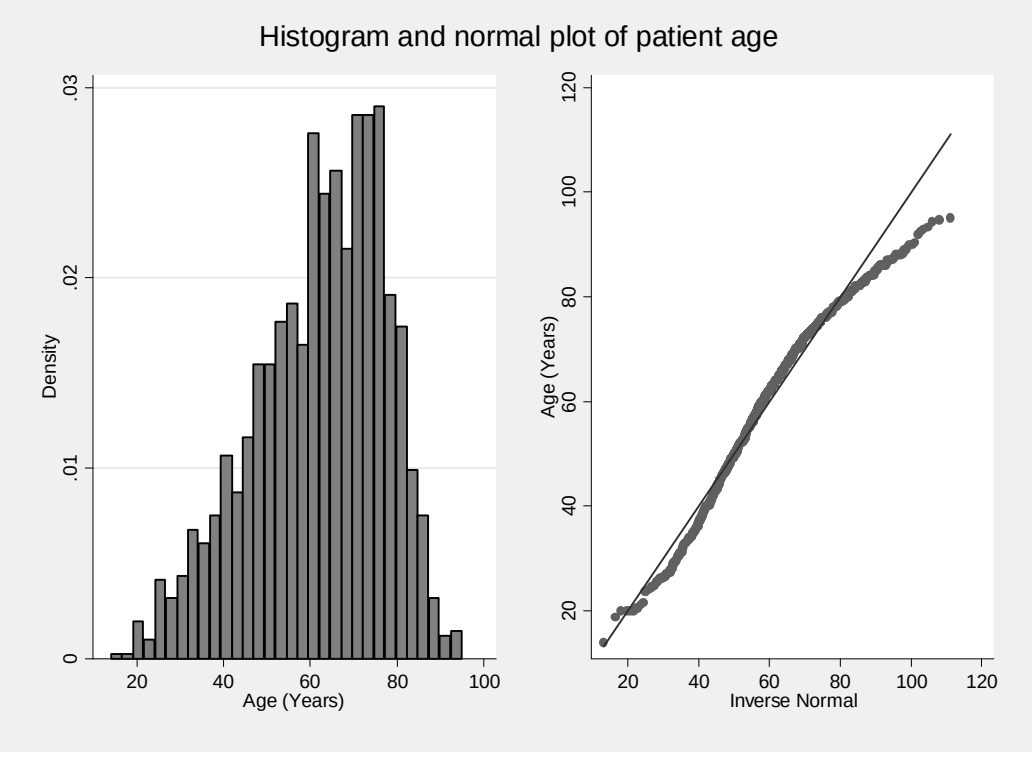


Figure 0.2 - Histogram & normal plot for patient age (years)

Figure 0.3 - Histogram & normal plot for patient age squared (years-squared)

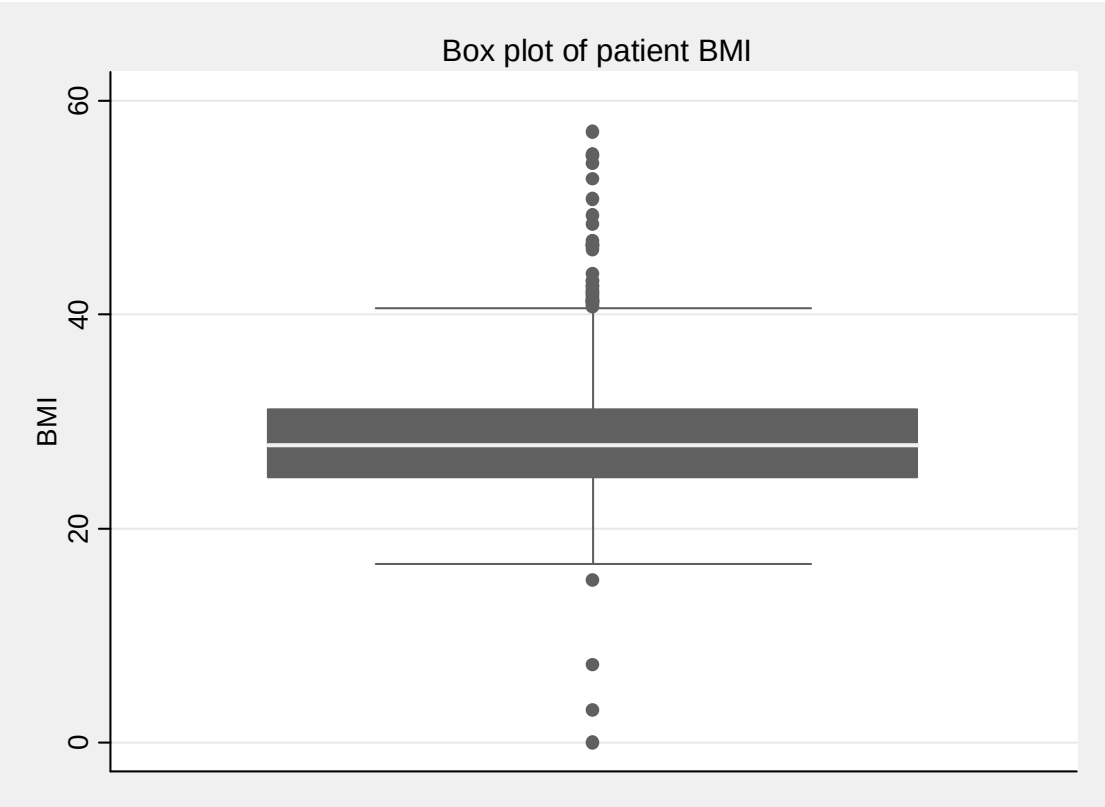


Figure 0.4 - Box plot for patient BMI

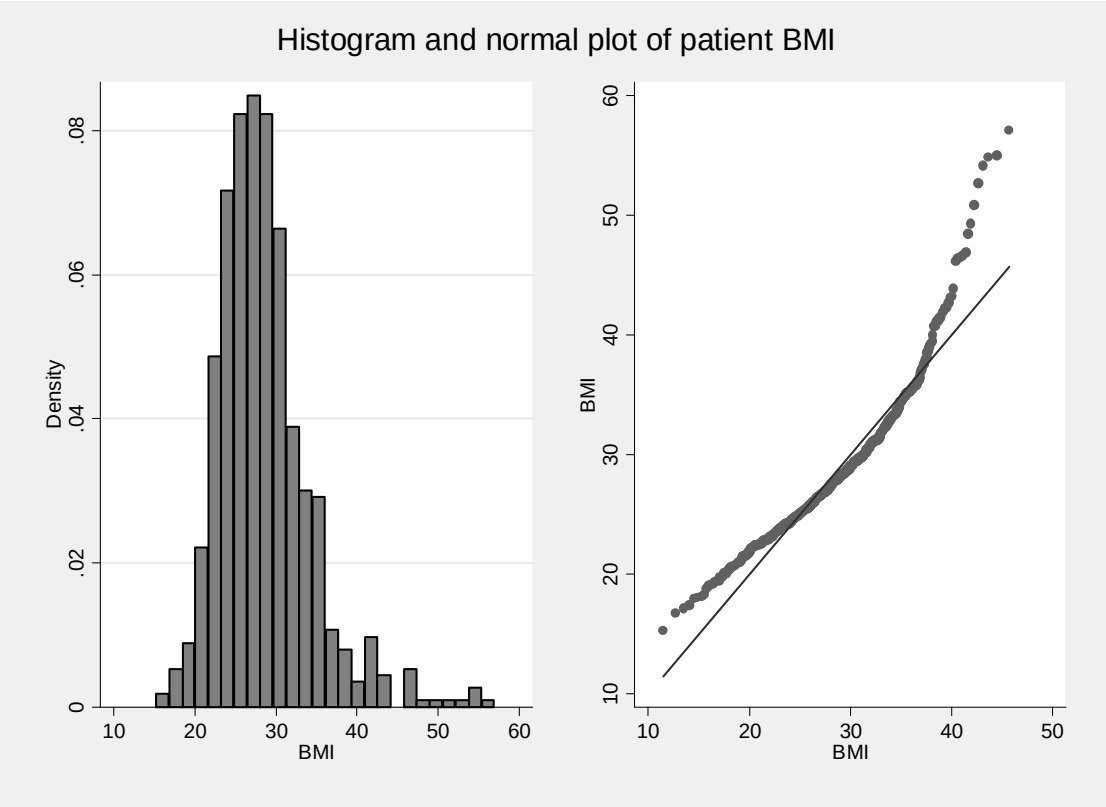


Figure 0.5 - Histogram & normal plot for patient BMI

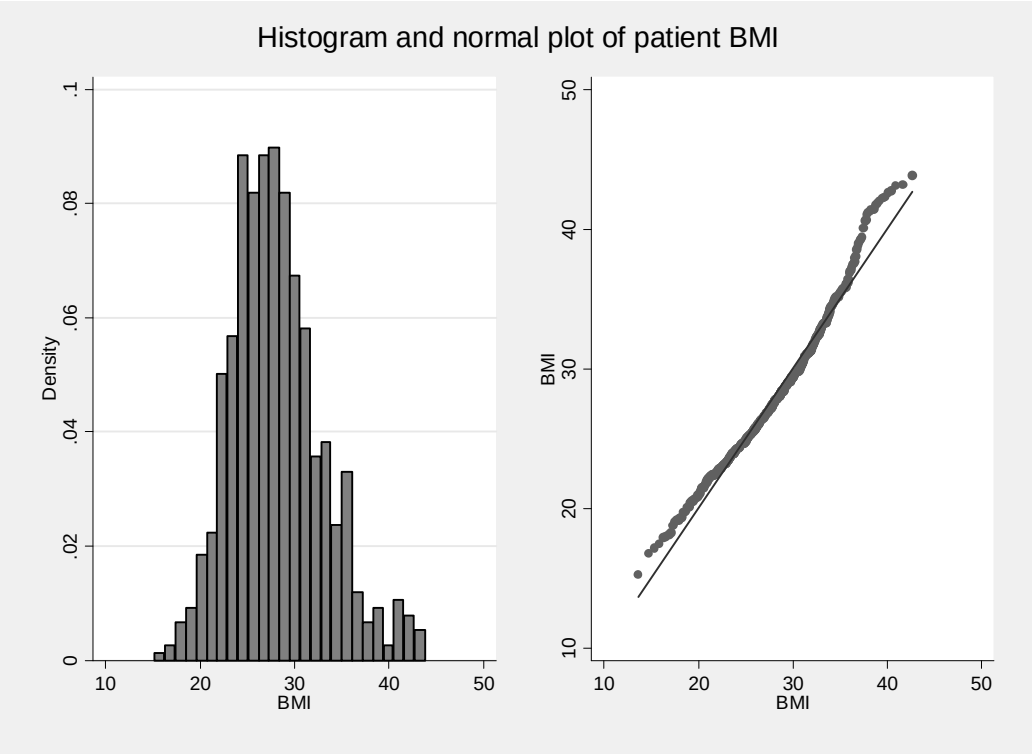


Figure 0.6 - Histogram & normal plot for patient BMI (BMI > 45 removed)

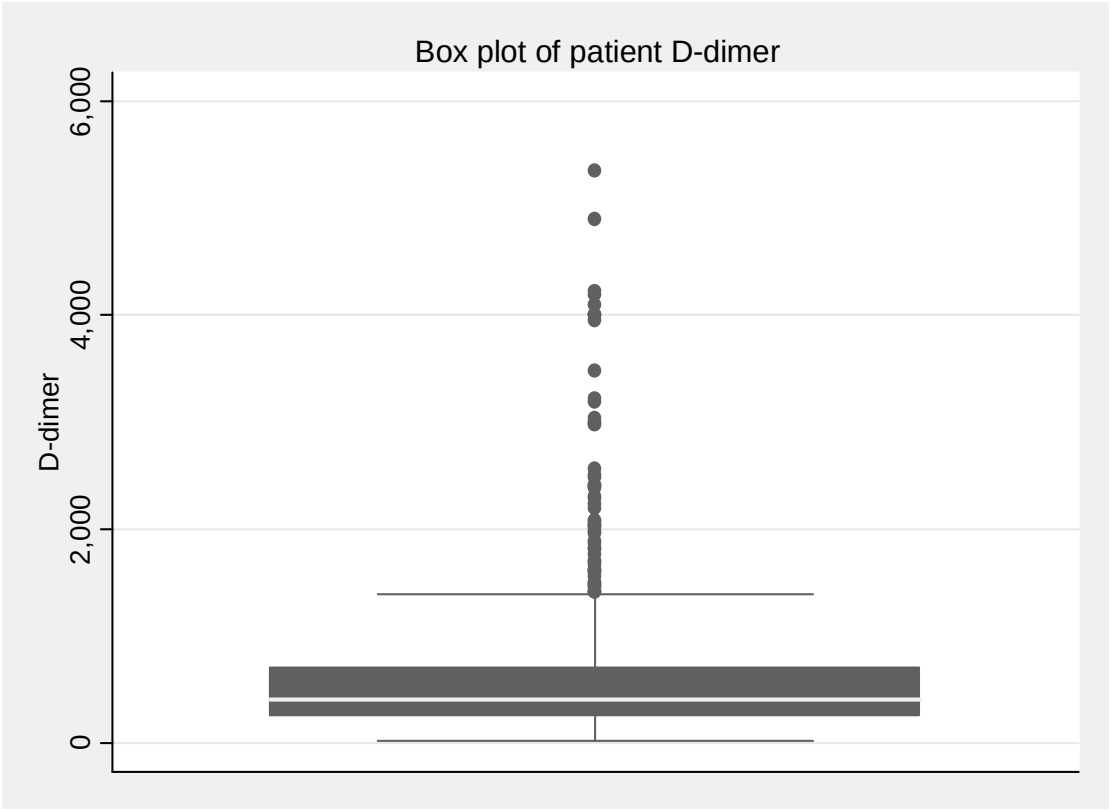


Figure 0.7 - Box plot for patient D-dimer score (ng/mL)

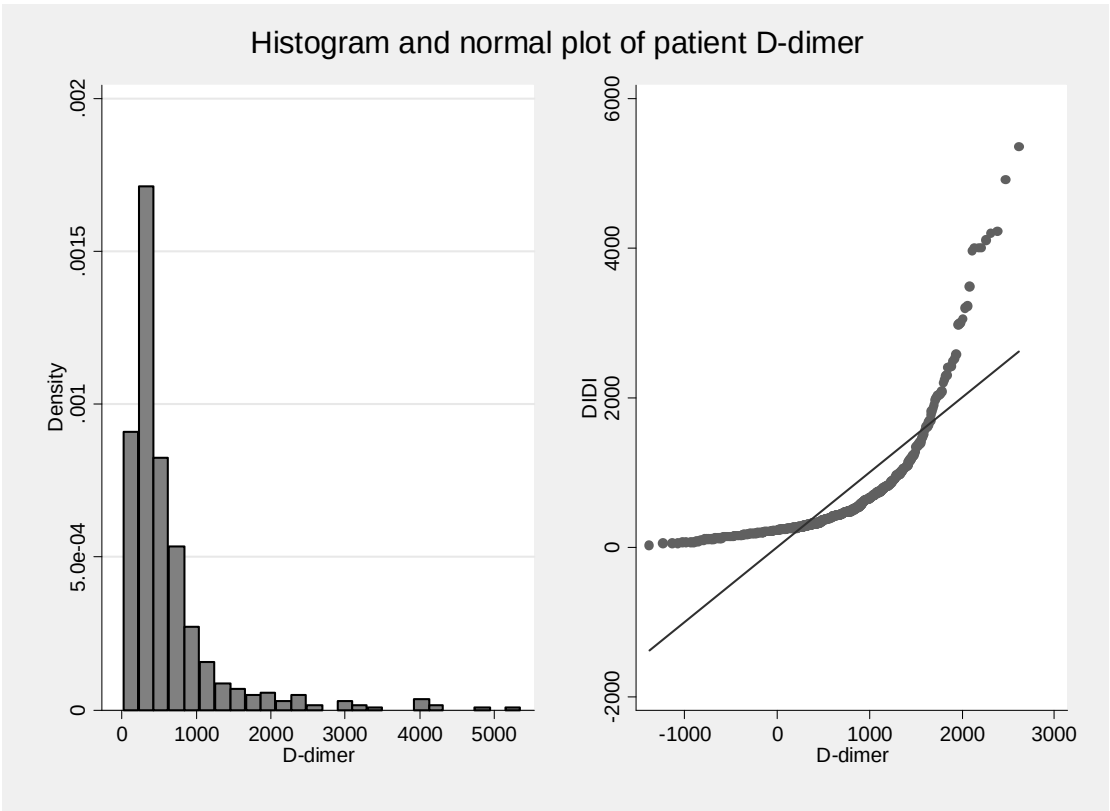


Figure 0.8 - Histogram & normal plot for patient D-dimer score (ng/mL)

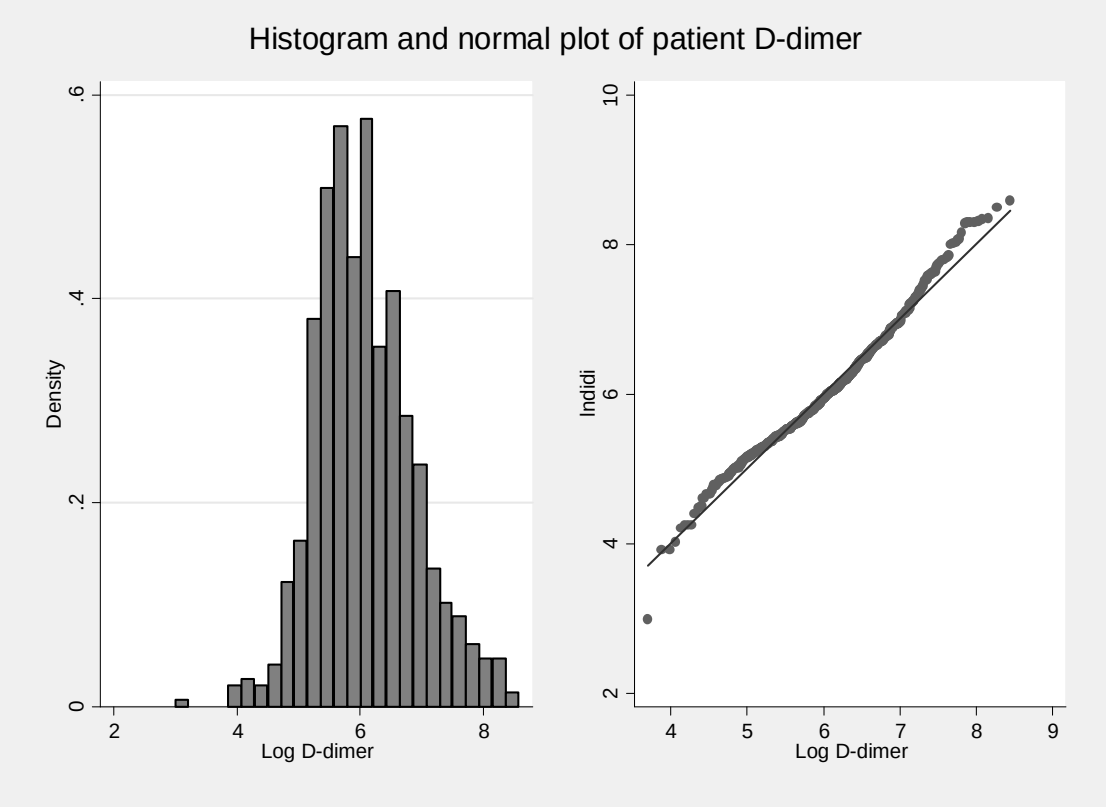


Figure 0.9 - Histogram & normal plot for patient Log D-dimer score (ng/mL) [Outlier - D-dimer=20]

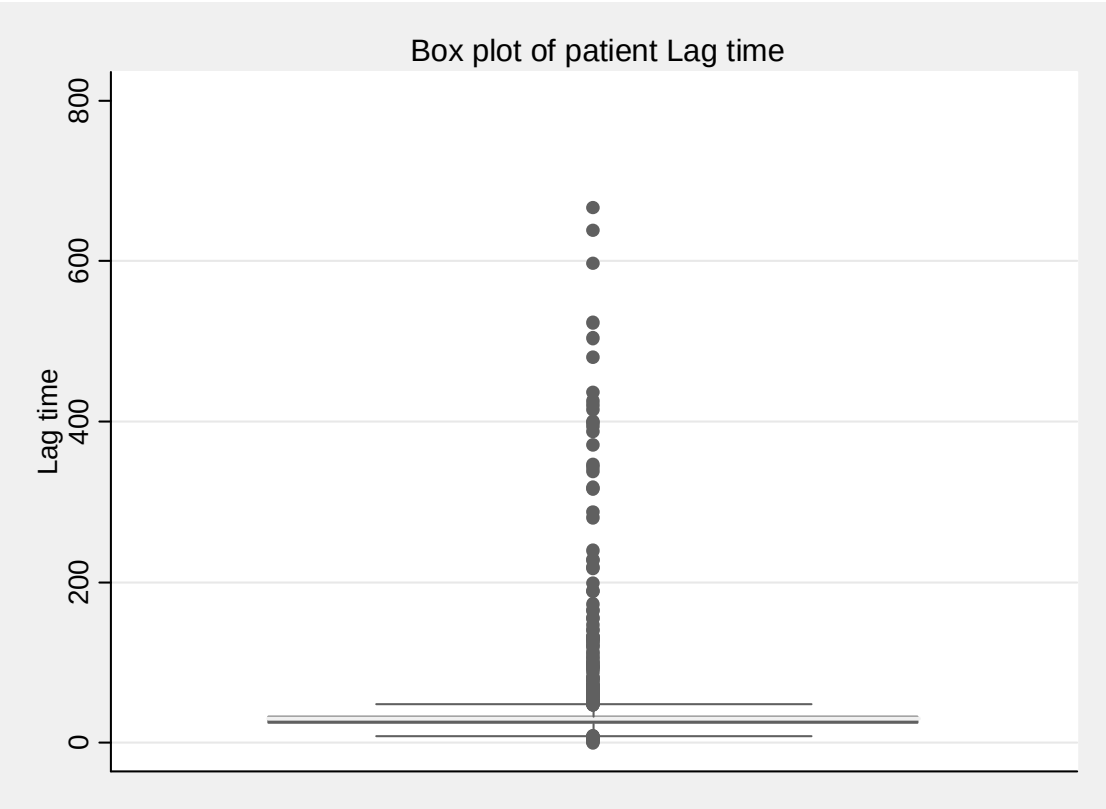


Figure 0.10 - Box plot for patient lag time (days)

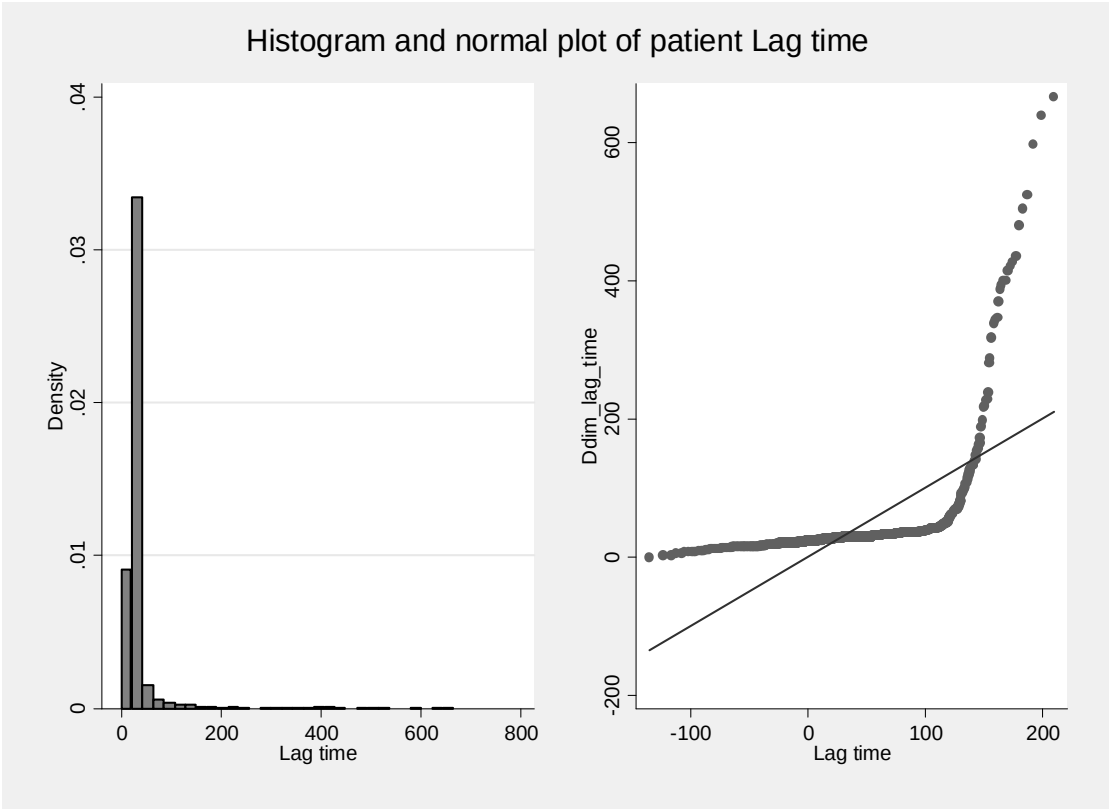


Figure 0.11 - Histogram & normal plot for patient lag time (days)

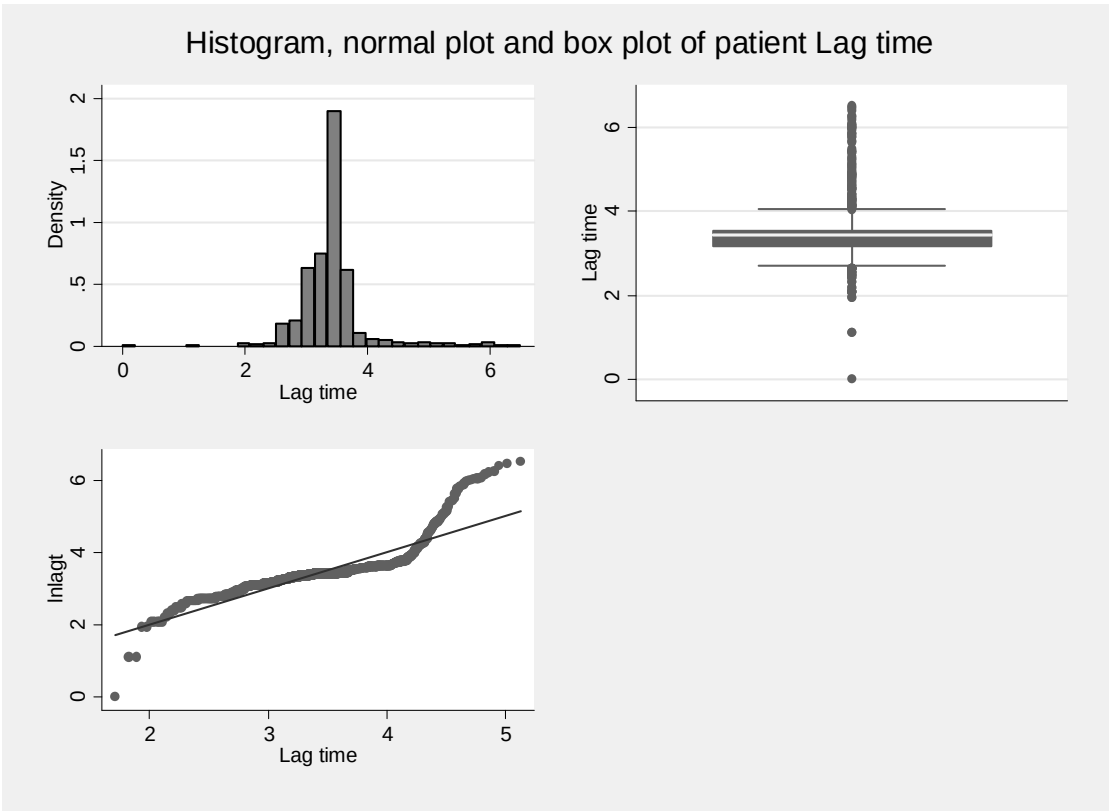


Figure 0.12 – Histogram, box plot & normal plot for patient Log lag time (days)

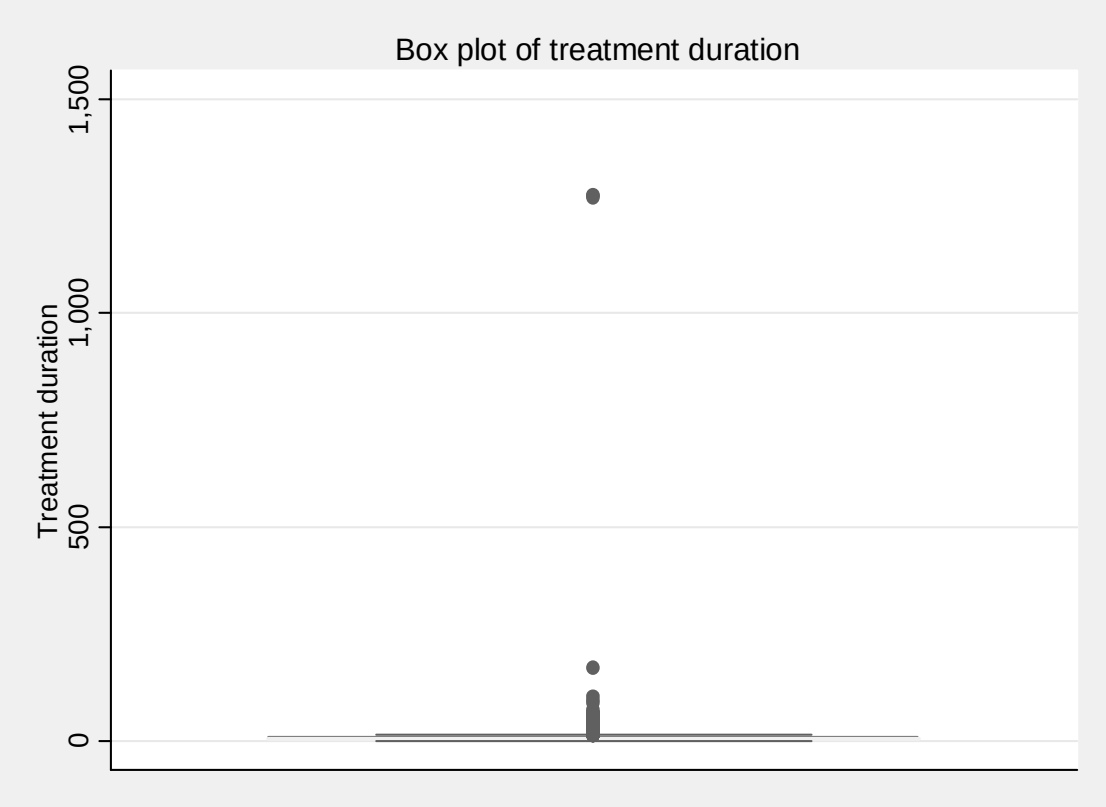


Figure 0.13 - Box plot for patients treatment duration (months)

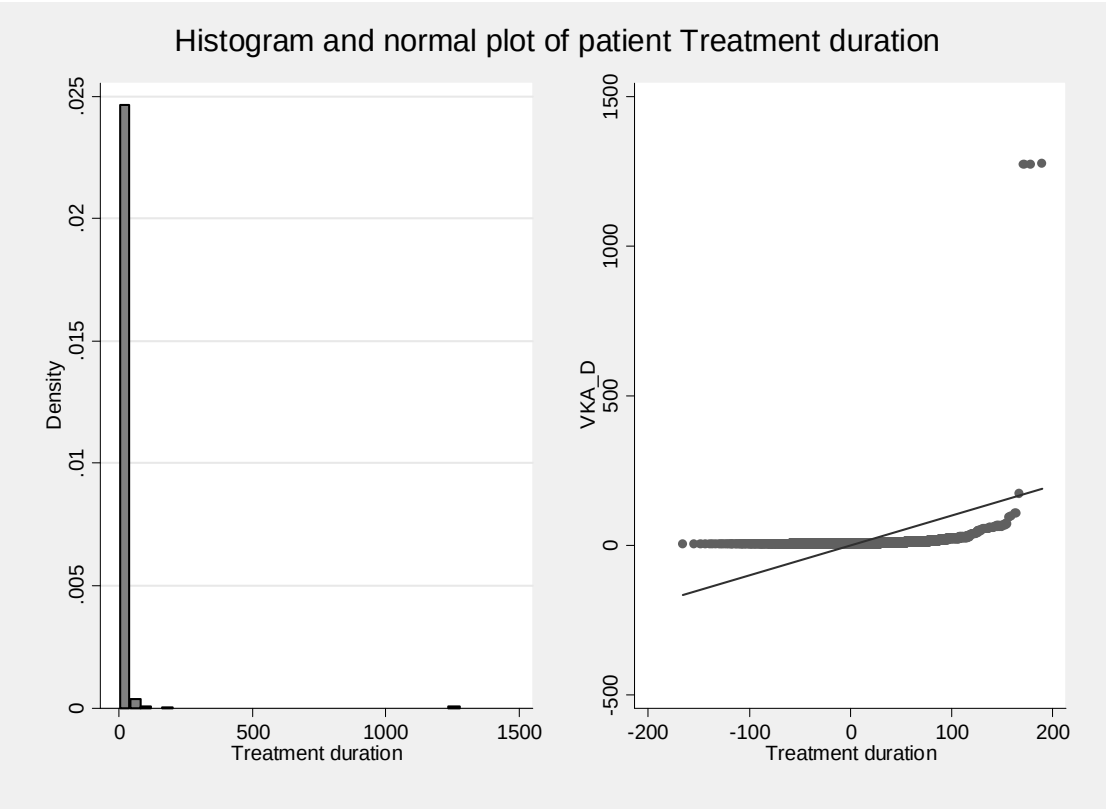


Figure 0.14 - Histogram & normal plot for patients treatment duration (months)

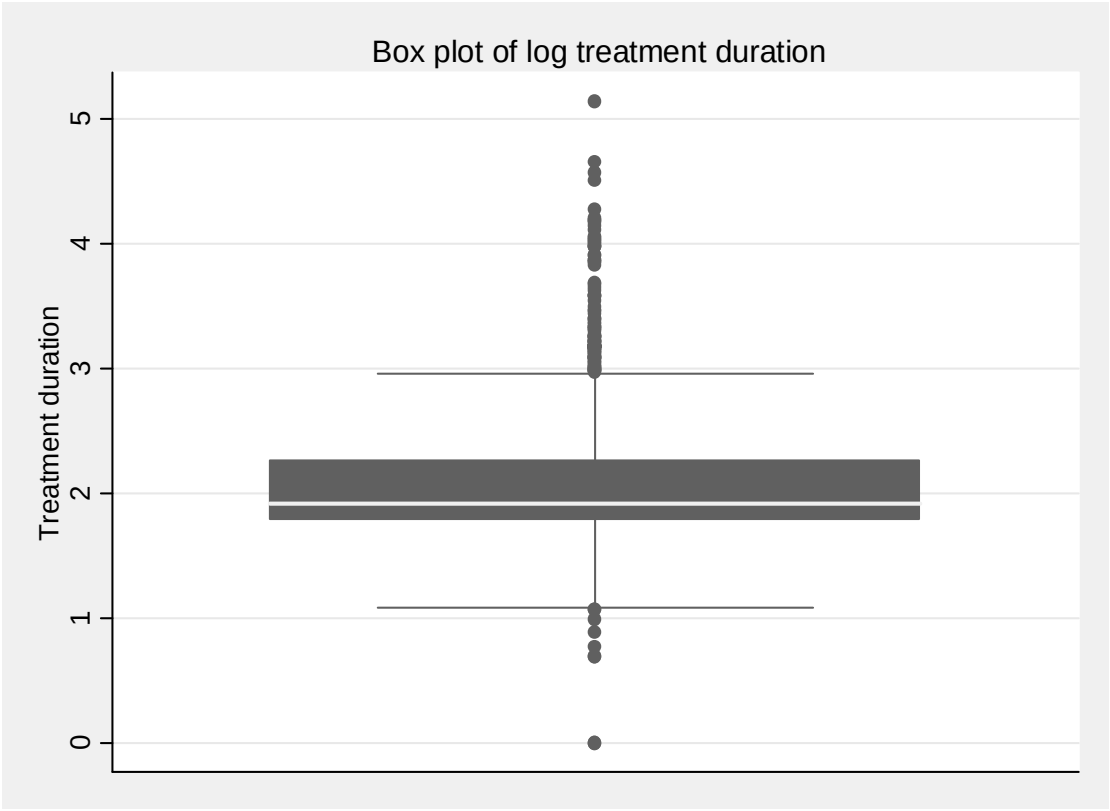


Figure 0.15 - Box plot for patients Log treatment duration (months)

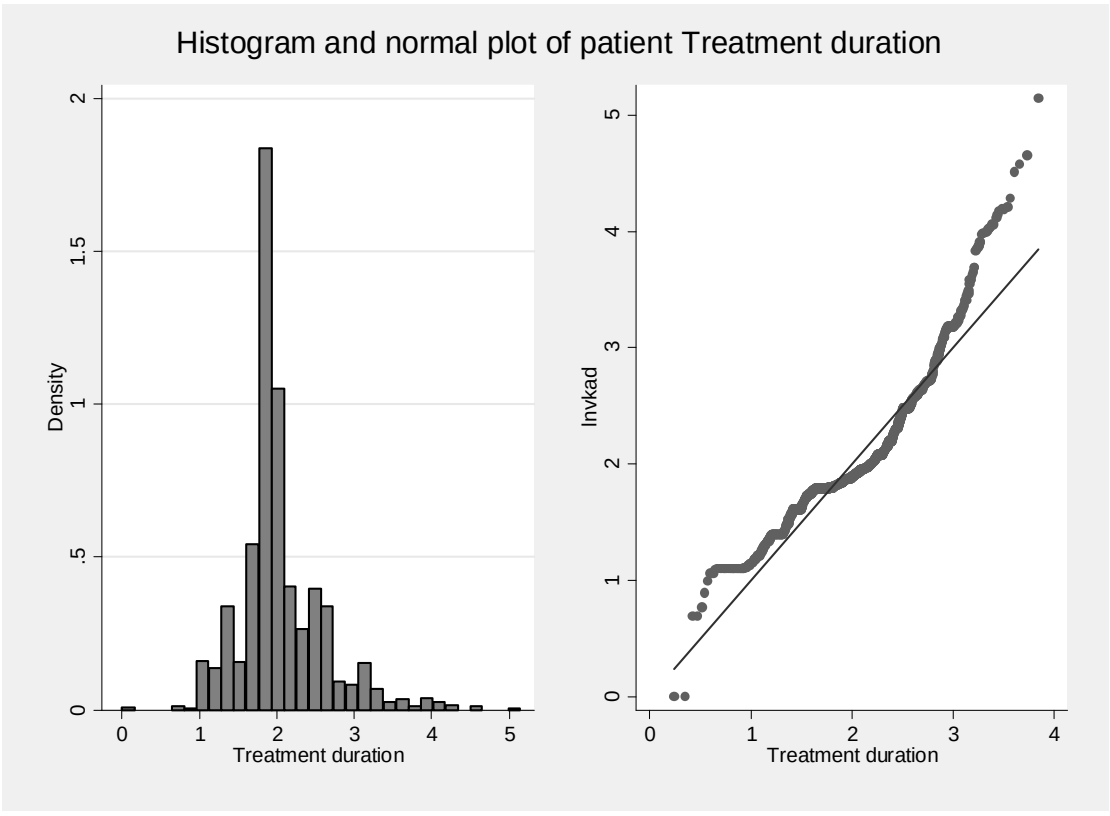


Figure 0.16 - Histogram & normal plot for patients Log treatment duration (months) [treatment durations > 1000 months removed]

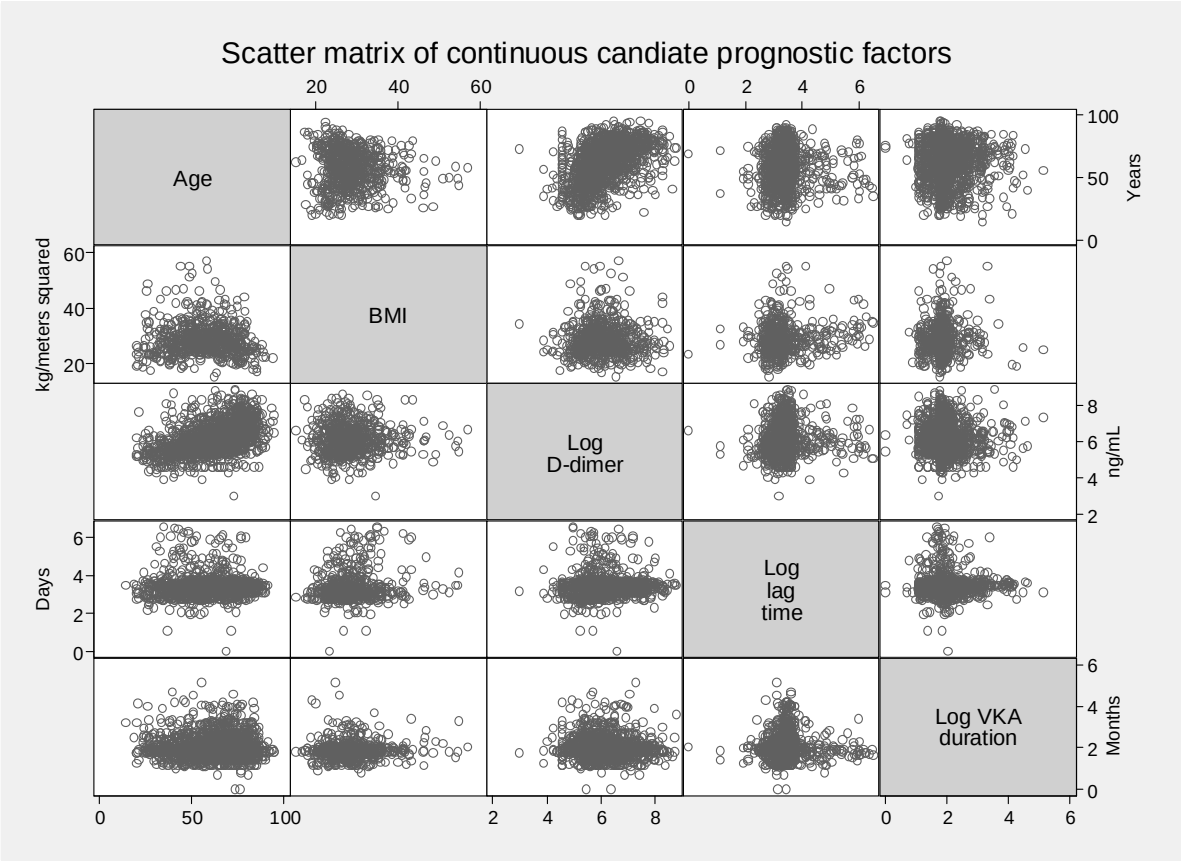


Figure 0.17 - Scatter plots of continuous candidate factors

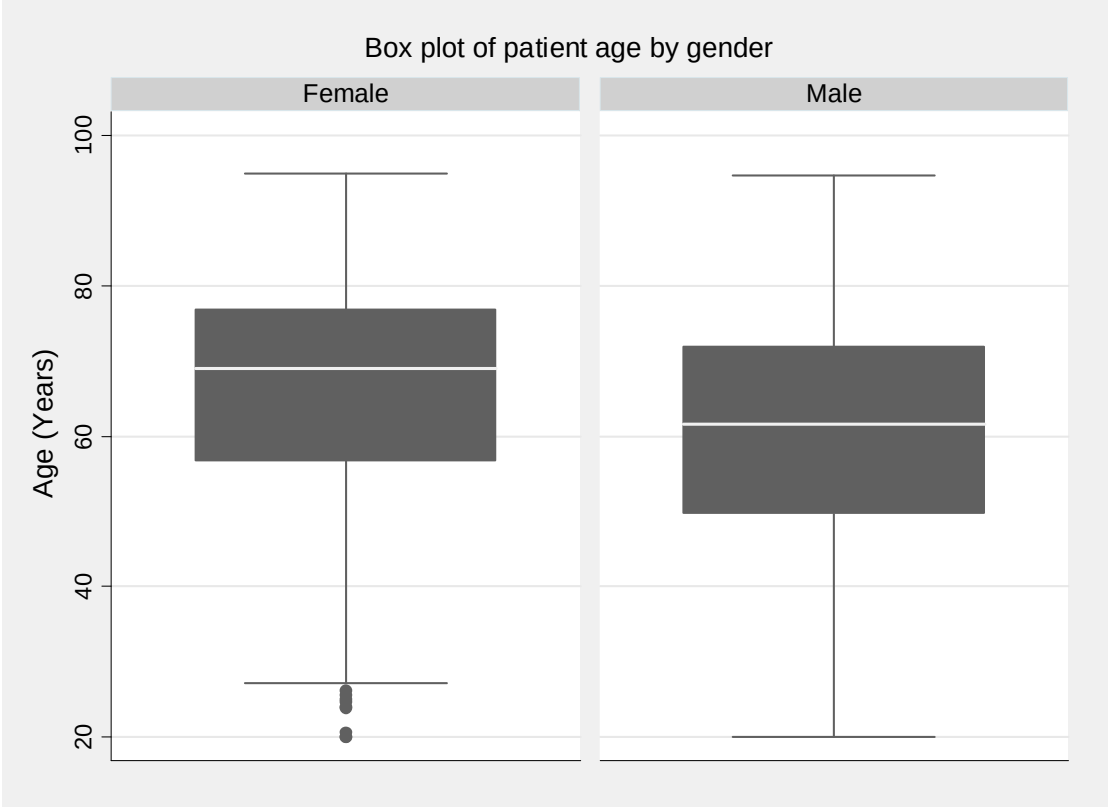


Figure 0.18 - Box plots for patient age (years) by gender

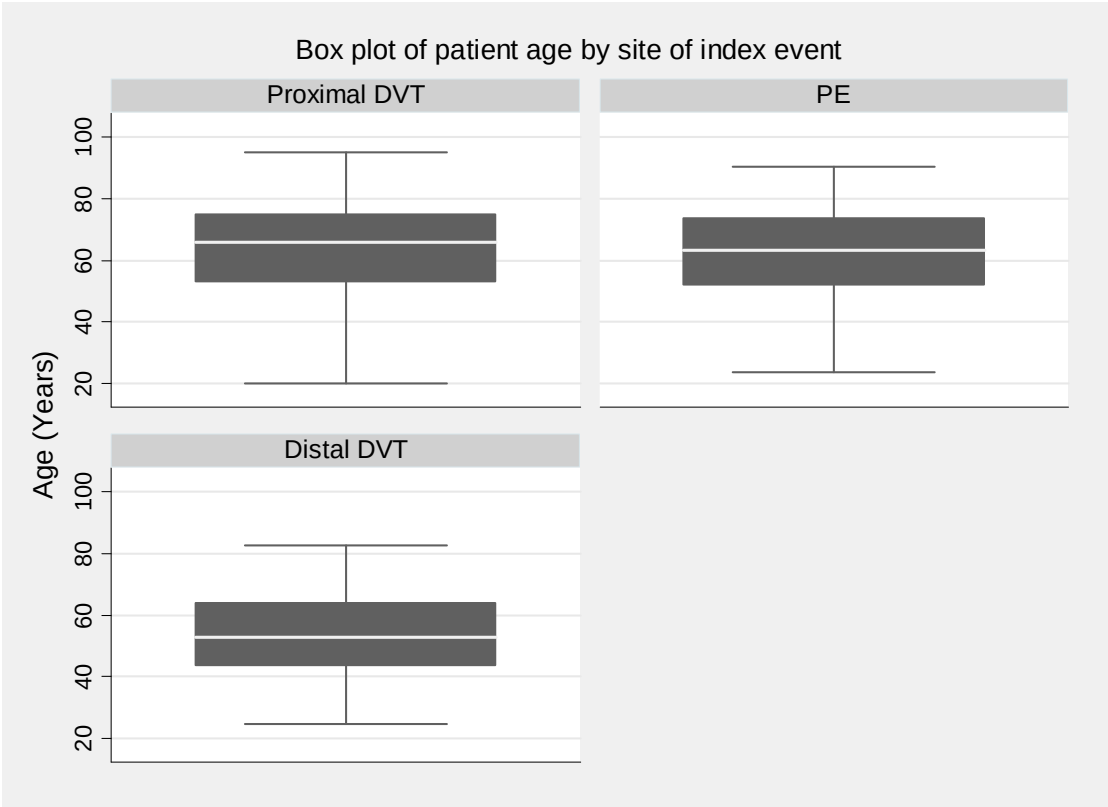


Figure 0.19 - Box plots of patient age (years) by site of index event

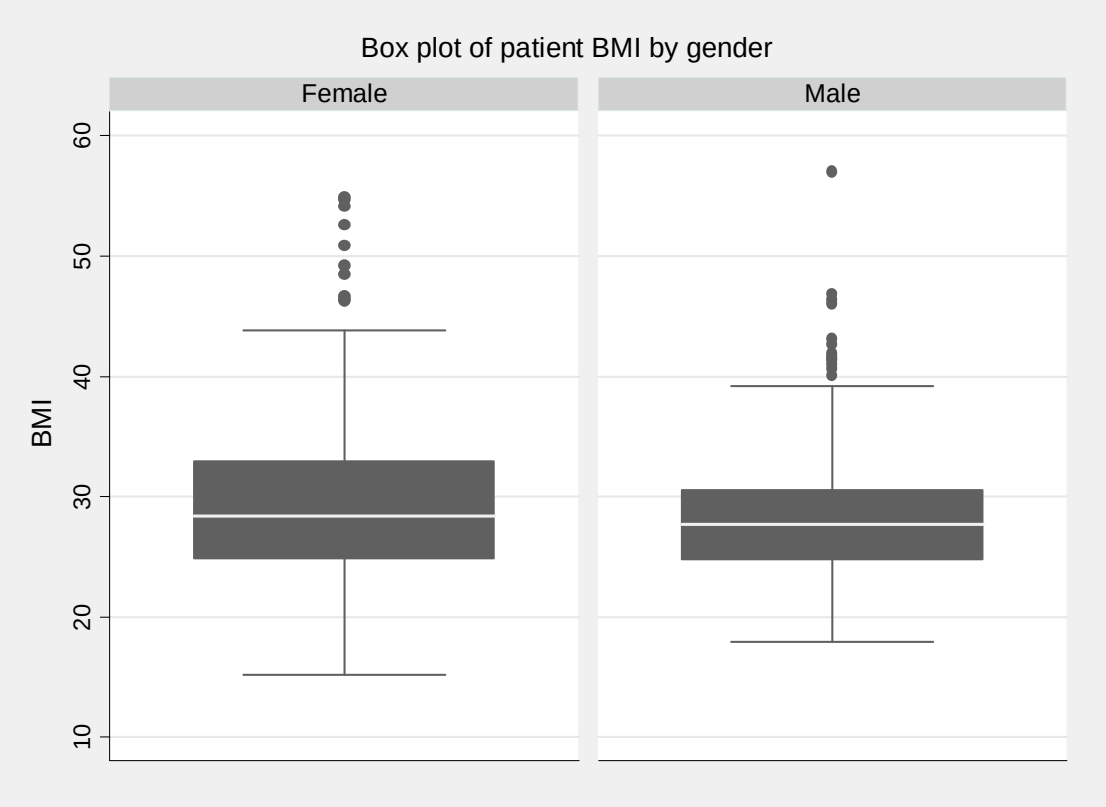


Figure 0.20 - Box plots of patients BMI by gender

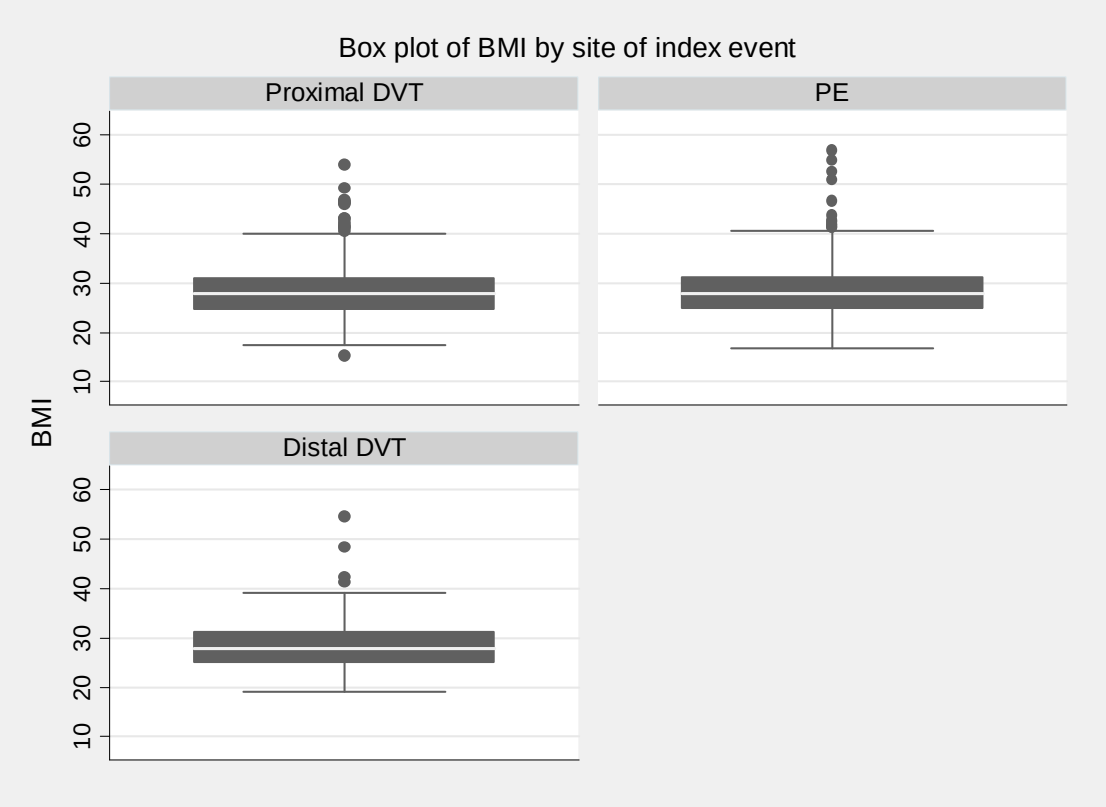


Figure 0.21 - Box plots of patients BMI by site of index event

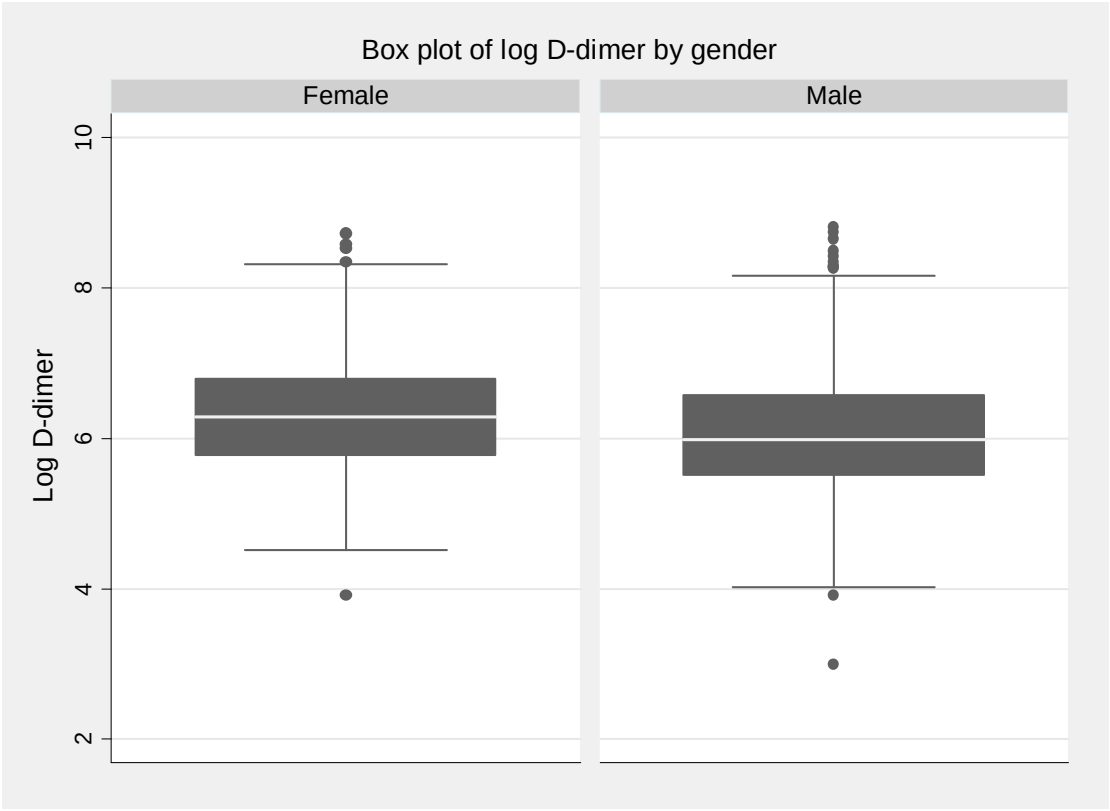


Figure 0.22 - Box plots of patients Log D-dimer score (ng/mL) by gender

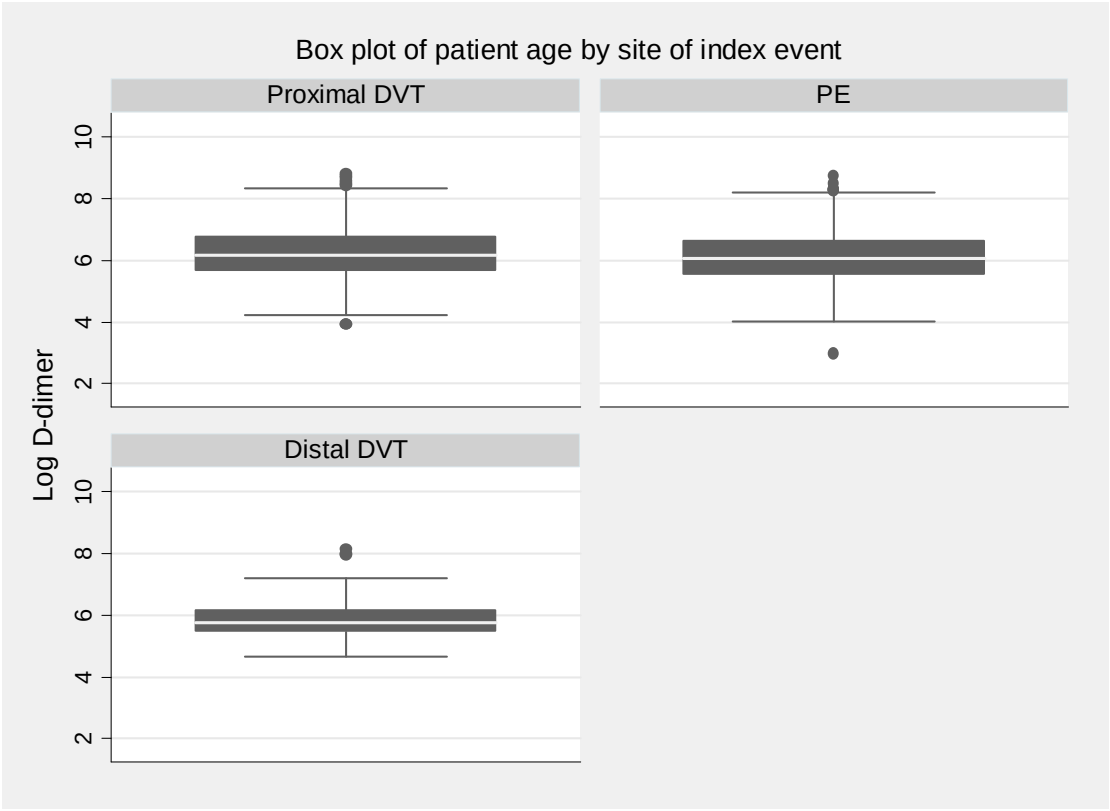


Figure 0.23 - Box plots of patients Log D-dimer score (ng/mL) by site of index event

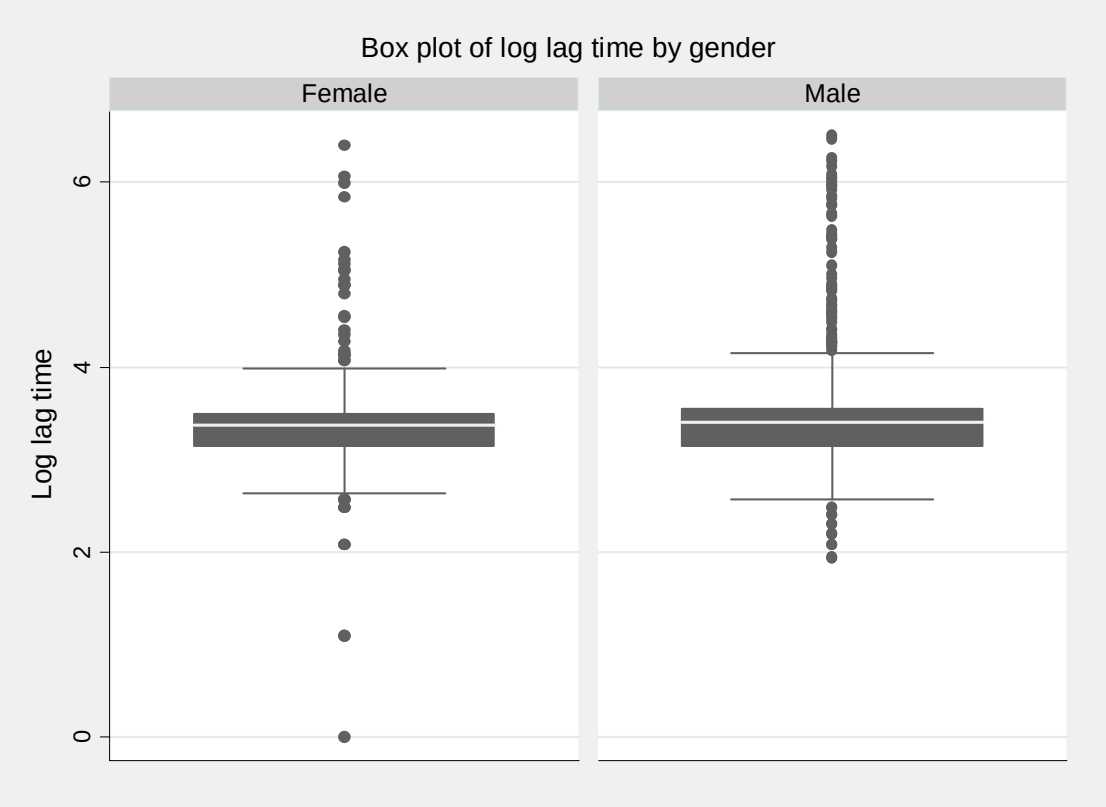


Figure 0.24 - Box plots of patient Log lag time (days) by gender

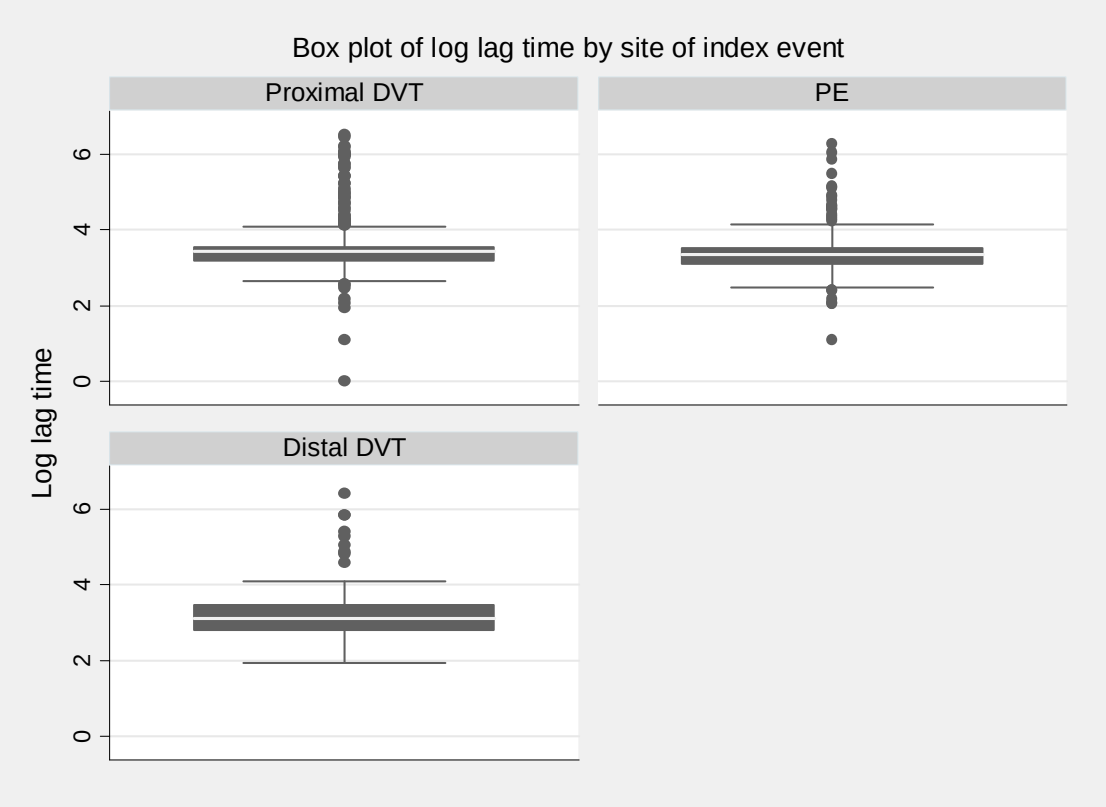


Figure 0.25 - Box plots of patient Log lag time (days) by site of index event

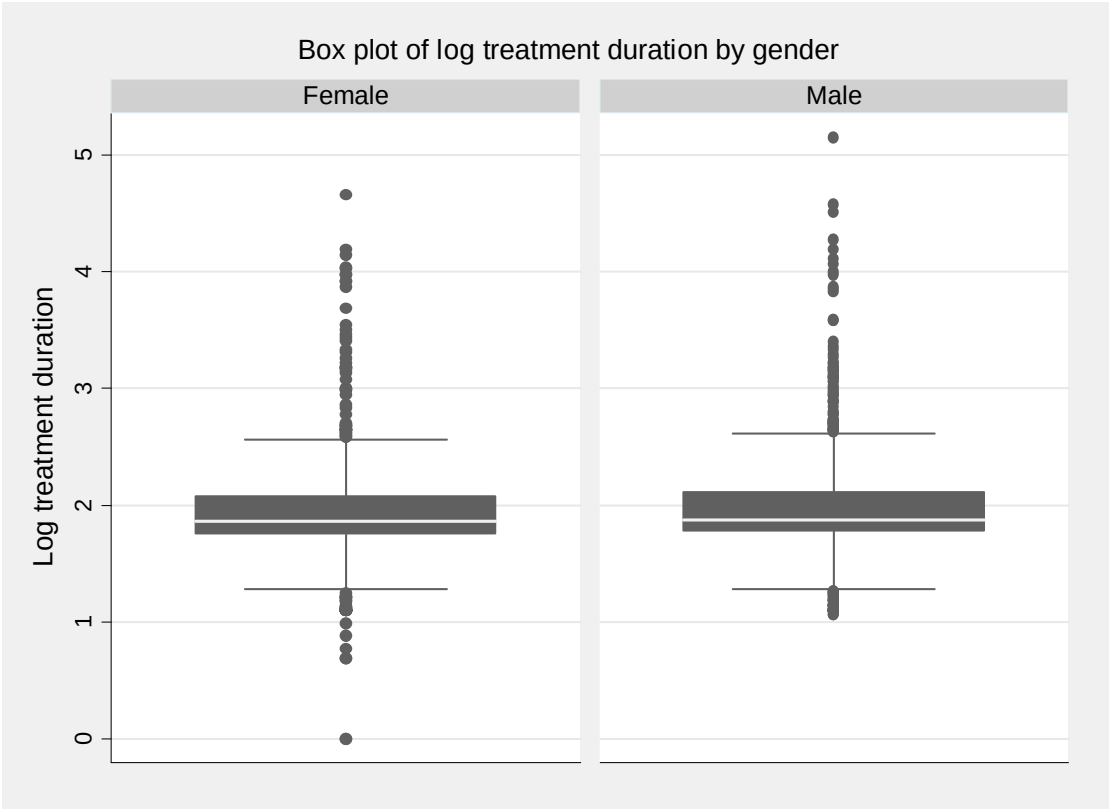


Figure 0.26 - Box plots of patient Log treatment duration (months) by gender

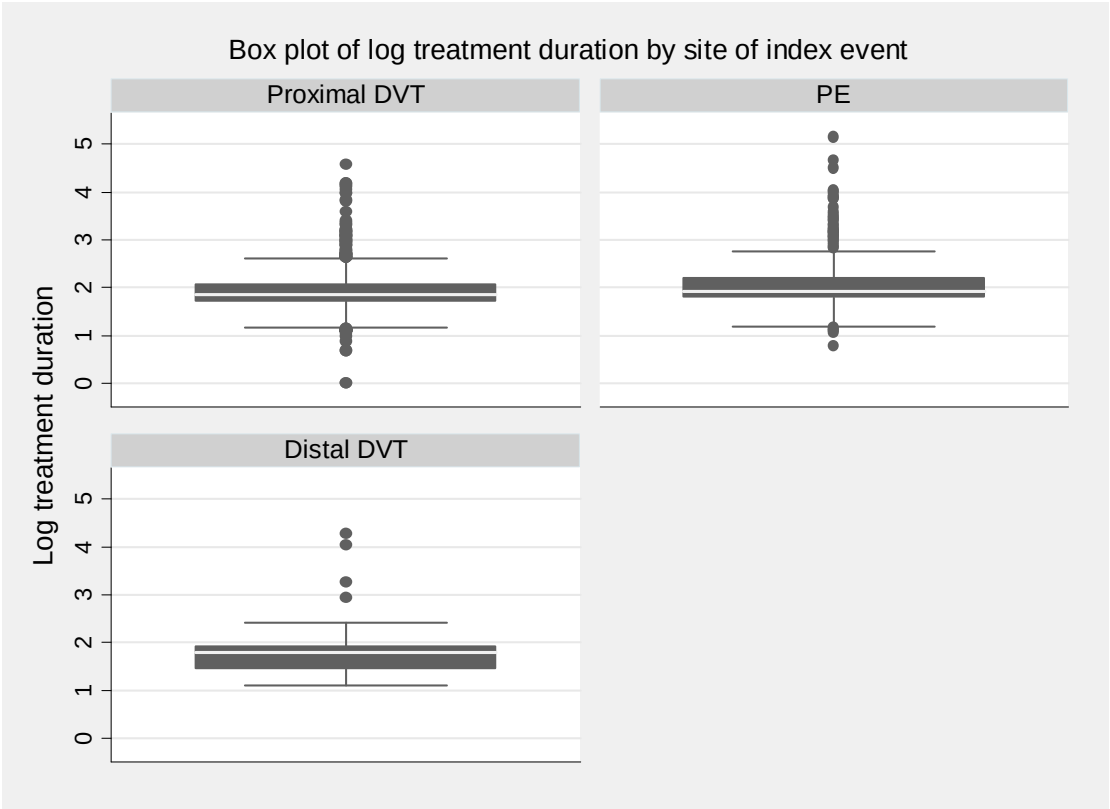


Figure 0.27 - Box plots of patient Log treatment duration (months) by site of index event

Figure 0.28 - Box plots of patient age x log D-dimer interaction by gender

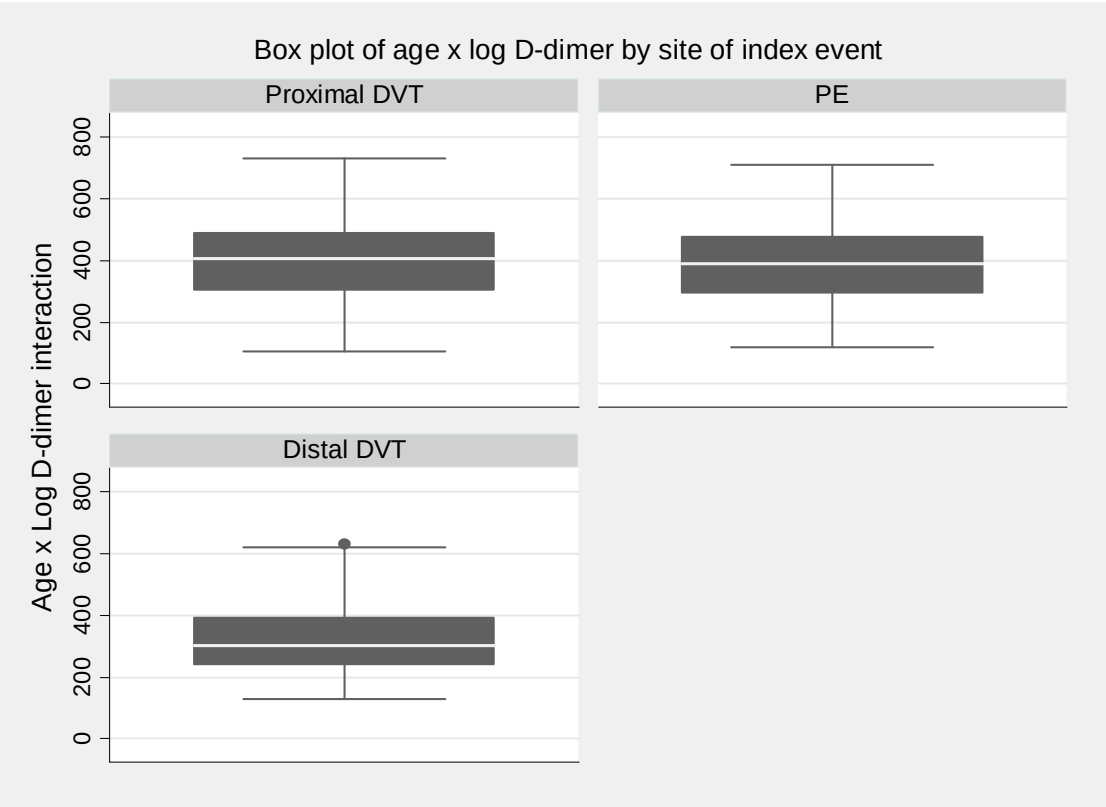


Figure 0.29 - Box plots of patient age x log D-dimer interaction by site of index event

APPENDIX B3: Sensitivity analysis results: Post D-dimer model

Interaction effects

Interaction effects quantify a differential effect of a predictor in a specific subgroup of the population. An interaction effect can be either an increased risk or decreased risk beyond that associated with a single characteristic. For example within the pre D-dimer model, both gender (being male) and site of index event (having a first PE) are associated with significant increases in recurrence rate; thus an interaction between gender and site of index event would imply that patients who are both male and have a PE are at increased risk beyond that associated with being male or having a PE alone.

As genuine interaction effects are rare and hard to identify, and because data dredging to identify interactions may find spurious results, the clinical team were asked for their guidance regarding which interaction terms are most important to examine. The clinical team suggested investigating an interaction of D-dimer and age, as it was felt plausible that the predictive effect of D-dimer value (a measure of general coagulability) may change as age increases.

To test for an interaction between age and D-dimer, the final Post D-dimer model was re-fitted including a term for the multiplication of age and D-dimer score. The interaction effect was shown to be insignificant at the 5% level, with a 95% confidence interval ranging from 0.98 to 1.01, and a P-value of 0.3 (*see Table 0.5*). Thus no interaction term was included in the final model.

Table 0.5 - Model specification including an Age x D-dimer interaction effect (The post D-dimer model)

Predictor	Beta coefficient (95% CI)	Hazard ratio (95% CI)	P-value
Age	0.026 (-0.048, 0.101)	1.03 (0.95, 1.11)	0.49
Gender			
Male	0.539 (0.185, 0.894)	1.71 (1.2, 2.44)	0.003
Site of index event			
Proximal DVT	1.633 (0.623, 2.643)	5.12 (1.86, 14.05)	0.002
PE	1.671 (0.651, 2.691)	5.32 (1.92, 14.74)	0.001
D-dimer (Log)	1.045 (0.309, 1.781)	2.84 (1.36, 5.94)	0.01
Lag time in days (Log)	-0.371 (-0.675, -0.067)	0.69 (0.51, 0.93)	0.02
Age x D-dimer interaction term	-0.006 (-0.018, 0.006)	0.99 (0.98, 1.01)	0.298

Further to this, an interaction effect between D-dimer levels and lag time was examined as the two predictors are inextricably linked; it is plausible that the prognostic importance of D-dimer levels varies over lag time (the time taken between cessation of therapy and the measurement of D-dimer levels). As previously, the final Post D-dimer model was re-fitted

including a term for the multiplication of D-dimer level and lag time. The interaction effect was shown to be insignificant at the 5% level, with a 95% confidence interval for the hazard ratio ranging from 0.79 to 1.57, and a P-value of 0.552 (*see Table 0.6*). Thus no interaction term for D-dimer and lag time was included in the final model.

Table 0.6 - Model specification including an D-dimer x Lag time interaction effect (The post D-dimer model)

Predictor	Beta coefficient (95% CI)	Hazard ratio (95% CI)	P-value
Age	-0.012 (-0.024, -0.001)	0.988 (0.976, 0.999)	0.037
Gender			
Male	0.55 (0.2, 0.91)	1.74 (1.22, 2.48)	0.002
Site of index event			
Proximal DVT	1.65 (0.64, 2.66)	5.19 (1.89, 14.24)	0.001
PE	1.68 (0.66, 2.7)	5.38 (1.94, 14.92)	0.001
D-dimer (Log)	0.31 (-0.86, 1.49)	1.37 (0.42, 4.45)	0.601
Lag time in days (Log)	-1.02 (-3.23, 1.18)	0.36 (0.04, 3.27)	0.364
D-dimer x Lag time interaction term	0.11 (-0.24, 0.45)	1.11 (0.79, 1.57)	0.552

Time-dependent effects

Allowing for time-dependent predictor effects might improve the performance of the model if it better fits the underlying data. Non-proportional hazards can be a sign of a time-dependent effect and as such including time-dependent effects can account for departures from the proportional hazards assumption. The validity of the proportional hazards assumption was assessed for predictors above (*see section 0*), and the assumption was considered appropriate for all predictors. It was therefore not expected that any time-dependent effects would be found to significantly improve the performance of either final model.

To further check this, a procedure proposed by Royston and Lambert (46) was used to identify potential time-dependent effects. The procedure first identifies the p-value associated with including each predictor in the model as a time-dependent effect using a likelihood-ratio test. A time-dependent effect is included for the predictor with the smallest p-value, providing the p-value is less than a pre-defined alpha significance level. The process is repeated until no time-dependent effects are significant at the chosen alpha level.

A 1% significance level was selected to test for time-dependent effects so as to account for multiple testing. The baseline spline function for the post D-dimer model used 3d.f. (*see section 0*) and therefore 3d.f. were used for the time-dependent effects to allow more flexibility. After one cycle of the procedure no predictors in the post D-dimer model were found to be significantly time-dependent at the 1% level, though log D-dimer was close to significance with a p-value from the likelihood-ratio test of 0.02 (*see Table 0.7*). Given the

lack of formal significance, and the aim for a more parsimonious model, the time-dependent effect was excluded.

Table 0.7 – First cycle of stepwise forward selection of time-dependent effects (The post D-dimer model)

Predictors	Deviance difference	P-value vs. Null
Age	2.49	0.477
Gender (Male)	4.491	0.213
Site of index event (Proximal DVT)	0.658	0.883
Site of index event (PE)	2.495	0.476
Log D-dimer	9.68	0.022
Log lag time	3.98	0.264

Multiple imputation of missing data

As the RVTE dataset used for model development included some missing data for some of the potential predictors, a complete case analysis was perform for model development, excluding any patient with missing data from the analysis. Sensitivity analysis was performed using multiple imputation, to evaluate how model estimates compared to those from the complete case analysis.

Of the included predictors only D-dimer and lag time had missing values, with 15% and 11.4% incomplete data respectively, across the whole dataset of six trials (*see Table 0.3*). It was therefore possible to consider multiple imputation for these two factors within the post D-dimer model. As the RVTE dataset consisted of multiple trial populations for development of the post D-dimer model it was important to account for this clustering when imputing missing observations; imputation across trials can lead to bias where the association between factors differs by trial (211). As such missing data for lag time which was 100% incomplete within the Baglin trial (198) could not be imputed (*see Table 0.3*), and so the same set of six trials (excluding Baglin) were used as in the complete data analysis for the post D-dimer model.

Imputation models were selected to include all included predictors from the final post D-dimer model as well as predictors for the observed recurrences (event indicator) and the baseline hazard to account for the time-to-event outcome (57); imputation was performed within trial populations. The largest proportion of incomplete data observed within individual trial populations was 48.1% missing D-dimer observations within the Poli trial (214), therefore 50 imputed datasets were created to provide the greatest reproducibility (57). Imputation was performed for ten cycles within each of the 50 imputed datasets to stabilise the results.

Box plots were used to check that the distributions of the observed and imputed data broadly matched, large differences could indicate an inappropriate imputation model (57).

On inspection the imputed distributions for both D-dimer and lag time (*see Figure 0.30 and Figure 0.31*) appeared to be very similar to the corresponding observed distributions (indicated as zero on the box plots). Therefore the imputation process appeared to be appropriate.

The model including all predictors identified as important in the complete case analysis (age, site of index event, gender, D-dimer and lag time) was fitted to the 50 imputed datasets and the coefficients of each were combined using Rubins rules (56) (*see Table 0.8*). In comparison with the specification of the post D-dimer model under a complete case analysis (*see Table 0.8*) the estimated hazard ratios after imputation were reasonably similar. The effect of each factor within the model did not have a dramatically different interpretation between the complete case and multiple imputation models. In particular the effects of D-dimer and lag time are relatively unchanged with hazard ratios of 1.93 and 0.74 compared to 2.01 and 0.75 for the complete case model. In general the 95% confidence intervals were similar with the exception of site of index event where the multiple imputation model estimated slightly smaller 95% confidence intervals, showing greater precision, likely due to the increased number of observations. The effect of age and lag time were borderline significantly different from null in the complete case model, but appeared to be significant in the multiple imputation model. This adds further weight to the inclusion of age and lag time factors in the prognostic model.

The inclusion of treatment duration as a predictor in the multiple imputation model was investigated given the increased complete patient data; treatment duration did not reach significance within the imputation model with a hazard ratio of 1.07 (95% CI: 0.85, 1.35), and p-value of 0.574 providing confirmatory evidence toward the exclusion of treatment duration within the complete case post D-dimer model.

Table 0.8 - The post D-dimer model specification following imputation of missing variable data. P=P-value.

Predictor	Imputation model		Original model		FMI*
	Beta coefficient (95% CI)	P	Beta coefficient (CI)	P	
Age	-0.017 (-0.026, -0.007)	0.001	-0.0105 (-0.022, 0.0011)	0.075	0.032
Gender					
Male	0.666 (0.352, 0.98)	<0.001	0.55 (0.19, 0.89)	0.002	0.007
Site of index event					
Proximal DVT	1.662 (0.627, 2.697)	0.001	1.74 (0.67, 2.79)	0.001	0.001
PE	1.639 (0.596, 2.683)	0.002	1.76 (0.68, 2.83)	0.001	0.002
Log D-dimer	0.657 (0.485, 0.829)	<0.001	0.7 (0.51, 0.89)	<0.001	0.169
Log lag time	-0.298 (-0.572, -0.025)	0.044	-0.29 (-0.58, 0.002)	0.051	0.167

** FMI = Fraction of Missing Information = B /(W + B); where B is the between-imputations variance, and W is the within imputation variance*

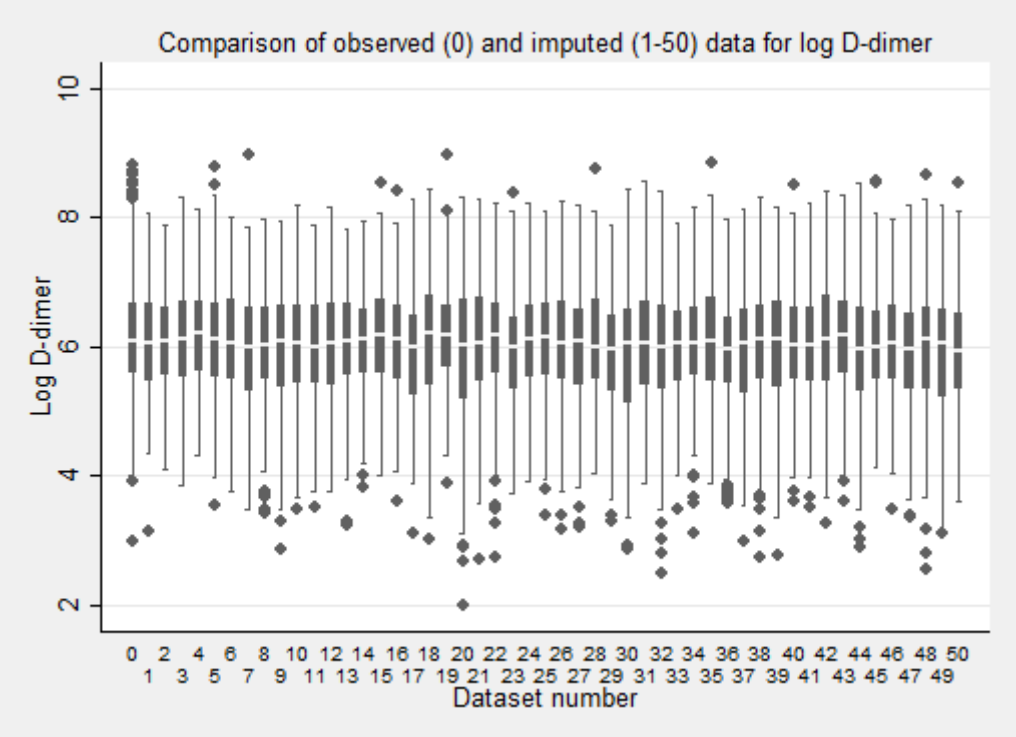


Figure 0.30 - Comparison of observed and imputed data for log D-dimer (The post D-dimer model)

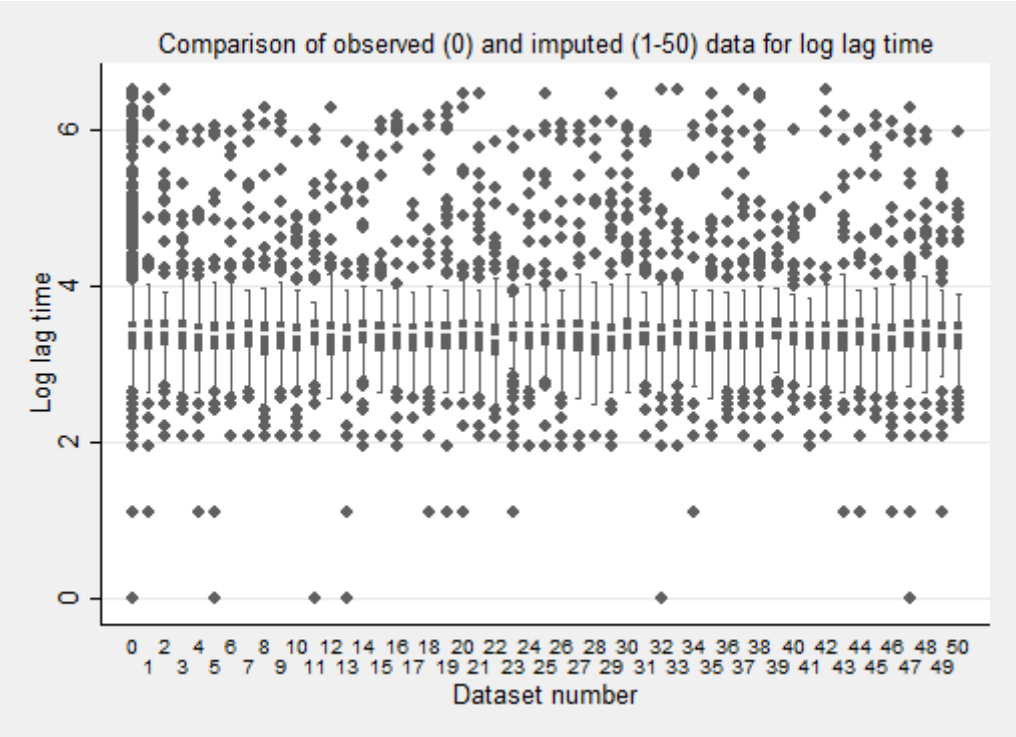


Figure 0.31 - Comparison of observed and imputed data for log lag time (The post D-dimer model)

Regarding the reproducibility of the multiple imputation results to check that similar conclusions could be drawn from an identical imputation approach, the MC error of each estimated hazard ratio and the corresponding standard errors was checked. The MC error was measured as a percentage of the standard error of the estimated hazard ratio, where an MC error lower than 10% of the standard error was considered appropriate. The MC errors observed from the imputation procedure used were all lower than 10%, with the greatest being 5.74% for D-dimer, meaning that it is highly likely that the results of multiple imputation procedure would lead to consistent conclusions across the imputed datasets (*see Table 0.9*).

Table 0.9 - Monte Carlo error acceptability for analysis based on 50 imputed datasets

Predictor	Hazard ratio	St. error	P-value	Lower 95% CI	Upper 95% CI	FMI*
Age	0.98	0.005	0.001	0.97	0.99	0.03
MC error	0.00	6.90E-07	0.000	1.26E-04	1.21E-04	0.01
% of s.e.	2.50%	0.01%				
Gender (Male)	1.95	0.302	0.000	1.43	2.64	0.01
MC error	0.00	0.0003	0.000	0.003	0.005	0.00
% of s.e.	1.16%	0.12%				
Site (Proximal DVT)	5.27	2.694	0.001	1.93	14.37	0.00
MC error	0.01	0.003	0.000	0.01	0.04	0.00
% of s.e.	0.54%	0.12%				
Site (PE)	5.15	2.653	0.002	1.87	14.15	0.00
MC error	0.01	0.003	0.000	0.01	0.04	0.00
% of s.e.	0.56%	0.13%				
Log D-dimer	1.93	0.180	0.000	1.61	2.32	0.17
MC error	0.01	0.001	0.000	0.01	0.02	0.03
% of s.e.	5.74%	0.78%				
Log lag time	0.74	0.110	0.044	0.56	0.99	0.17
MC error	0.01	0.002	0.006	0.006	0.009	0.03
% of s.e.	5.70%	1.37%				

** FMI = Fraction of Missing Information = B /(W + B); where B is the between-imputations variance, and W is the within imputation variance*

APPENDIX B4: Model checking results: Post D-dimer model

Proportional hazards assumption

A scatter plot of the scaled Schoenfeld residuals against log time, with a lowess smoother, was used to check the proportional hazards assumption for factors in the post D-dimer model as described previously (*see section 3.2.6*). Plots for log D-dimer (*see Figure 0.32*) and log lag time (*see Figure 0.33*) show that the proportional hazards assumption is valid for the

post D-dimer model; the lowess smoothed line roughly follows the reference line for each covariates log hazard ratio, indicating proportionality. Similar plots testing the proportional hazards assumption were inspected for the remaining covariates in the post D-dimer model, the proportional hazards assumption was valid for all predictors.

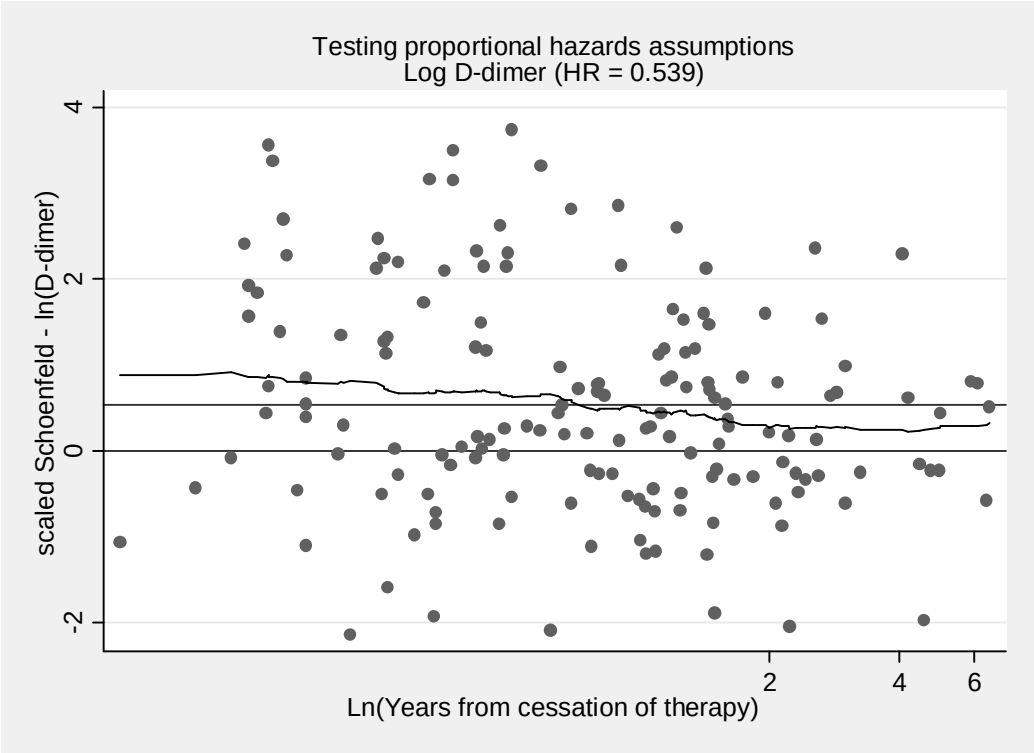


Figure 0.32 - Scaled Schoenfeld residuals vs. Log time from cessation of therapy for log D-dimer

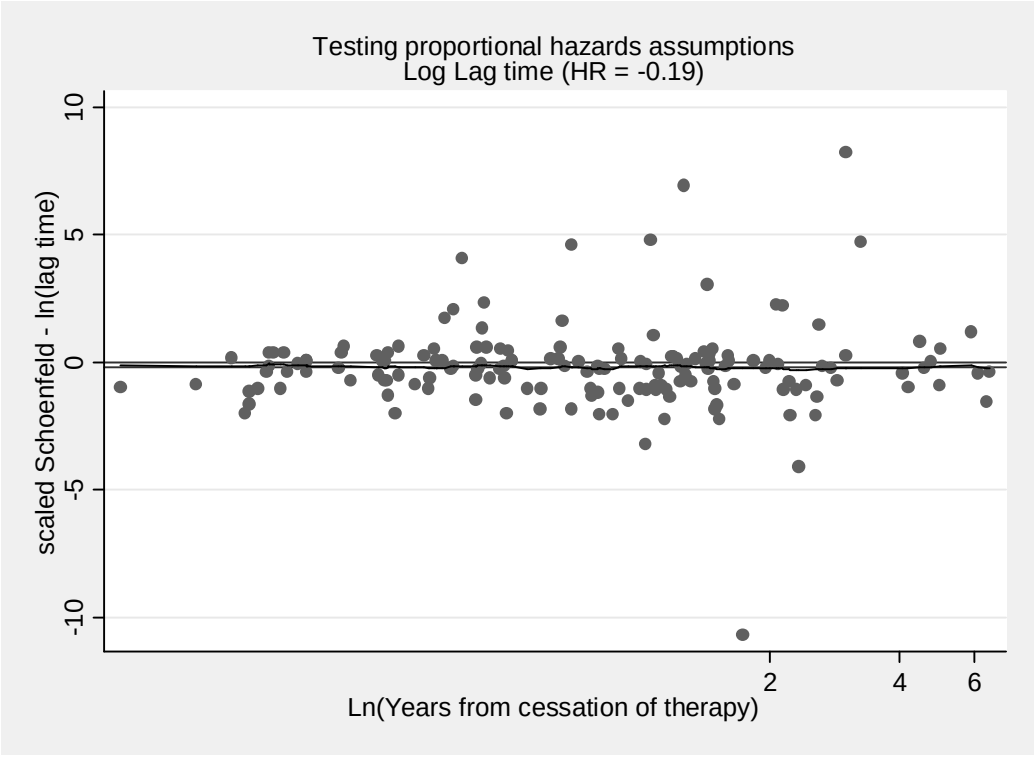
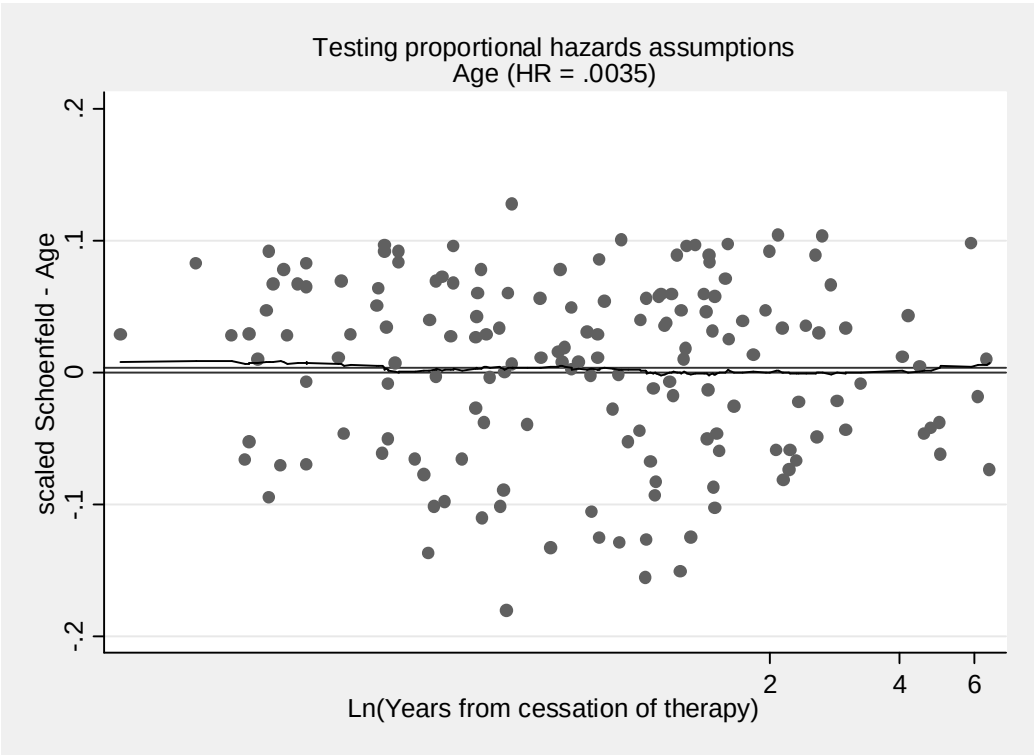
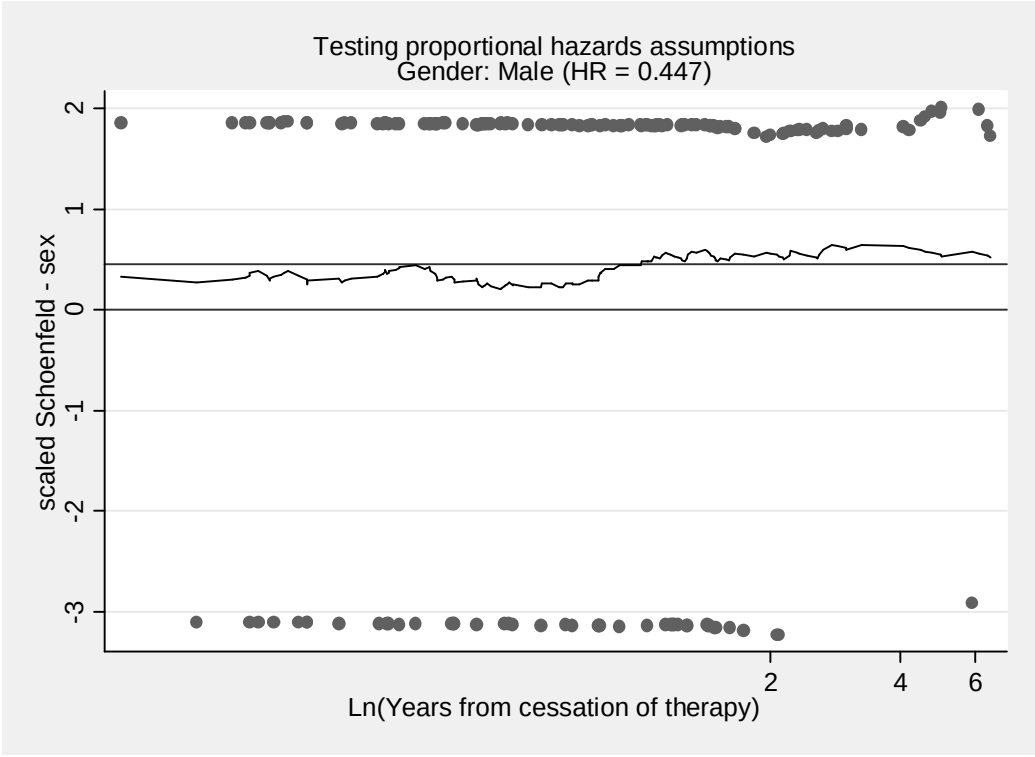
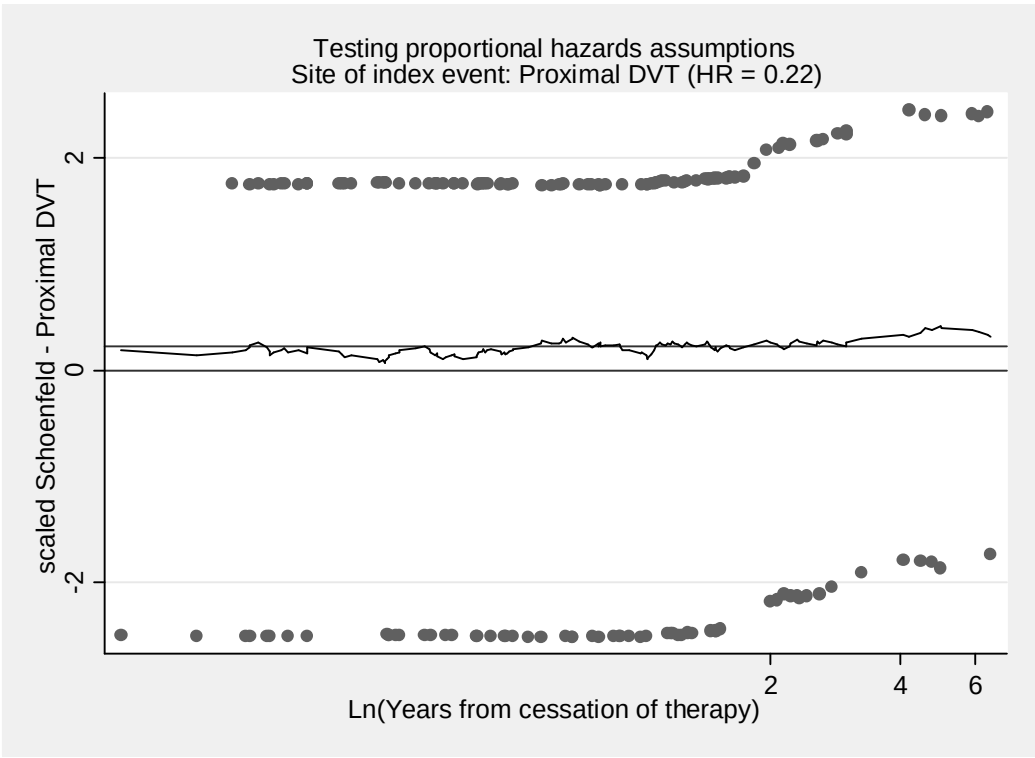
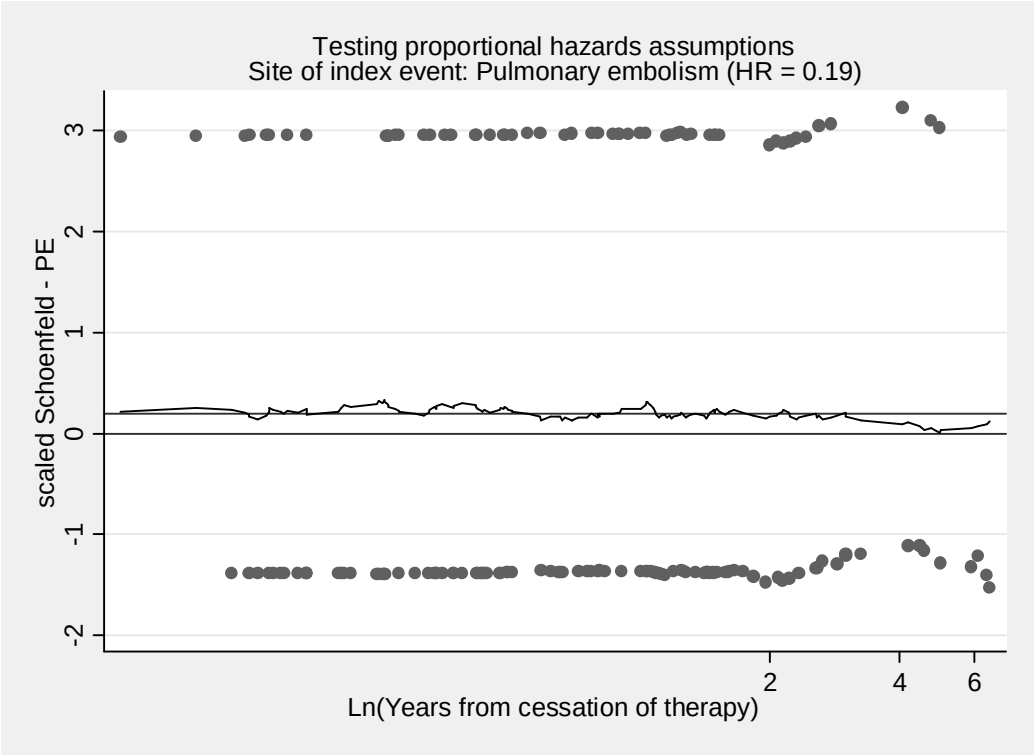


Figure 0.33 - Scaled Schoenfeld residuals vs. Log time from cessation of therapy for log lag time




Testing proportional hazards assumptions
Gender: Male (HR = 0.447)
scaled Schoenfeld - sex
2
1
0
-1
-2
-3
2
4
6
Ln(Years from cessation of therapy)


Testing proportional hazards assumptions
Site of index event: Proximal DVT (HR = 0.22)
scaled Schoenfeld - Proximal DVT
2
0
-2
2
4
6
Ln(Years from cessation of therapy)



Functional form

To check that continuous predictors were included in the model with appropriate functional form, scatter plots of Martingale residuals against the predictors with a lowess smoother applied were inspected. Patient age, log D-dimer and lag time were the only continuous predictors included in the post D-dimer model and the functional form of these covariates was checked using Martingale residuals. Figure 0.34 and Figure 0.35 show a lowess smoother applied to a scatter of martingale residuals against log D-dimer and log lag time, respectively. In both cases the smoother appears to follow a linear trend over the covariate values, indicating a linearity assumption (on the log scale) for both factors was appropriate.

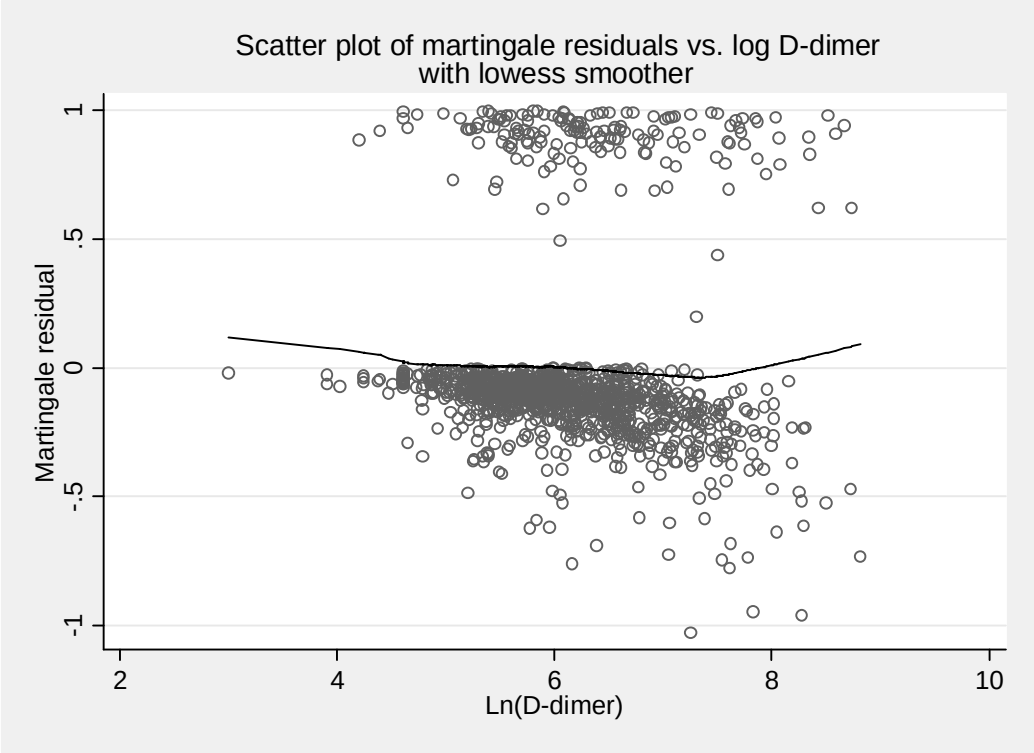


Figure 0.34 - Scatter plot of martingale residuals against log D-dimer (The post D-dimer model)

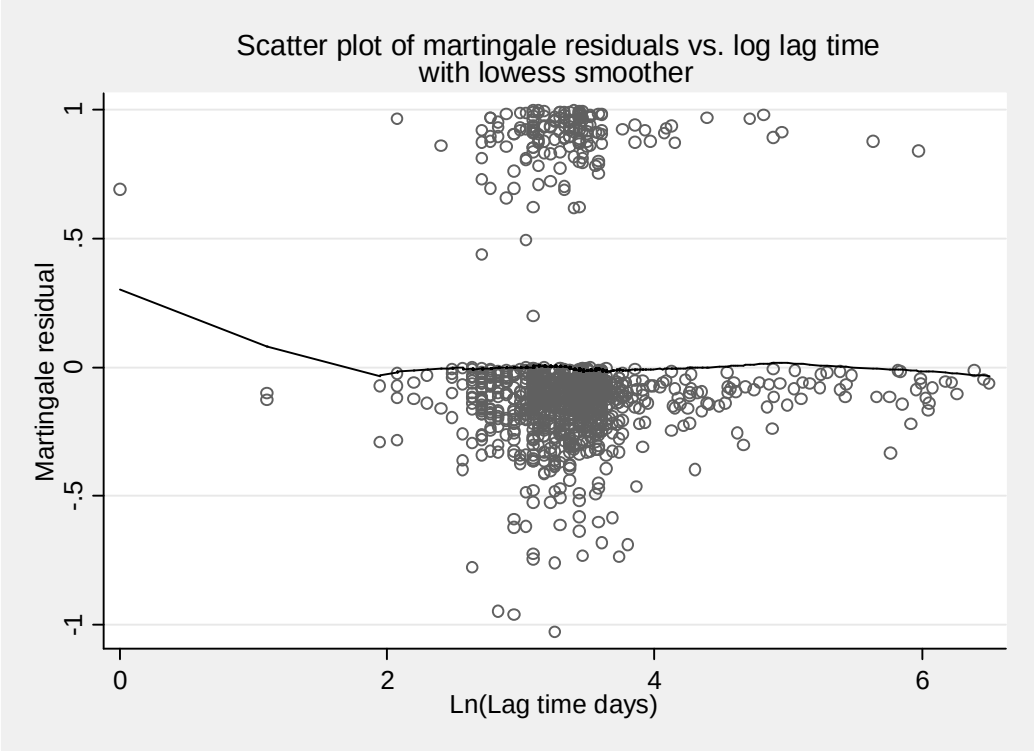
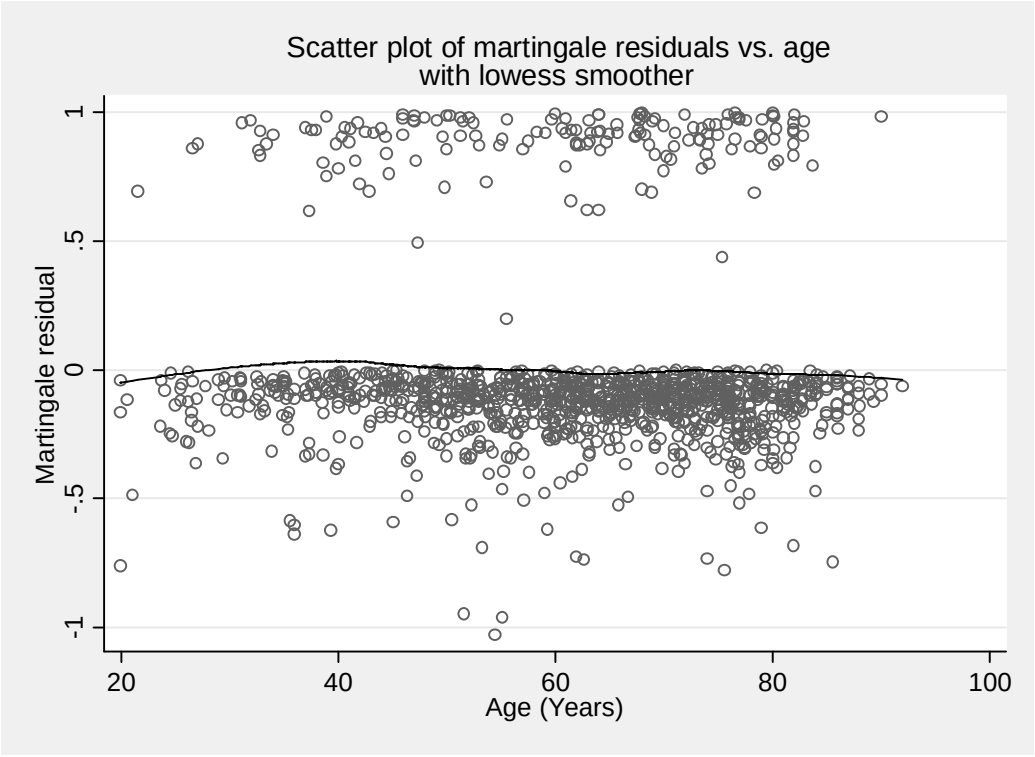


Figure 0.35 - Scatter plot of martingale residuals against log lag time (The post D-dimer model)



Outliers

As seen for the pre D-dimer model, plots of the deviance residuals against a patient indicator (*see Figure 0.36*) and against time (*see Figure 0.37*) indicate some outlying individuals. Figure 0.36 illustrates a scatter of the deviance residuals for the post D-dimer model, they clearly do not follow a normal distribution and this may again be due to heavy censoring in the dataset, a small number of individuals fall above the 1.96 critical Z value.

A plot of the deviance residuals against years from cessation of therapy allows investigation of any trend in the deviance residuals. In Figure 0.37 for the post D-dimer model there is again a trend in the deviance residuals over time based on the cumulative hazard at the event time (or censoring time). The deviance residuals which lie in the top left of the plot are, as for the pre D-dimer model, likely to be those individuals who had a recurrence early and therefore did not accumulate much hazard.

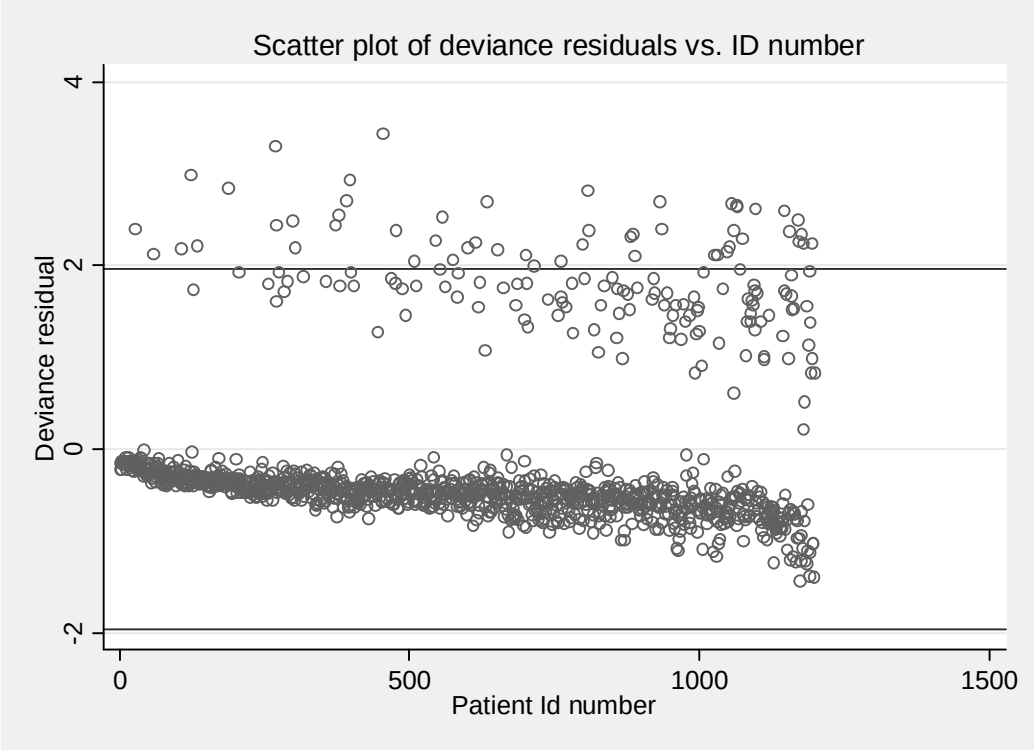


Figure 0.36 - Scatter plot of deviance residuals vs. patient ID (The post D-dimer model)

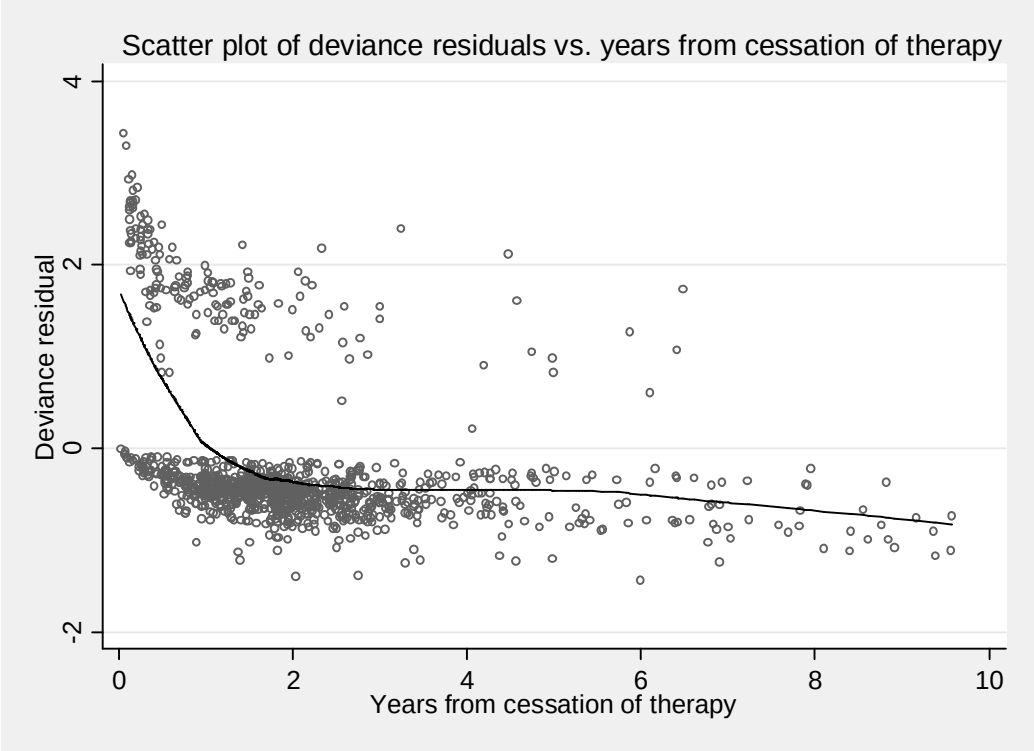


Figure 0.37 - Scatter plot of deviance residuals vs. years from cessation of therapy (The post D-dimer model)

Leverage

To check the influence of individuals on the parameter estimates, leverage can be assessed using delta-beta changes for each covariate as seen in the pre D-dimer model. Scatter plots of delta-betas for log D-dimer (*see Figure 0.38*) and log lag time (*see Figure 0.39*) show that even individuals with the greatest leverage on these parameter estimates, have very small effects on the log hazard ratio as seen for the pre D-dimer model. Similar, small delta-beta changes were observed for the other covariates included in the post D-dimer model.

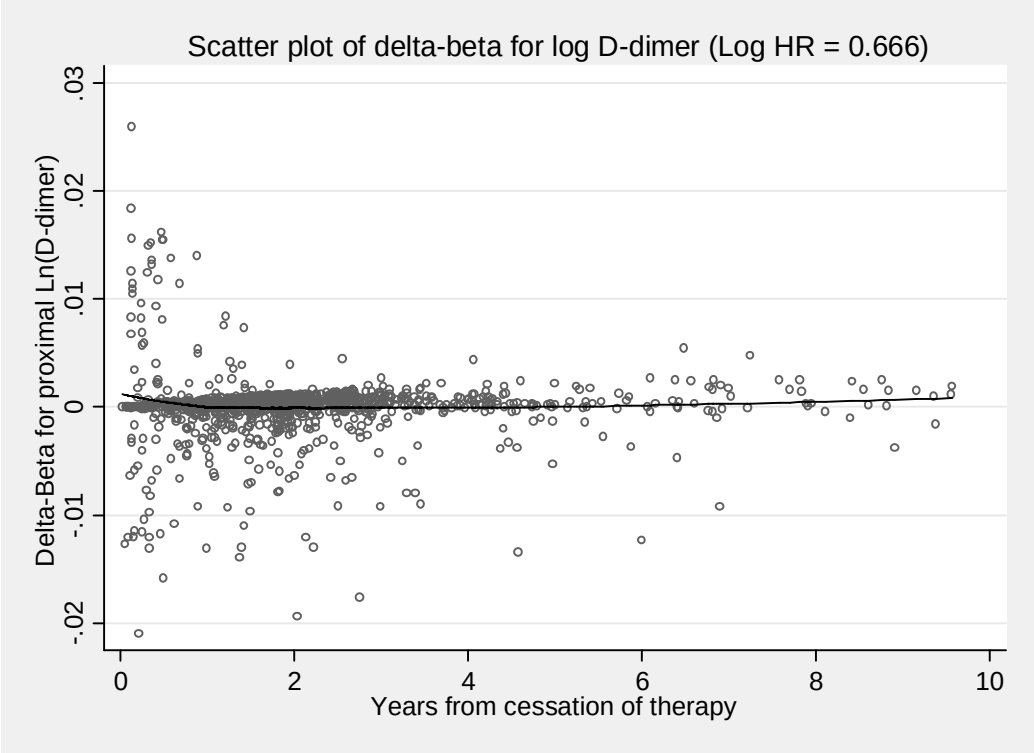


Figure 0.38 - Scatter plot of Delta-Beta for log D-dimer vs. years from cessation of therapy

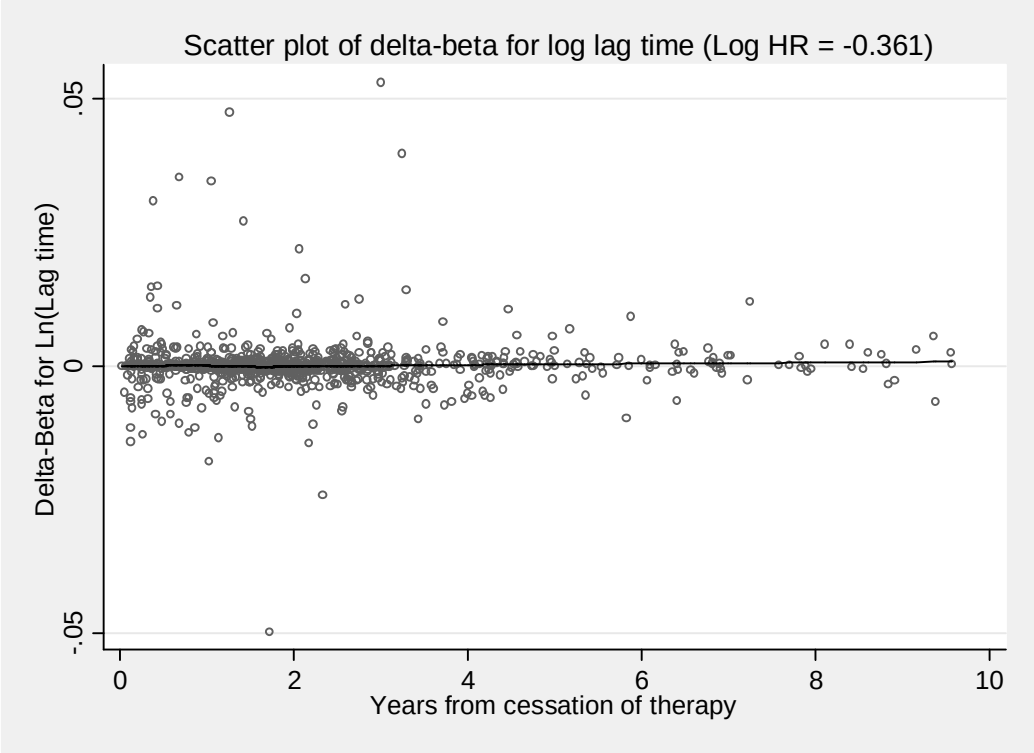
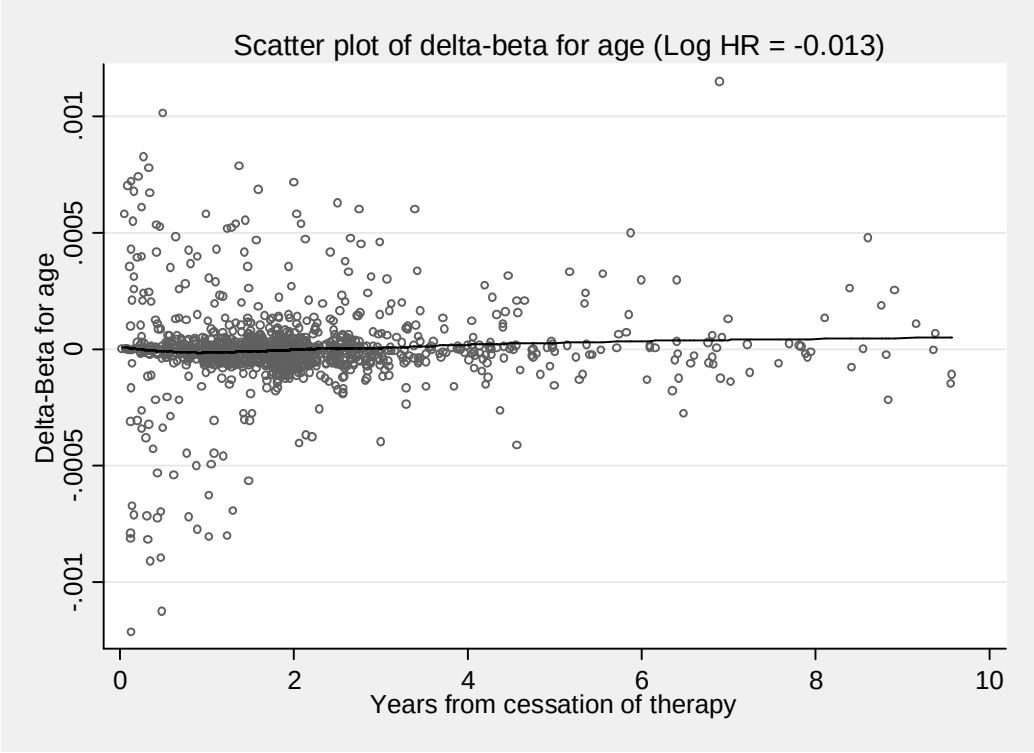
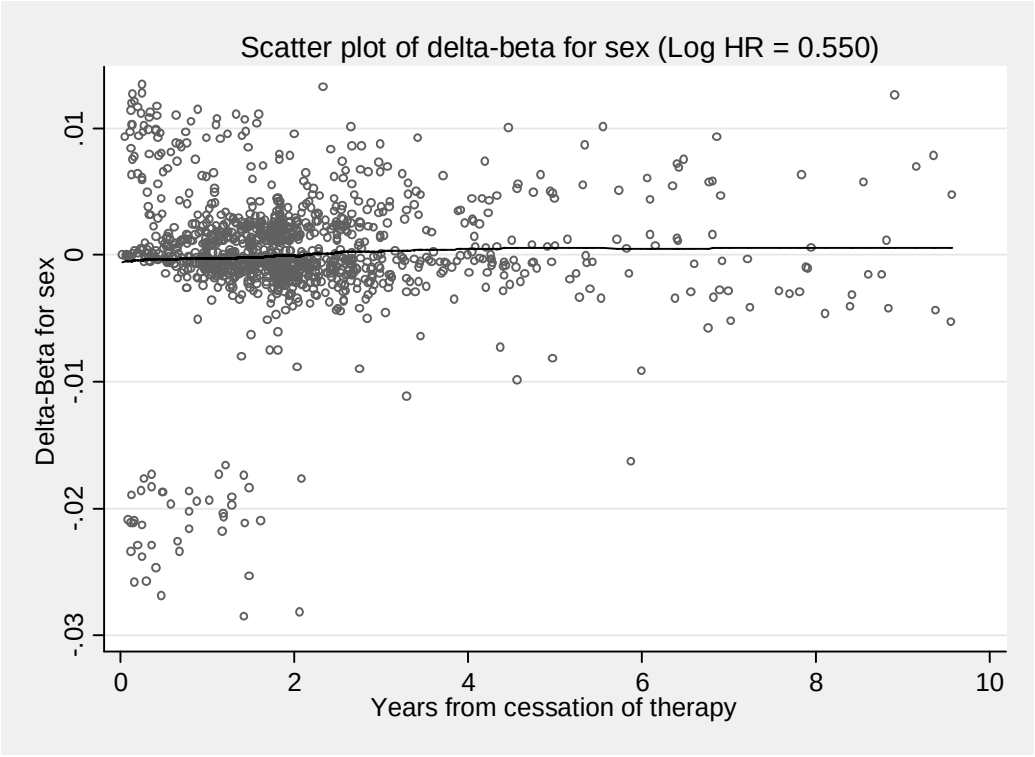
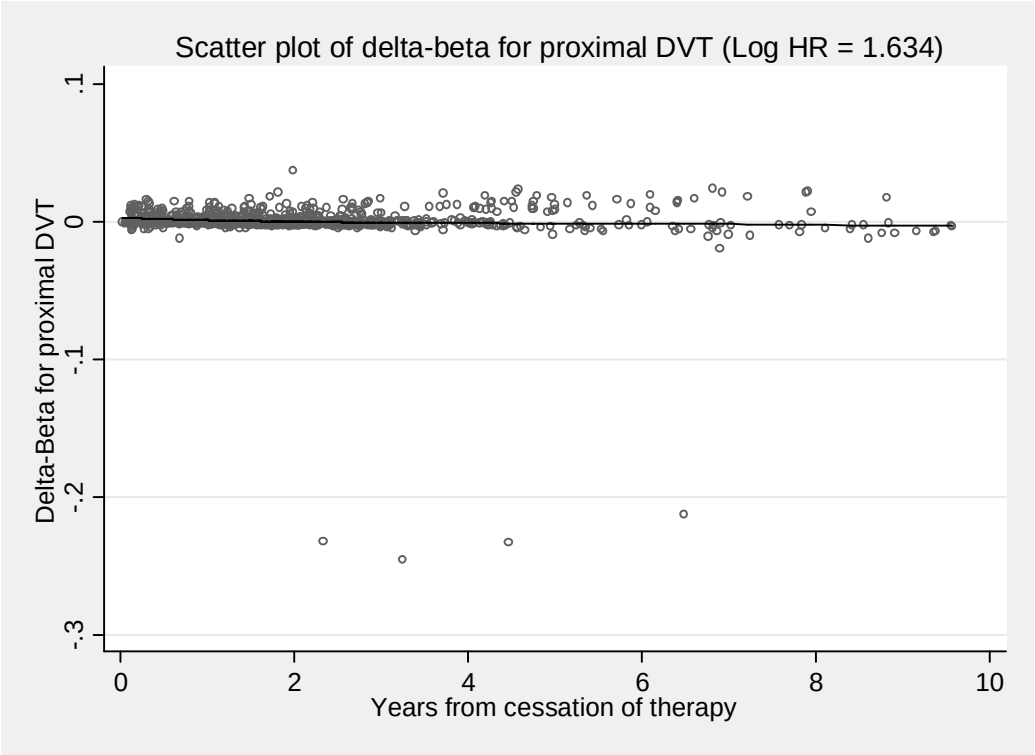
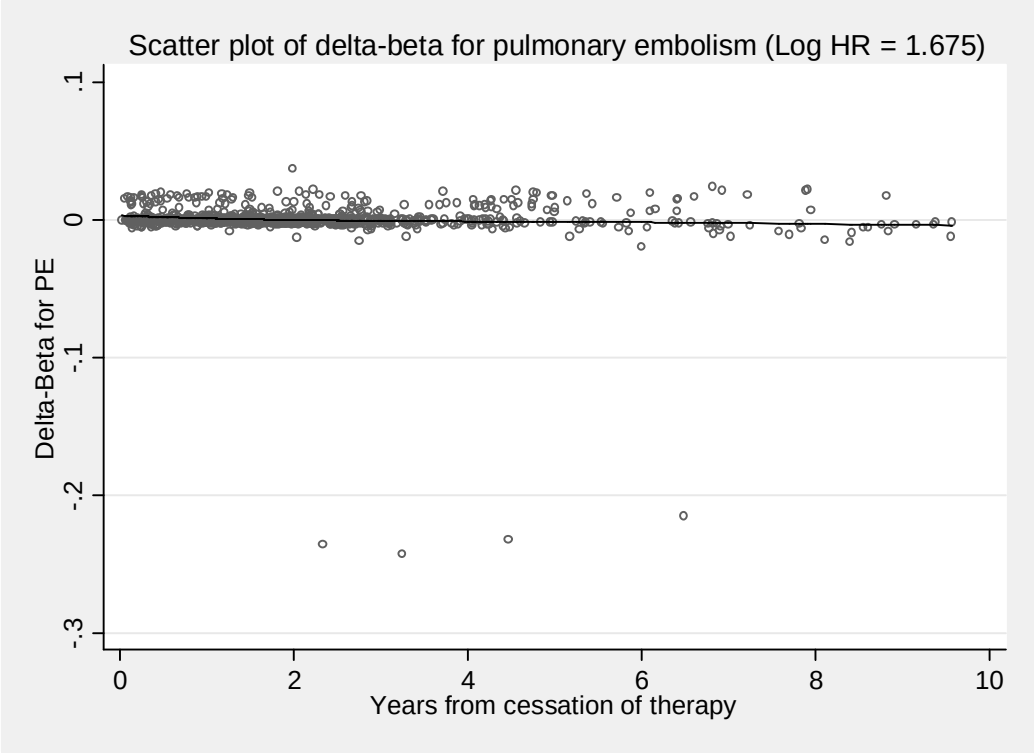


Figure 0.39 - Scatter plot of Delta-Beta for log lag time vs. years from cessation of therapy




Scatter plot of delta-beta for sex (Log HR = 0.550)
Delta-Beta for sex
.01
0
-.01
-.02
-.03
0
2
4
6
8
10
Years from cessation of therapy


Scatter plot of delta-beta for proximal DVT (Log HR = 1.634)
Delta-Beta for proximal DVT
.1
0
-.1
-.2
-.3
0
2
4
6
8
10
Years from cessation of therapy



APPENDIX B5: Pre D-dimer model validation performance

Model validation

The final step of the IECV approach (*see section 3.2.7*) is to assess model performance within the validation trial, at each cycle of the IECV approach. As the validation trial was excluded from model development the performance of the model within this dataset can be deemed as external validation. Model performance is now assessed in terms of both discrimination and calibration (*see section 3.2.7*).

Discrimination and calibration results for each cycle of the IECV for the pre D-dimer model are presented in Table 0.10, under a random-effects assumption on the baseline hazard. C-statistic estimates for the developed model range from 0.47 in the Tait (218) trial, to 0.58 in both the Palareti 2006 (204) and Poli (214) trials. A random-effects meta-analysis of the C-statistics from each validated model (each cycle of the IECV) provides a pooled estimate of the performance across all developed models (*see Figure 0.40*). The pooled C-statistic of 0.56 (95% CI: 0.51, 0.6) represents the overall weighted average C-statistic from all validation trials, showing poor discriminatory ability of the models developed in the cycles of the IECV approach. However, as this is a weighted average of the performance within each validation trial, it is expected that the discrimination would average out to that of a model built using the whole dataset. In this case it is of more interest to examine the heterogeneity across the trials, and the 95% prediction interval (112). The prediction interval provided is a useful tool

for interpreting the potential range of performance of the new model in a new setting (where the model will be applied), by accounting for the uncertainty in the pooled estimate, the heterogeneity between trials and the between trial standard deviation (113). The interval suggests that the C-statistic for the model used in a new setting could vary anywhere between 0.49 and 0.62, which represents a potentially broad range of performance from awful discrimination to a higher but still quite poor level. The heterogeneity, or variability, across the trial populations appears to be minimal (I-squared statistic = 0), indicating that the discrimination of the model appears consistent in new populations, and that any variation is due to chance (113). However, as this zero heterogeneity is only an estimate, its uncertainty is propagated in the 95% prediction interval, which is why the prediction interval is so wide. The prediction interval for the pooled C-statistic is entirely below the estimated prediction interval for the post D-dimer model, indicating that the pre D-dimer model may be inferior in all other possible settings.

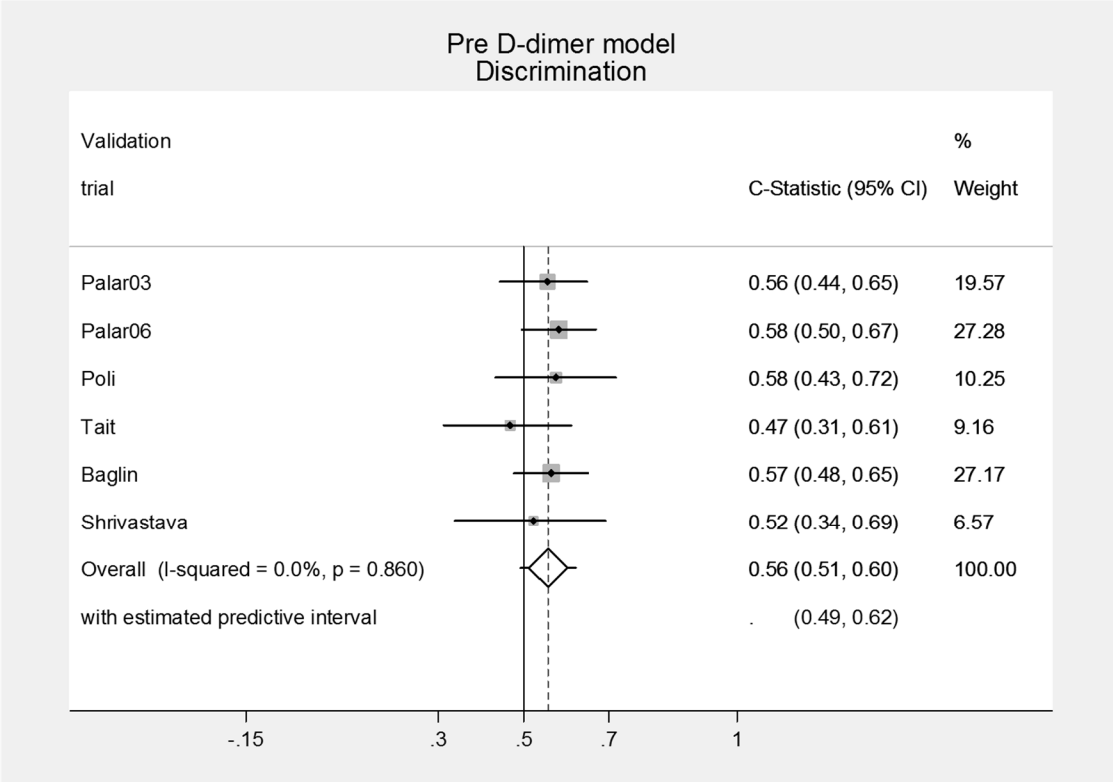


Figure 0.40 - Random-effects meta-analysis of C-statistic estimates obtained from each external validation of the Pre D-dimer models from the IECV cycle

Calibration is examined visually across all time-points in Figure 0.41. It is then quantified further in Table 0.10 at four time points: six months, one year, two years and three years after cessation of therapy. The model appears to be well calibrated (*see Table 0.10*), with expected minus observed, $S(t) - \hat{S}(t)$, probabilities with a recurrence very close to zero, and 95% confidence intervals including zero across all cycles of the IECV. Plots of the observed

probability of recurrence (based on the Kaplan-Meier survival estimates) compared to the expected probability of recurrence (based on the predictions of the model) are presented for each validation trial in Figure 0.41. A perfectly calibrated model would give a predicted curve very similar to the observed Kaplan-Meier curve, which can be seen for validation trials Palareti 2003 (217) and Poli (214). Within the remaining validation trials the developed model either over, or under predicted the probability of recurrence, compared to the observed probabilities within the validation trial. For example over prediction can be seen in the Palareti 2006 (204) trial beyond six months post cessation of therapy (*see Figure 0.41*). Plots of the S(t) – $\hat{S}$(t) statistic and 95% confidence intervals for each validation trial can be seen in Figure 0.42, showing the difference in proportion survived remains close to zero over time from cessation of therapy.

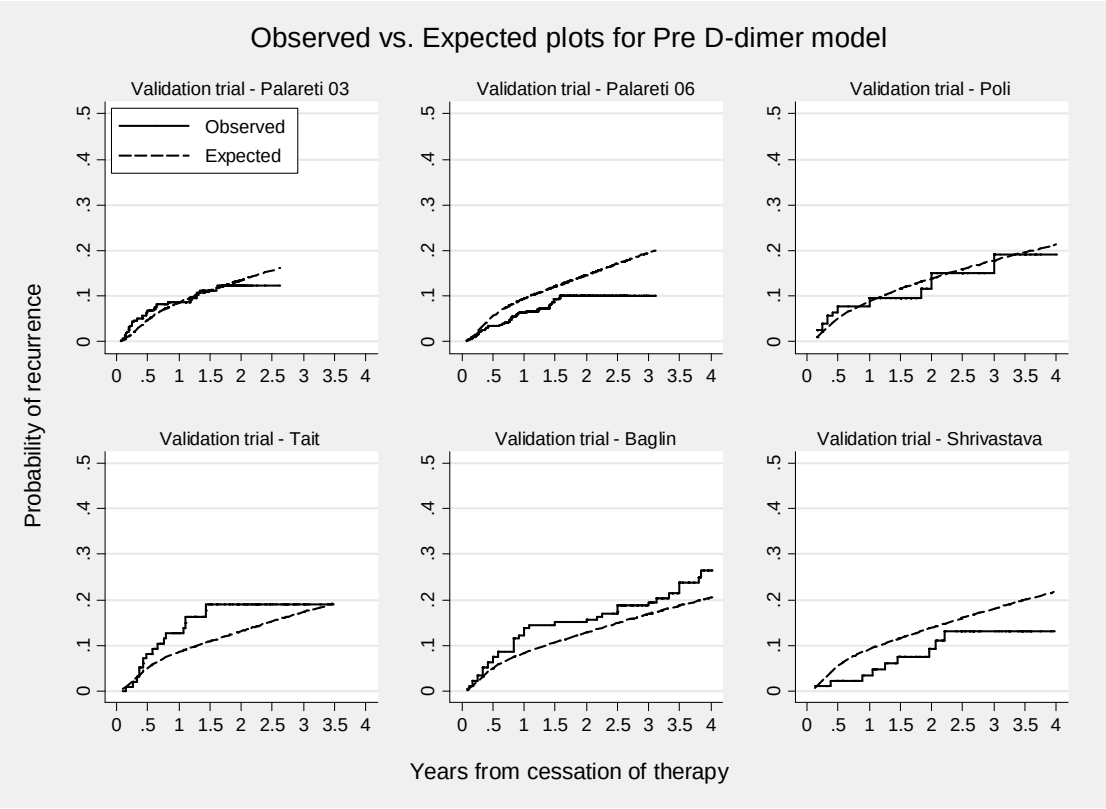


Figure 0.41 - Observed vs. Expected recurrence probabilities over time, obtained from each external validation of the Pre D-dimer models from the IECV cycle

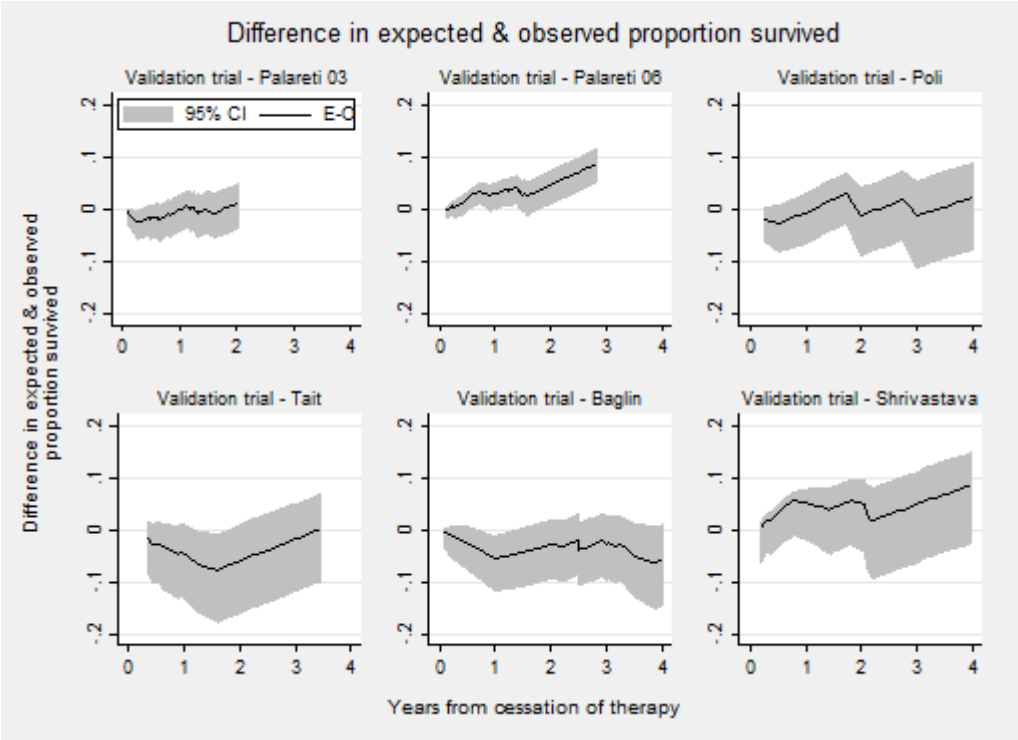


Figure 0.42 - Expected minus Observed probabilities with a recurrence for each validation trial for the pre D-dimer model

Table 0.10 - Summary statistics for discrimination and calibration of the pre D-dimer model

	External validation trial Estimate (95% CI)					
	Palareti 03	**Palareti 06**	**Poli**	**Tait**	**Baglin**	**Shrivastava**
Recurrences/Total patients	31/280	38/434	26/156	17/99	40/175	9/91
C-statistic	0.56 (0.44, 0.65)	0.58 (0.50, 0.67)	0.58 (0.43, 0.72)	0.47 (0.31, 0.61)	0.57 (0.48, 0.65)	0.52 (0.34, 0.69)
S(t) – Ŝ(t) statistic (6 months)	0.02 (-0.01, 0.05)	-0.02 (-0.04, 0)	0.03 (-0.02, 0.07)	0.03 (-0.02, 0.09)	0.02 (-0.02, 0.06)	-0.03 (-0.06, 0)
S(t) – Ŝ(t) statistic (1 year)	0 (-0.03, 0.03)	-0.03 (-0.05, -0.01)	0.01 (-0.04, 0.05)	0.04 (-0.03, 0.11)	0.06 (0, 0.11)	-0.06 (-0.1, -0.02)
S(t) – Ŝ(t) statistic (2 year)	-0.01 (-0.05, 0.03)	-0.05 (-0.08, -0.02)	0.01 (-0.05, 0.08)	0.06 (-0.02, 0.14)	0.03 (-0.03, 0.08)	-0.05 (-0.11, 0.02)
S(t) – Ŝ(t) statistic (3 year)	-0.05 (-0.09, -0.01)	-0.1 (-0.13, -0.06)	0.01 (-0.07, 0.09)	0.02 (-0.07, 0.1)	0.03 (-0.03, 0.09)	-0.05 (-0.13, 0.03)

NB: S(t) is the probability of recurrence

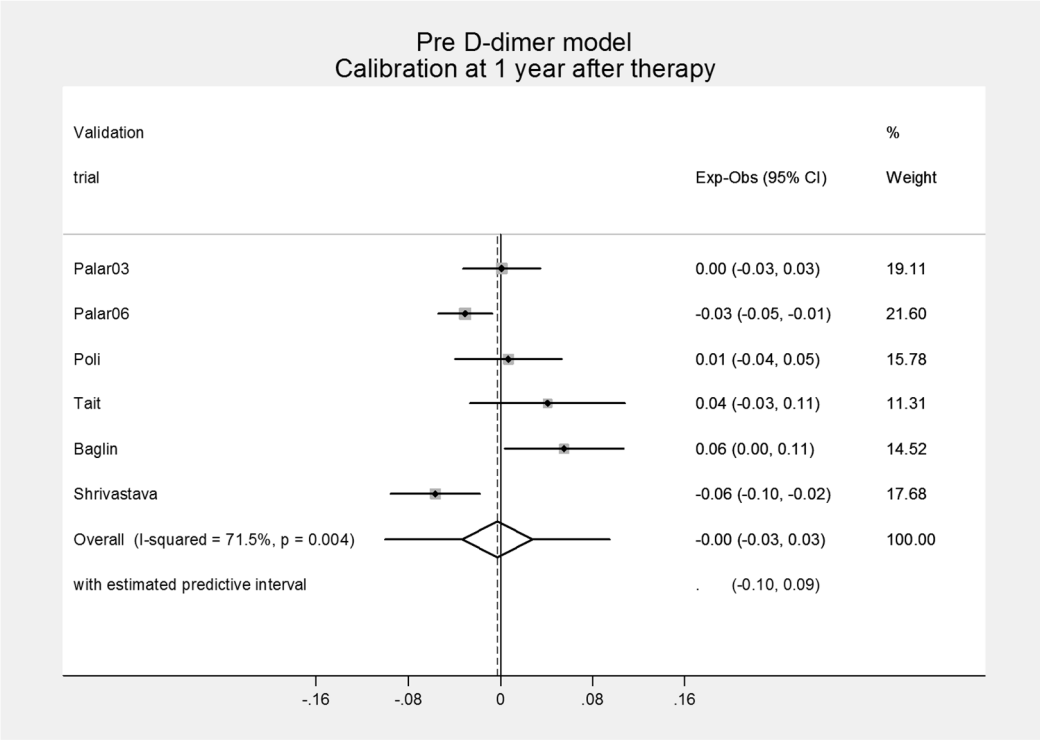


Figure 0.43 - Random-effects meta-analysis of calibration performance (at 1 year post therapy) estimates from each external validation trial in the IECV cycles for the pre D-dimer model

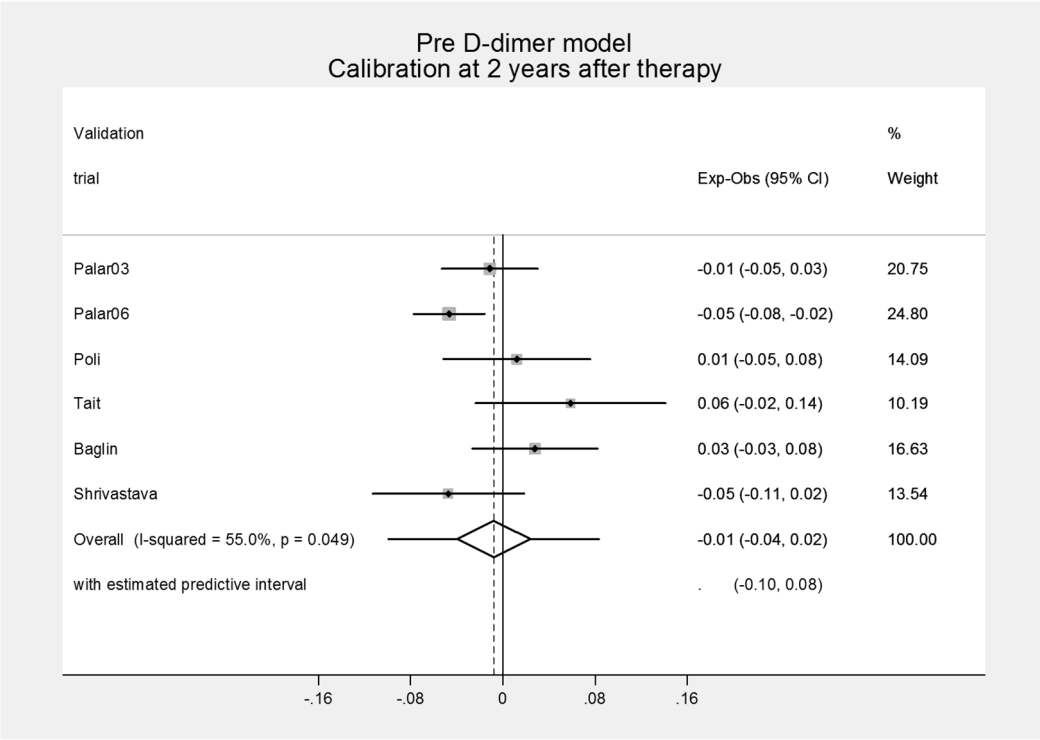


Figure 0.44 - Random-effects meta-analysis of calibration performance (at 2 years post therapy) estimates from each external validation trial in the IECV cycles for the pre D-dimer model

The pooled calibration from a random-effects meta-analysis gives an overall S(t) – $\hat{S}$(t) statistic of zero (95% CI: -0.03, 0.03) at one year post cessation of therapy (*see Figure 0.43*), showing excellent calibration on average. However there appears to be large heterogeneity across trials, with an I-squared statistic of 71.5% suggesting that the calibration of the model is not consistent in all populations. Indeed, the 95% prediction interval ranges from 0.1 to -0.09, indicating that the discrepancy in the predicted and true observed S(t) could range from 0.1 to -0.09, in a particular population. The wide confidence interval is also a reflection of uncertainty in the heterogeneity estimate. Similar results can be seen for a random-effects meta-analysis of calibration statistics at two years post cessation of therapy (*see Figure 0.44*) showing consistent agreement on average in the validation trials at two years.

In summary, discrimination of the model developed is generally poor with C statistics ranging from 0.47 to 0.58 (*see Table 0.10 and Figure 0.40*); other published clinical prediction models have shown stronger discriminatory ability (24). Furthermore, although on average across all trials calibration appears good, there is a large amount of heterogeneity in calibration performance across the different trial populations.

In particular the post D-dimer model showed stronger discriminatory performance, with the prediction interval for the post D-dimer model entirely above the performance indicated for the pre D-dimer model, as such the pre D-dimer model in its current form should not be considered useful for estimating patient's risk of recurrence. Despite the short-comings of the current model, there would be a distinct benefit to making predictions at the time of stopping OAC therapy, and so future work could aim to enhance the pre D-dimer model. Of interest would be the performance of the pre D-dimer model should on treatment D-dimer measurements be included as an additional factor. To aid in such potential future work, the pre D-dimer is presented in full (similar to the post D-dimer model) in the appendix. In this way researchers could easily investigate the added value of updating the pre D-dimer model with predictors such as on therapy D-dimer measurements.

APPENDIX B6: Final pre D-dimer model

Although model performance was generally weak, it was considered important to present a final pre D-dimer model for future research to build on. The final model therefore used the data from all trials, and estimated predictor effects and the baseline hazard, with a random-effect on the baseline hazard to allow for trial differences.

A description, performance and sensitivity analyses of the final model are now presented.

Specification and parameter estimates

The pre D-dimer model was fitted to the whole dataset, with the candidate predictors for patient age, gender, treatment duration and site of index event (distal DVT, proximal DVT and PE) considered. A random-effects model on the baseline hazard was estimated using a FP model with 3d.f. on the proportional hazards scale. The MFP algorithm was used to perform predictor selection as described previously (*see section 3.2.6*); subsequently only

gender and site of index event were selected for inclusion in the final pre D-dimer model. The estimated hazard ratios for included predictors remained similar to those seen throughout the IECV cycles as expected (*see Table 0.11*). Patient's gender and site of index event had large hazard ratios consistent with the literature (40, 196, 272). Male gender was associated with an almost 80% increase in recurrence rate compared to females (HR=1.79, 95% CI: 1.33, 2.41), while proximal DVT and PE were associated with around a six-fold increase in recurrence rate compared to patients with a first distal DVT (*see Table 0.11*).

Table 0.11 - Final specification and estimates for the pre D-dimer model after fitted to all trial data, with a random effect on the baseline hazard

Predictor	Beta coefficient (95% CI)	Hazard ratio (95% CI)	P-value
Gender			
Male	0.58 (0.29, 0.88)	1.79 (1.33, 2.41)	< 0.001
Site of index event			
Proximal DVT	1.82 (0.76, 2.88)	6.17 (2.13, 17.86)	0.001
PE	1.71 (0.64, 2.79)	5.55 (1.9, 16.23)	0.002

To make predictions from the model, the following equation is required

Equation 0.1 - Equation to predict probability of recurrence free survival at time t

$$S(t) = S_0(t)^{\exp(\beta x)}$$

where for the pre D-dimer model, βx within Equation 0.1 is the risk score which is equal to

Equation 0.2 - Risk score equation for the pre D-dimer model

$$\beta x = (0.58 \times Gender\ if\ Male) + (1.82 \times Site\ if\ Proximal\ DVT) + (1.71 \times Site\ if\ PE)$$

and where $S_0(t)$ is the average baseline survival function at a specific time t, which is shown below in Figure 0.45 up to four years post cessation of therapy. Values of $S_0(t)$ can be read from the Kaplan-Meier plot at specific time points (*see Figure 0.45*), as presented in Table 0.12 for six months, one, two and three years post cessation of therapy.

Equation 0.1 allows the prediction of a recurrence free survival probability at a particular time point after cessation of therapy, meaning that the probability of recurrence by a specific time point, R(t), is equal to;

$$R(t) = 1 - S(t)$$

Table 0.12 - Baseline (recurrence free) survival at particular time points to combine with patient specific predictor values for individual risk prediction (Pre D-dimer model)

Model predictor	Time from cessation of therapy			
	6 months	1 year	2 years	3 years
$S_0(t)$	0.9938	0.9895	0.9835	0.9780

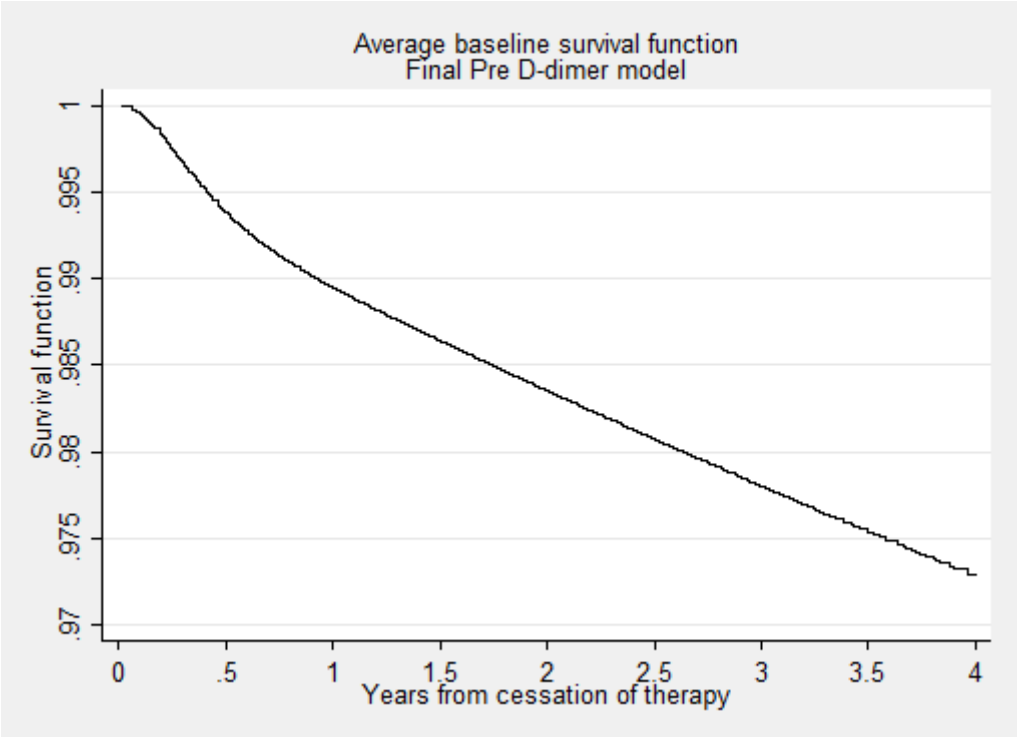


Figure 0.45 - Average baseline (recurrence free) survival function ($S_0(t)$) for the pre D-dimer model

The apparent calibration of the predicted probability of recurrence to the observed probabilities (Kaplan-Meier estimates) within this whole trial dataset appeared under visual inspection to calibrate well up to four years from cessation of therapy (*see Figure 0.46*). This is expected, as the model is estimated on the same dataset, so the apparent calibration is naturally a good fit.

The probability of recurrence over time from cessation of therapy varies across the risk spectrum, illustrating what happens to individuals at the edges of the risk spectrum (46). It can be seen that individuals in the 90th centile of the distribution of the prognostic index having higher probability of recurrence compared to those in the 10th centile of the prognostic index (*see Figure 0.47*). However, the range of discrimination for the model appears to be limited, with little gap between some centiles, which corresponds with the discrimination statistics observed during model development (*see section 0*). This is expected, as the IECV showed the discrimination is low, with the average C-statistic of 0.56 across all cycles (*see Figure 0.40*). The superior discrimination in the post D-dimer model compared to the pre D-dimer model is illustrated by far larger separation in the centiles of risk predictions from the model (*see Figure 0.47 and Figure 3.11*).

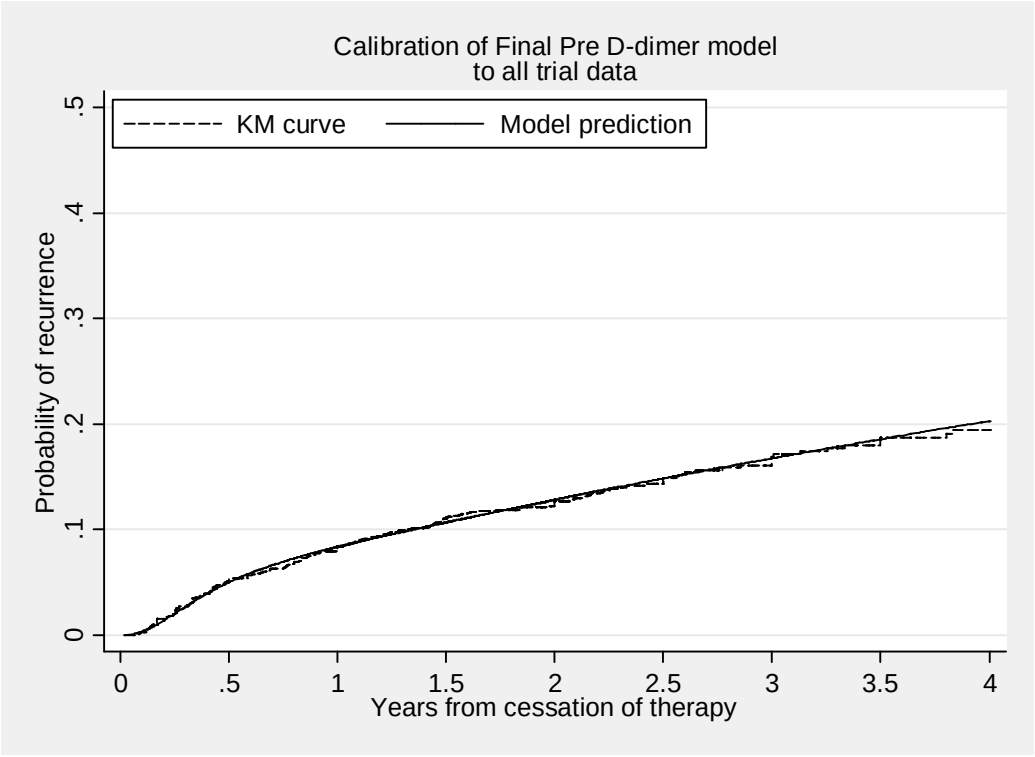


Figure 0.46 - Calibration of the pre D-dimer model fit to all trial data

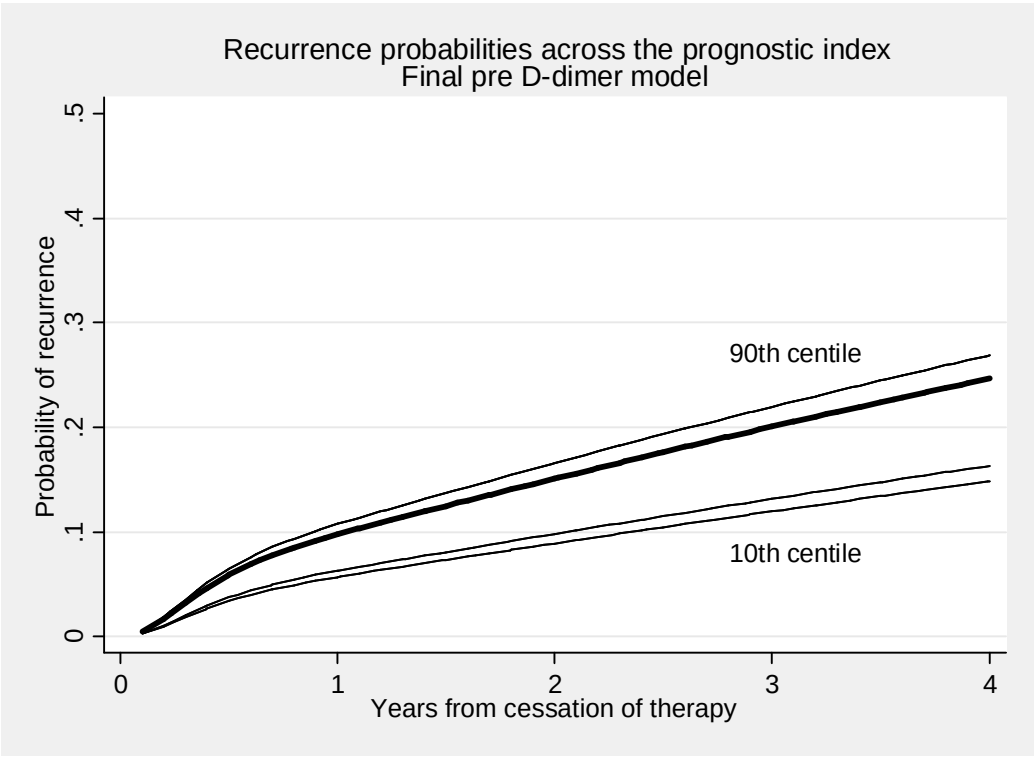


Figure 0.47 - Probability of recurrence across the risk spectrum (The pre D-dimer model)

Model checking & sensitivity analyses

The final model above was checked in terms of proportional hazards assumptions, functional form of continuous predictors (non-linear trends), outliers, leverage, interactions and time-dependent effects. There was no evidence of any concerns, and no indication that the model could be improved or modified in regard these aspects. None of the predictors had missing observations (*see Table 0.3*), and as such a sensitivity analysis using multiple imputation was not required. Sensitivity analyses showed no requirement for the addition of interaction or time dependent effects to the model.

Summary

The final pre D-dimer model proposed in the appendix contained site of index event and gender as predictors. It forms a starting point for individual recurrence risk prediction at the time of stopping therapy, to help inform immediate decisions on the need for extended therapy. However throughout the IECV approach and through external validation of the final model, the performance of the model was rather poor in terms of discrimination and there was heterogeneity in calibration performance across populations. Thus the pre D-dimer model should not currently be recommended for use in practice, and needs improving. One way the model performance may be improved is through the inclusion of more candidate predictors, which may better explain individuals risk and the variation between patients. As such future work may look to investigate the addition of D-dimer measurements taken on therapy as an additional predictor. D-dimer has been shown to be predictive of recurrence throughout the literature (15, 40, 198-203, 217).

APPENDIX B7: Sensitivity analysis on D-dimer assays

A small sensitivity analysis was conducted to crudely assess the impact of differences in the continuous scale of D-dimer assays on the predicted risk of recurrent VTE from the post D-dimer model. Assuming that there could be a potential discrepancy of up to 10% in D-dimer values across assays, the change in predicted risk of recurrence was assessed using example patients with true D-dimer values at the 25th, 50th and 75th percentiles of the distribution of D-dimer values within the RVTE population. All other predictor values were forced to be constant in the model for the predictions. These were varied by 10% either greater or lower, and the resulting predicted survival probabilities were plotted over time (*see Table 0.13 and Figure 0.48 and Figure 0.49 and Figure 0.50*). The figures show very little difference in predicted recurrence free survival, indicating that in practice a similar treatment decision would be made regardless of such a discrepancy in D-dimer measurements.

Table 0.13 - Values of log D-dimer used in post D-dimer model to assess 10% change in D-dimer value

Values of Log(D-dimer) used in the post D-dimer model	Percentile of the dataset		
	25th	**50th**	**75th**
D-dimer 10% lower	247.5	375.75	672.3
Log (D-dimer) 10% lower	5.51	5.93	6.51
D-dimer	275	417.5	747
Log (D-dimer)	5.55	6.03	6.62
D-dimer 10% higher	302.5	459.25	821.7
Log (D-dimer) 10% higher	5.71	6.13	6.71

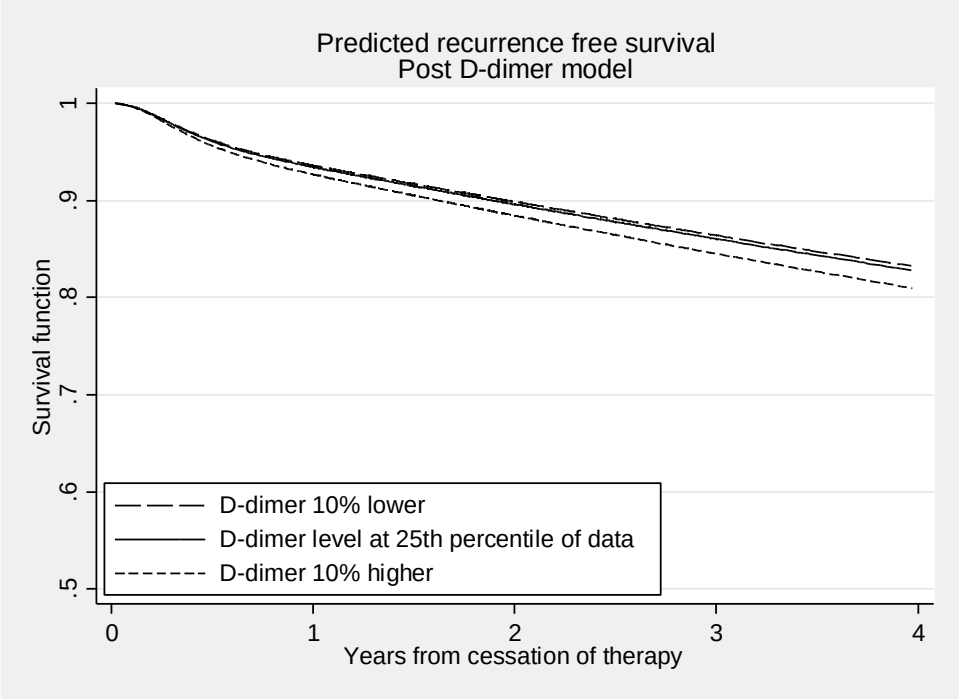


Figure 0.48 - Predicted recurrence free survival for the 25th percentile of D-dimer values & 10% change in D-dimer values

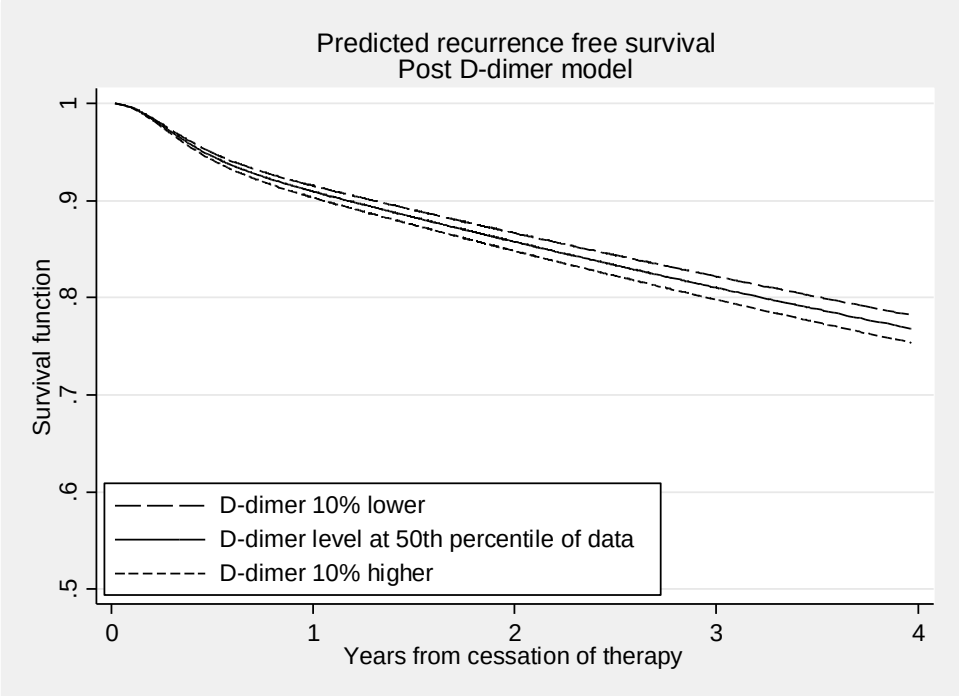


Figure 0.49 - Predicted recurrence free survival for the 50th percentile of D-dimer values & 10% change in D-dimer values

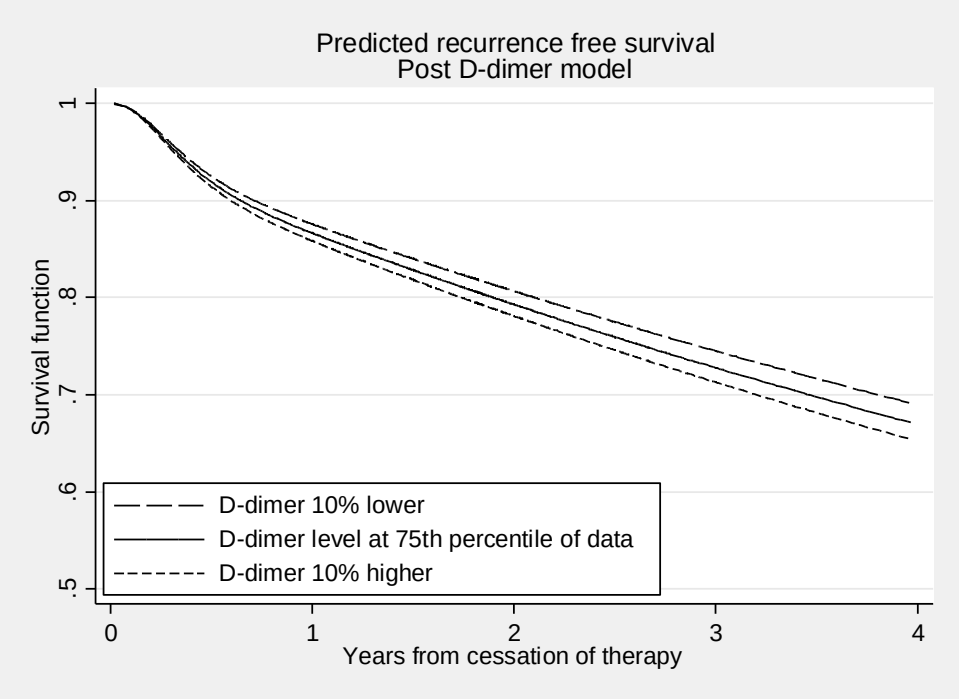


Figure 0.50 - Predicted recurrence free survival for the 75th percentile of D-dimer values & 10% change in D-dimer values

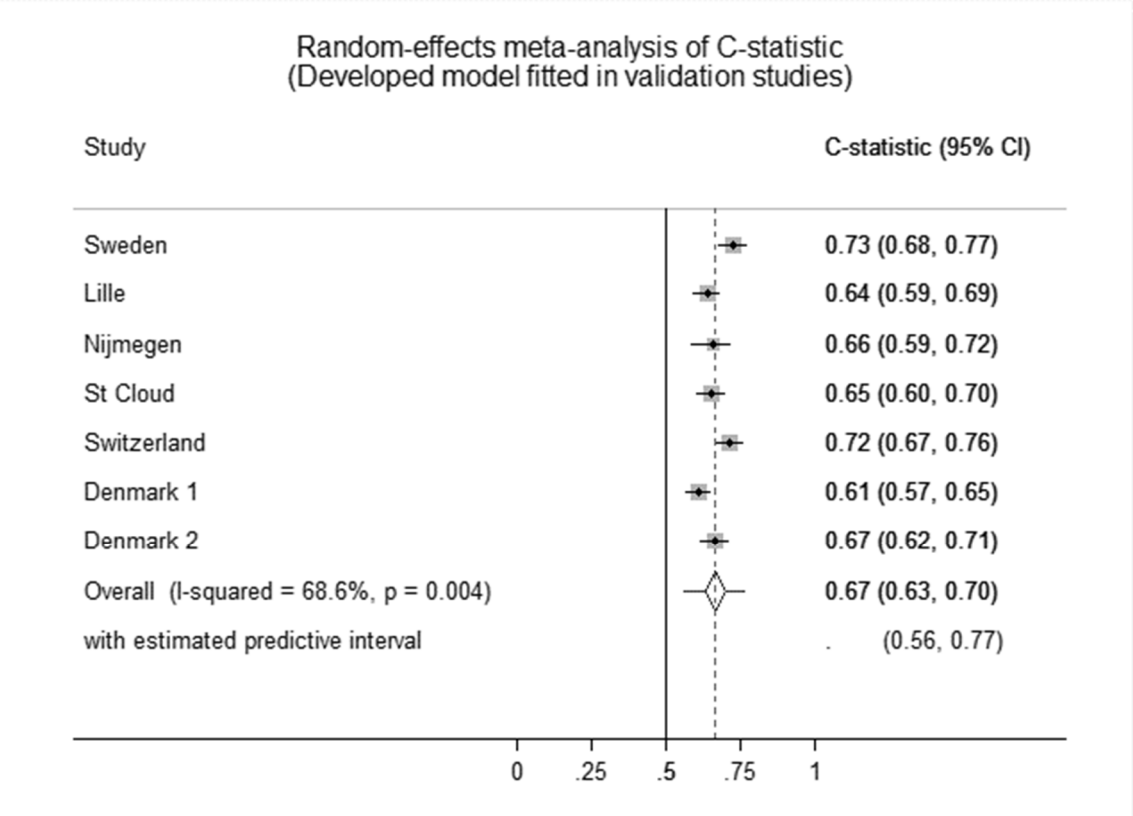
APPENDIX C: Chapter 4 Appendices

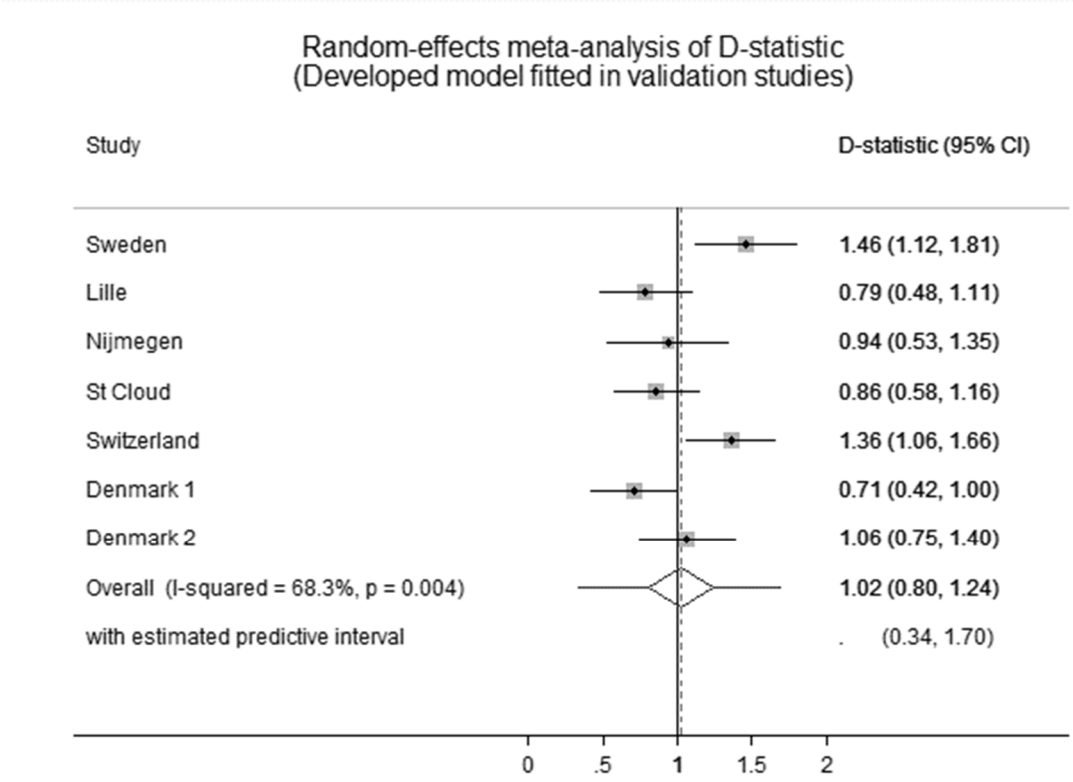
APPENDIX C1: Validation performance

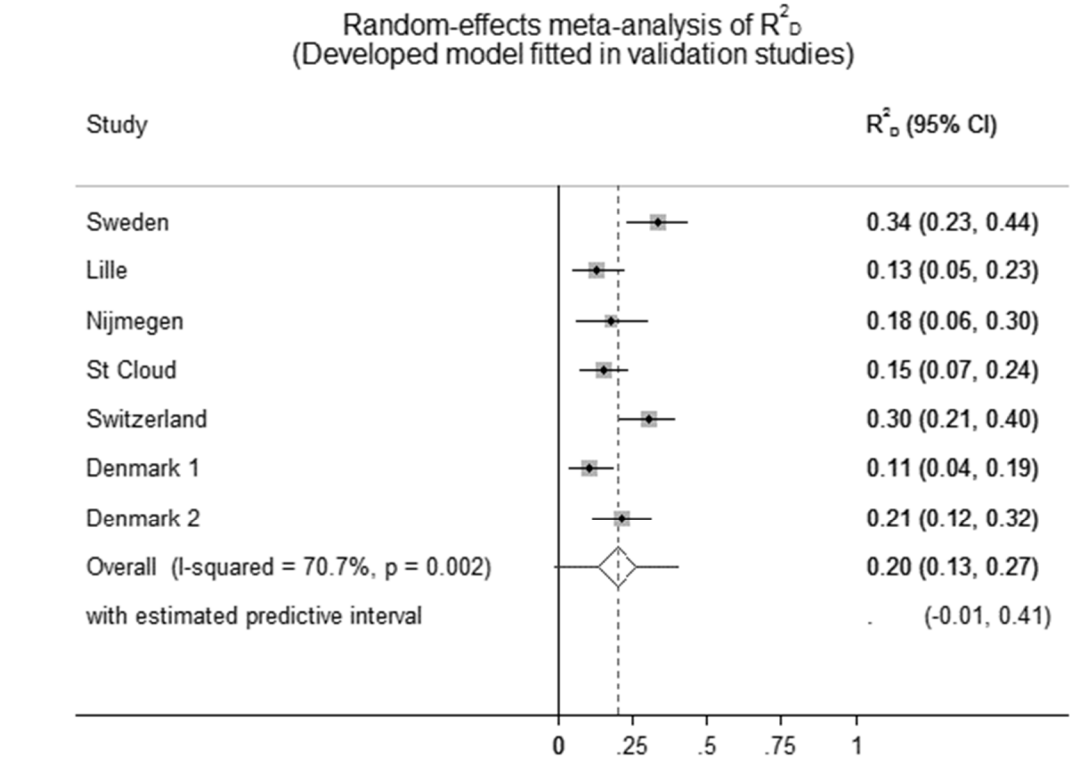
Performance of developed model in validation studies

Discrimination performance

Forest plots showing the meta-analysis of the developed models performance across the validations studies is presented below for the C-statistic, D and R^2_D statistics.

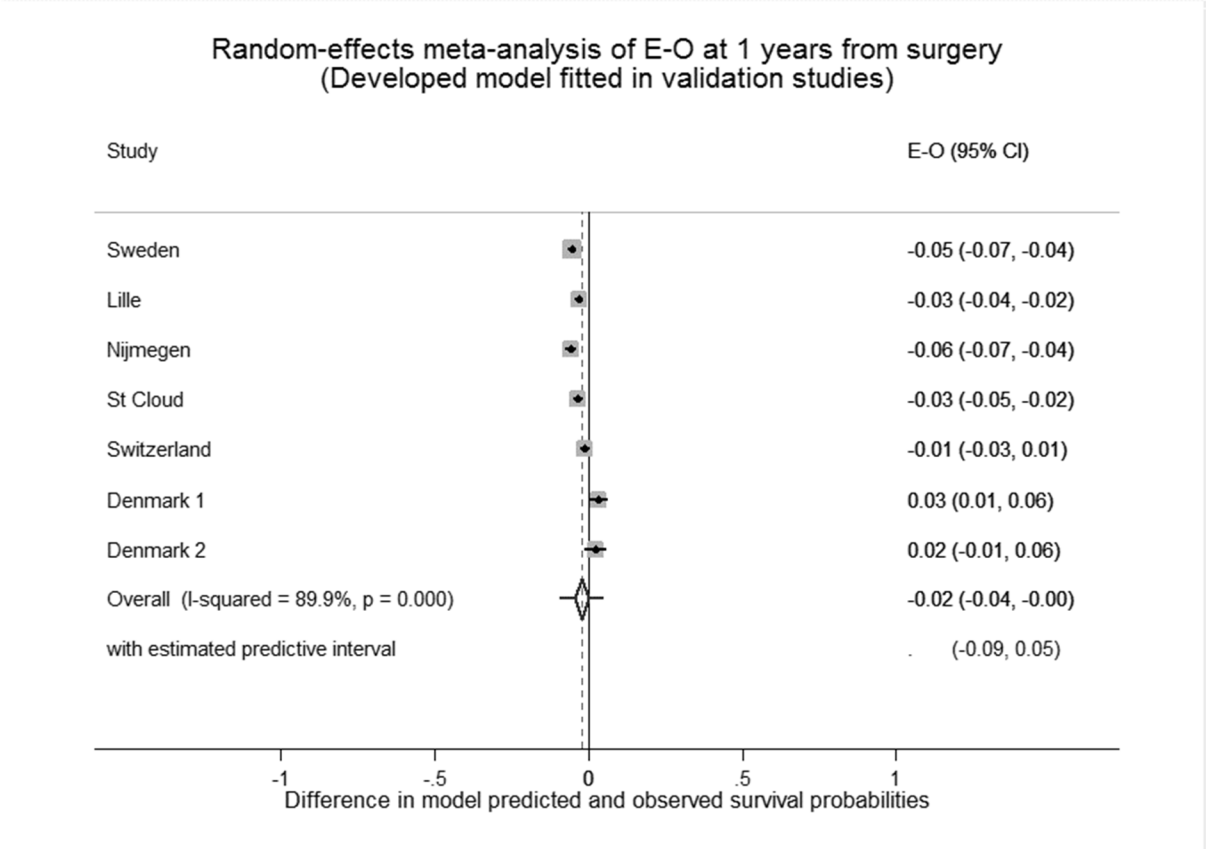


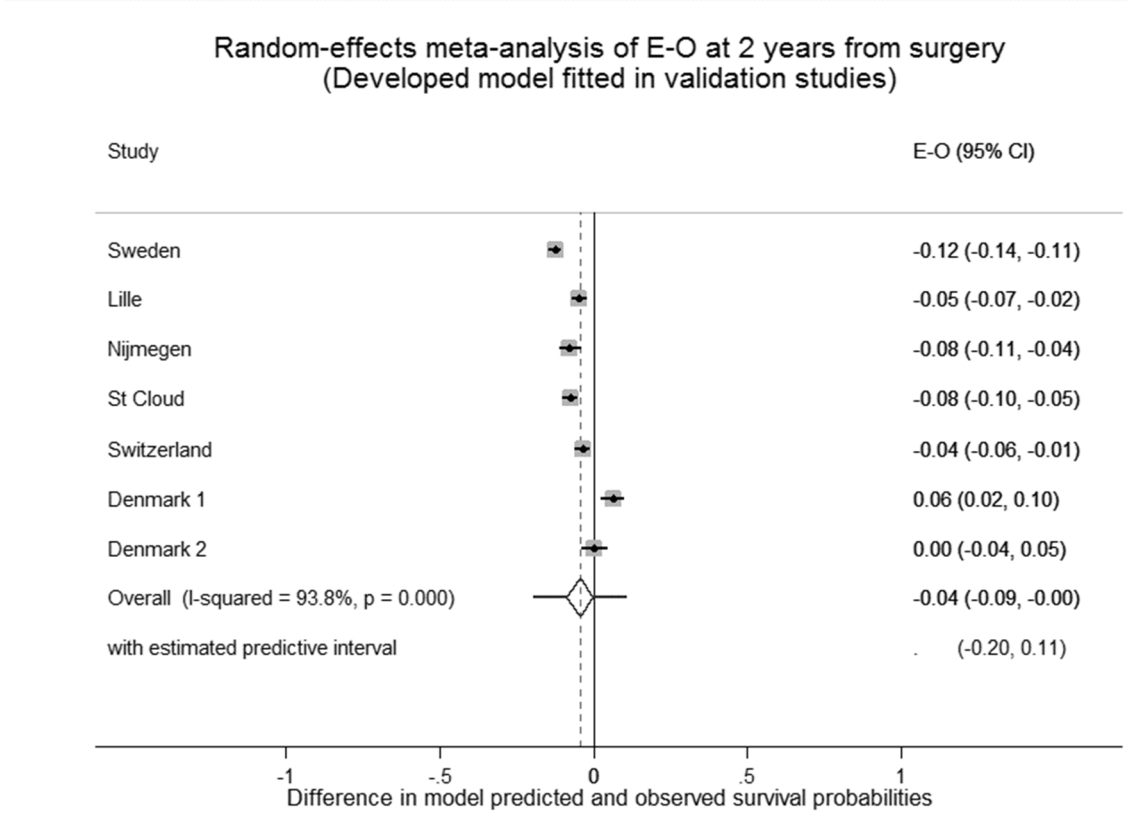

Random-effects meta-analysis of D-statistic
(Developed model fitted in validation studies)
Study
D-statistic (95% CI)
Sweden
1.46 (1.12, 1.81)
Lille
0.79 (0.48, 1.11)
Nijmegen
0.94 (0.53, 1.35)
St Cloud
0.86 (0.58, 1.16)
Switzerland
1.36 (1.06, 1.66)
Denmark 1
0.71 (0.42, 1.00)
Denmark 2
1.06 (0.75, 1.40)
Overall (I-squared = 68.3%, p = 0.004)
1.02 (0.80, 1.24)
with estimated predictive interval
. (0.34, 1.70)
0
.5
1
1.5
2

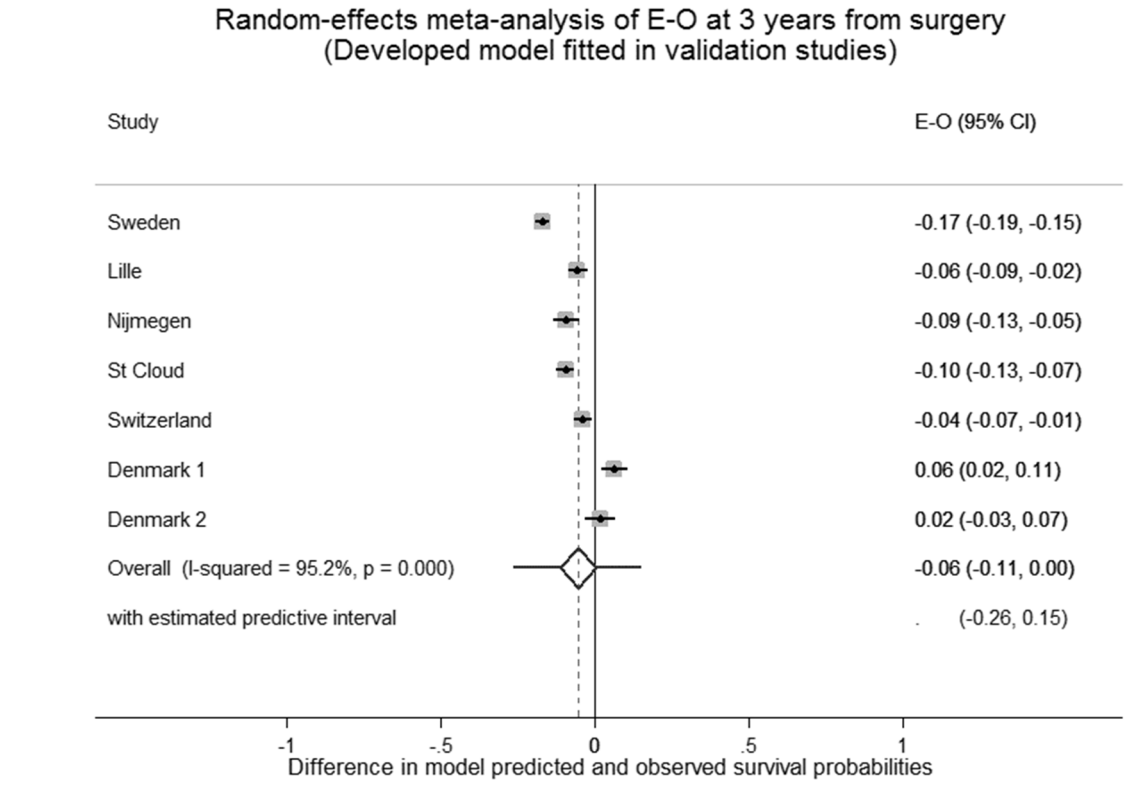

Random-effects meta-analysis of R^2_D
(Developed model fitted in validation studies)
Study
R^2_D (95% CI)
Sweden
0.34 (0.23, 0.44)
Lille
0.13 (0.05, 0.23)
Nijmegen
0.18 (0.06, 0.30)
St Cloud
0.15 (0.07, 0.24)
Switzerland
0.30 (0.21, 0.40)
Denmark 1
0.11 (0.04, 0.19)
Denmark 2
0.21 (0.12, 0.32)
Overall (I-squared = 70.7%, p = 0.002)
0.20 (0.13, 0.27)
with estimated predictive interval
. (-0.01, 0.41)
0
.25
.5
.75
1

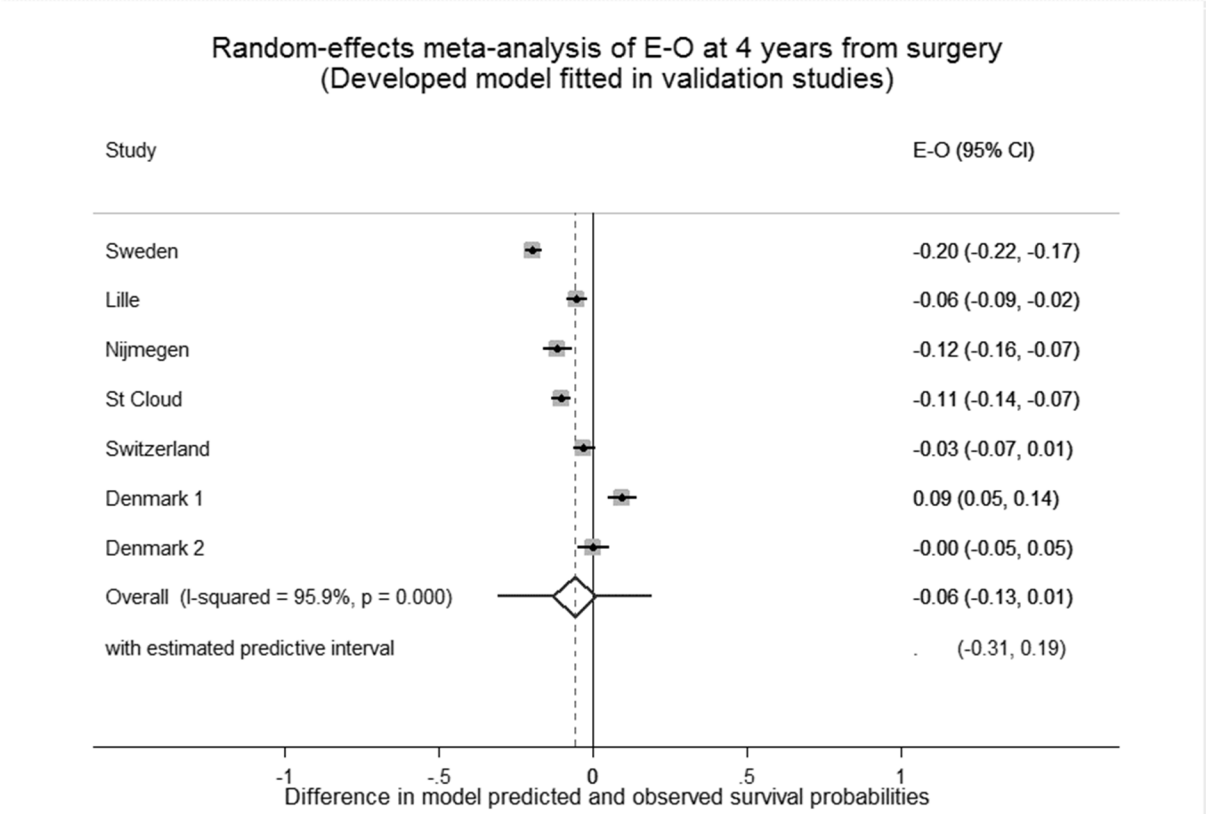
Calibration performance

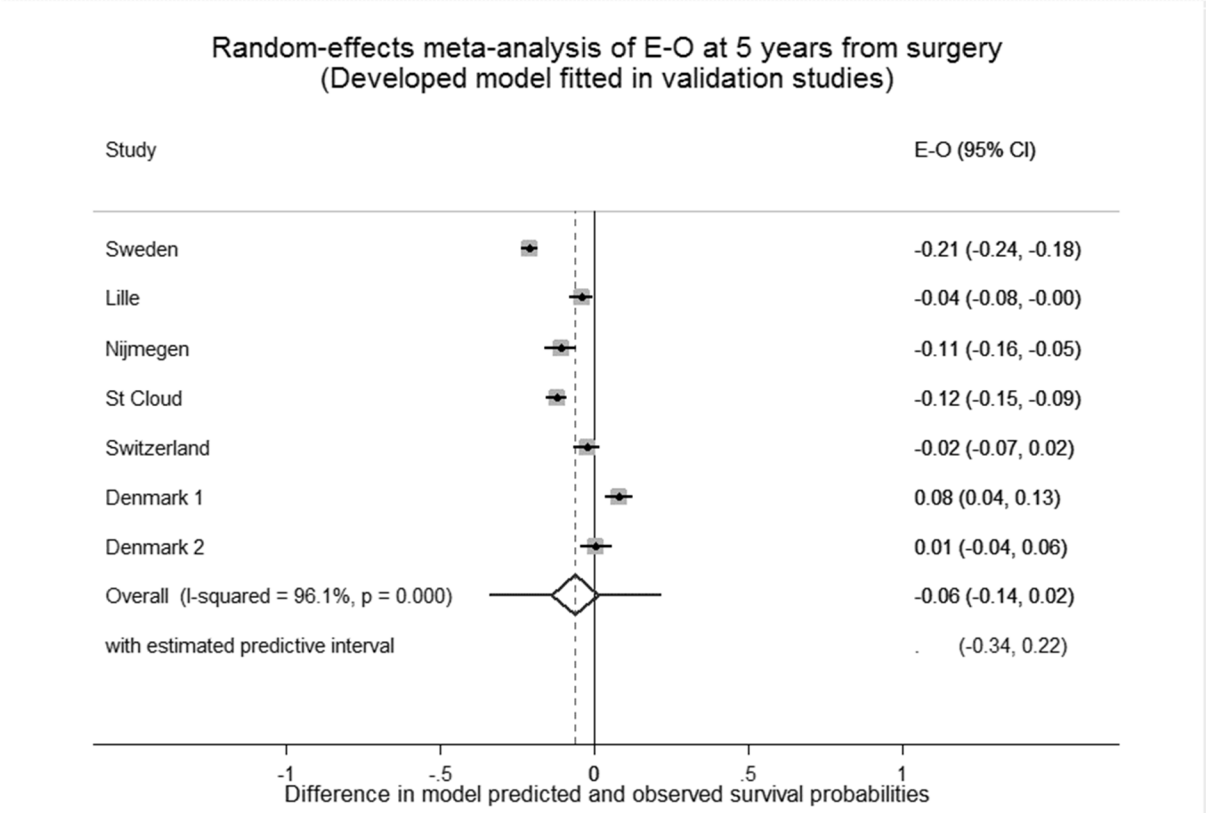
Forest plots showing the meta-analysis of the developed models performance across the validations studies is presented below for the Expected minus Observed statistic at time points including 6 months, and 1 to 6 years from surgery. The expected values describe the predicted survival probabilities from the model, while the observed values give the true observed event probabilities in the validations studies.

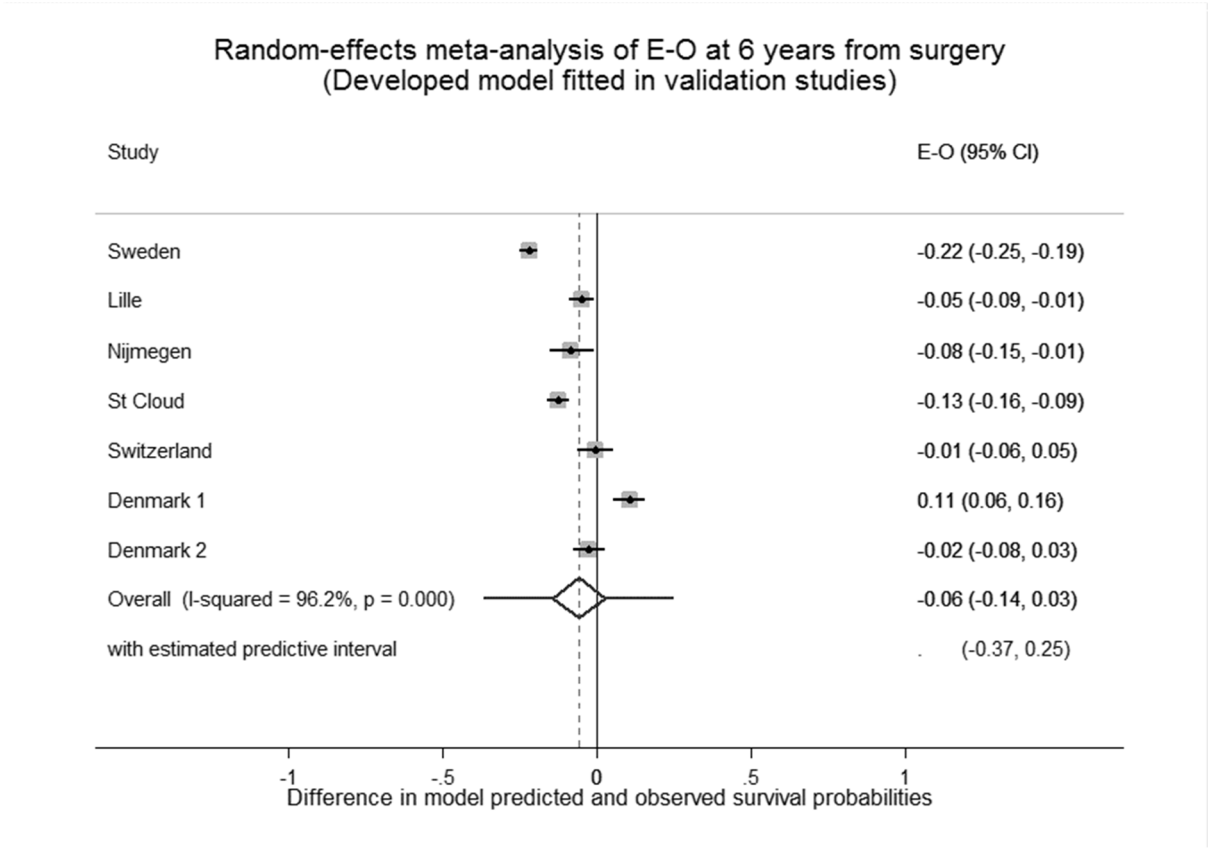



Random-effects meta-analysis of E-O at 2 years from surgery
(Developed model fitted in validation studies)
Study
E-O (95% CI)
Sweden
-0.12 (-0.14, -0.11)
Lille
-0.05 (-0.07, -0.02)
Nijmegen
-0.08 (-0.11, -0.04)
St Cloud
-0.08 (-0.10, -0.05)
Switzerland
-0.04 (-0.06, -0.01)
Denmark 1
0.06 (0.02, 0.10)
Denmark 2
0.00 (-0.04, 0.05)
Overall (I-squared = 93.8%, p = 0.000)
-0.04 (-0.09, -0.00)
with estimated predictive interval
. (-0.20, 0.11)
-1
-.5
0
.5
1
Difference in model predicted and observed survival probabilities


Random-effects meta-analysis of E-O at 3 years from surgery
(Developed model fitted in validation studies)
Study
E-O (95% CI)
Sweden
-0.17 (-0.19, -0.15)
Lille
-0.06 (-0.09, -0.02)
Nijmegen
-0.09 (-0.13, -0.05)
St Cloud
-0.10 (-0.13, -0.07)
Switzerland
-0.04 (-0.07, -0.01)
Denmark 1
0.06 (0.02, 0.11)
Denmark 2
0.02 (-0.03, 0.07)
Overall (I-squared = 95.2%, p = 0.000)
-0.06 (-0.11, 0.00)
with estimated predictive interval
. (-0.26, 0.15)
-1
-.5
0
.5
1
Difference in model predicted and observed survival probabilities


Random-effects meta-analysis of E-O at 4 years from surgery
(Developed model fitted in validation studies)
Study
E-O (95% CI)
Sweden
-0.20 (-0.22, -0.17)
Lille
-0.06 (-0.09, -0.02)
Nijmegen
-0.12 (-0.16, -0.07)
St Cloud
-0.11 (-0.14, -0.07)
Switzerland
-0.03 (-0.07, 0.01)
Denmark 1
0.09 (0.05, 0.14)
Denmark 2
-0.00 (-0.05, 0.05)
Overall (I-squared = 95.9%, p = 0.000)
-0.06 (-0.13, 0.01)
with estimated predictive interval
(-0.31, 0.19)
-1
-.5
0
.5
1
Difference in model predicted and observed survival probabilities


Random-effects meta-analysis of E-O at 5 years from surgery
(Developed model fitted in validation studies)
Study
E-O (95% CI)
Sweden
-0.21 (-0.24, -0.18)
Lille
-0.04 (-0.08, -0.00)
Nijmegen
-0.11 (-0.16, -0.05)
St Cloud
-0.12 (-0.15, -0.09)
Switzerland
-0.02 (-0.07, 0.02)
Denmark 1
0.08 (0.04, 0.13)
Denmark 2
0.01 (-0.04, 0.06)
Overall (I-squared = 96.1%, p = 0.000)
-0.06 (-0.14, 0.02)
with estimated predictive interval
(-0.34, 0.22)
-1
-.5
0
.5
1
Difference in model predicted and observed survival probabilities



Performance of recalibrated model

Random effects meta-analysis plots are given below showing model performance at external validation across all the validation studies, for all recalibration methods.

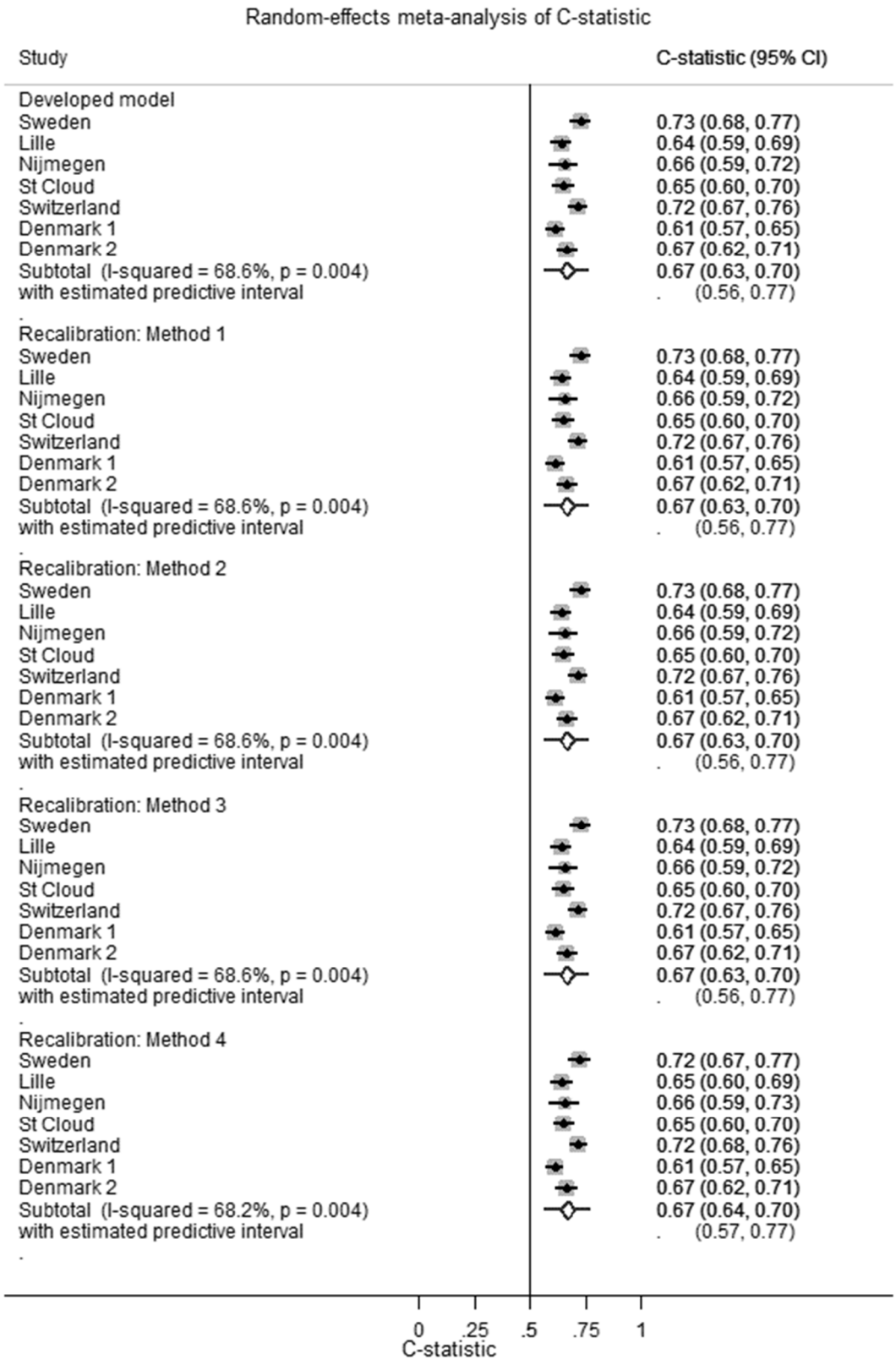


Figure 0.51 - Random effects meta-analysis of discrimination performance (C-statistics) of the model in all validation studies split by recalibration method. Top panel shows performance of the original model in the validation studies.

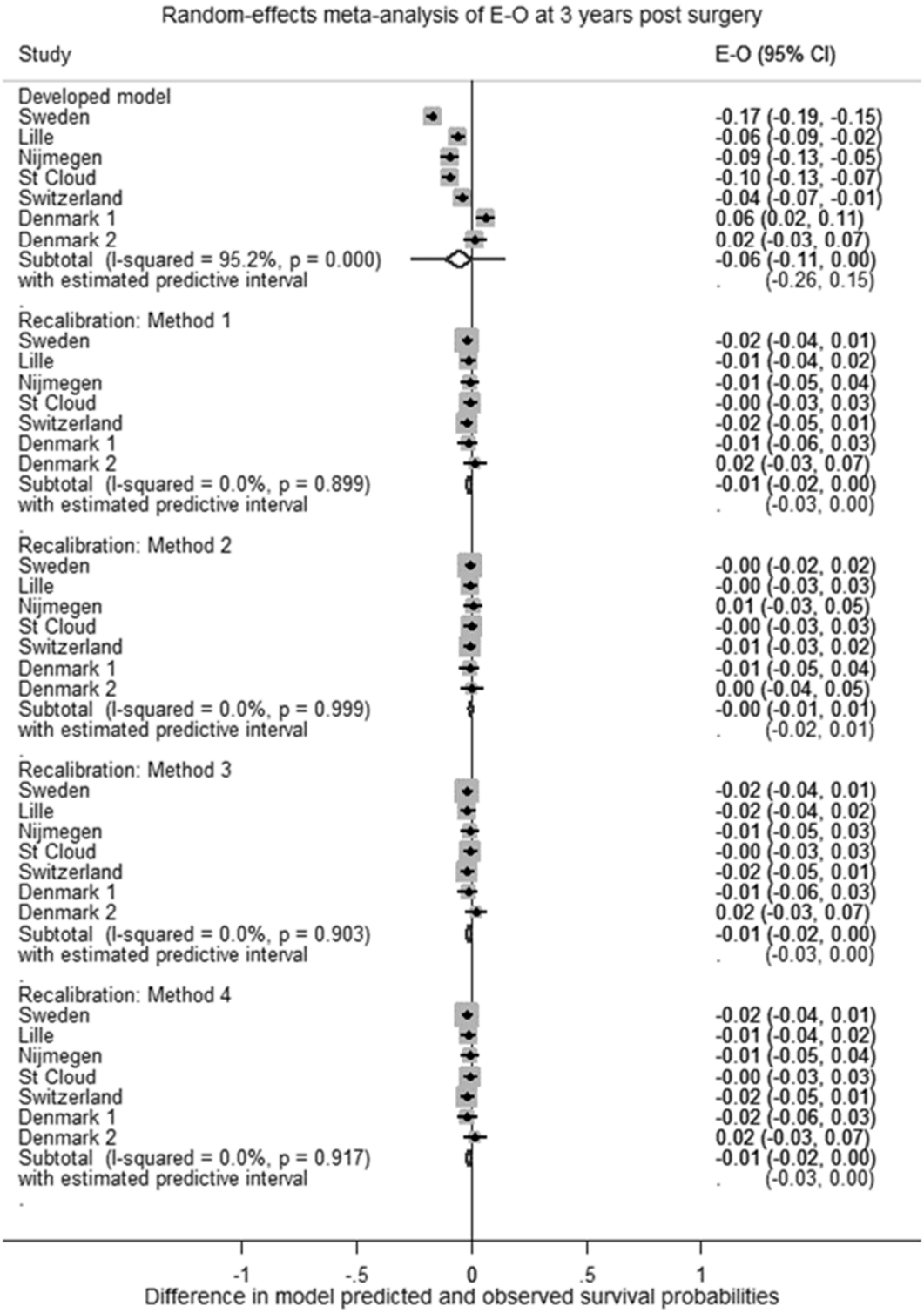


Figure 0.52 - Random effects meta-analysis of calibration performance (E-O at 3 years post-surgery) of the model in all validation studies split by recalibration method. Top panel shows performance of the original model in the validation studies.

APPENDIX C2: Stata code

Code for recalibration methods

The below is an excerpt of Stata 14 code showing how the model was fitted and then recalibrated for each method, it does not detail the process of calculating the performance statistics and 95% confidence intervals relating to each method as this was done using standard bootstrapping methods, looping across all validation studies (labo==`j').

```
// load and set up data and run developed model
clear
use BCdata, replace
// drop out unknown aduj trt patients, & T4 tumour sizes

drop if adjuv==1
drop if pt==4

// set up data as time-to-event
stset dfs, f(dfsi==1) scale(12) exit(time 72)

// Developed model in rotterdam derivation data
xi: stpm2 age i.pt i.np i.post2 i.adjuv if labo==1, all df(3)
scale(h) eform

// predict the linear predictor (for method 3)
predict LP, xbnobaseline

*****************************************************************
// save knots to ensure the same in model on external data
global bhknots `e(bhknots)'
global boundknots `e(boundary_knots)'

// build up constraints
local consval 1
foreach v in `e(varlist)' {
      constraint `consval' _b[`v'] = `=_b[`v']'
      local ++consval
}

foreach v in `e(rcsterms_base)' {
      constraint `consval' _b[`v'] = `=_b[`v']'
      local ++consval
}

constraint `consval' _b[_cons] = `=_b[_cons]'

// list stored constraints
constraint dir

*****************************************************
// METHOD 0
```

```
// make model predictions directly from fitted development model
above
predict xb_mean`j' if labo==`j', meansurv

*********************************************************************
// METHOD 1

// fit model allowing intercept to be re-estimated
qui stpm2 age _Ipt_2 _Ipt_3 _Inp_1 _Inp_2 _Inp_3 _Ipost2_1 _Iadjuv_2
if labo==`j', all ///
     scale(h) constraints(1(1)11) knots(${bhknots})
bknots(${boundknots}) eform


*********************************************************************
// METHOD 2

// fit model allowing intercept & baseline hazard shape to be re-
estimated
qui stpm2 age _Ipt_2 _Ipt_3 _Inp_1 _Inp_2 _Inp_3 _Ipost2_1 _Iadjuv_2
if labo==`j', all ///
     scale(h) constraints(1(1)8) knots(${bhknots})
bknots(${boundknots}) eform


*********************************************************************
// METHOD 3

// fit model allowing intercept to be re-estimated and scaling the
original LP
qui stpm2 LP if labo==`j', all ///
     scale(h) constraints(9(1)11) knots(${bhknots})
bknots(${boundknots})


*********************************************************************
// METHOD 4

// fit model allowing intercept and heterogeneous predictors to be
re-estimated
qui stpm2 age _Ipt_2 _Ipt_3 _Inp_1 _Inp_2 _Inp_3 _Ipost2_1 _Iadjuv_2
if labo==`j', all ///
     scale(h) constraints(1,4(1)11) knots(${bhknots})
bknots(${boundknots}) eform

*********************************************************************
******
```

Code for meta-analysis of recalibration performance

Example random-effects meta-analysis code in Stata version 14, for synthesising C-statistics and E-O statistics is presented below. In the following code fragment, `cstat` and `Em03`,

refer to the average c-statistic and E-O statistic from 1000 bootstrap samples, and the `lci` and `uci` parameters refer to the 2.5th and 97.5th percentiles of the 1000 bootstrap estimates respectively.

```
metan cstat clci cuci, random name(cstat_m2, replace) scheme(sj) ///
           lcols(Study) astext(40) texts(150) rfdist ///
           title("Random-effects meta-analysis of C-statistic"
"(Recalibration method 2)", size(medsmall)) ///
           graphr(col(white)) xlab(0,.25,.5,.75,1) null(.5) ///
           nowt nowarning effect(C-statistic) boxsca(150) ///
           xtitle("C-statistic", size(small))

metan EmO3 cal3lci cal3uci, random name(EmO3_m2, replace) ///
           scheme(sj) ///
           lcols(Study) astext(40) texts(150) rfdist ///
           title("Random-effects meta-analysis of E-O at 3 years
post surgery" "(Recalibration method 2)", size(medsmall)) ///
           graphr(col(white)) xlab(-.5,0,.5) null(0) ///
           nowt nowarning effect(E-O) boxsca(150) ///
           xtitle("Difference in model predicted and observed
survival probabilities", size(small))
```

APPENDIX D: Chapter 5 Appendices

Stata code for MIDC method

The following code fragment defines the MIDC program as described in section 5.3.

```
clear
capture program drop midc
program define midc, rclass

/* Syntax
          STudyvar = Name of study indicator variable (can be a
string or number eg.author )
          THresholdvar = Name of threshold indicator variable (can
be a string or number eg.threshold value)
          SENSitivity = Name of sensitivity variable OR name of new
variable to be created to hold sensitivity
          SPECificity = Name of specificity variable OR name of new
variable to be created to hold specificity
          COVariance = Variance-covariance structure of the random
effects (see help xtmelogit for further details)
                                [NB: Should be specified when
nometa option is omitted]
*/
     syntax , STudyvar(string) THresholdvar(string)
SENSitivity(string) ///
                      SPECificity(string) IMPutations(int)
COVariance(string)

/* Create a copy of the users dataset and then create a template
dataset based on the number of studies
     and thresholds input. Merge the user and template data
together. (This will create any missing
     threshold rows so that the program can look for missing
values)*/

timer on 1

// If convert/logitscale option is not specified then make "yes" the
defualt to ensure these options are 'on'
di _n "Note: Number of imputation datasets = `imputations'"
di  "Note: Bivariate random-effects model fitted with covariance
structure = `covariance'"

tempname st1 st2 th1 th2

// Create study and threshold indicators from string (or numerical)
variable user inputs
egen `st1' = group(`studyvar')
rename `studyvar' `studyvar'_`st2'
rename `st1' `studyvar'
egen `th1' = group(`thresholdvar')
```

```
rename `thresholdvar' `thresholdvar'_`th2'
rename `th1' `thresholdvar'
qui save userdata, replace

// Calculate number of unique studies and thresholds from the
corresponding user given variables
qui unique `studyvar'
local studies = r(sum)

qui unique `thresholdvar'
local thresholds = r(sum)

// Set obs of template dataset
clear
local obs = `studies'*`thresholds'
qui set obs `obs'

gen idnum=_n

// Create a study and threshold indicator in the template dataset
qui egen `studyvar' = seq(), b(`thresholds')
qui egen `thresholdvar' = seq(), t(`thresholds') by(`studyvar')

sort `studyvar' `thresholdvar'
qui save template, replace

// Merge user data into the template
qui merge 1:1 `studyvar' `thresholdvar' using userdata.dta
sort `studyvar' `thresholdvar'


/* If the convert option is on (default) then data is transformed
from a 2x2 table format into sensitivity
     and specificity, missing values are imputed and then finally
converted back to 2x2 table format*/


// Generate a missing values indicator variable
     qui gen missingind=0
     qui replace missingind=1 if tp==.
     qui gen impind=0

// Create true diseased and true non-diseased variables and store
them in a vector
     qui gen td = tp+fn
     qui gen tnd = fp+tn

     local listD = "0"
     local listND = "0"
     local counter = 1
          forvalues k=1/`studies' {
               forvalues y=1/`thresholds' {
                    if missingind[`counter']==0 {
```

```
                        local trueD`k' = td[`counter']
                        local trueND`k' = tnd[`counter']
                        }
                  local counter = `counter'+1
                  }
            local listD = "`listD',`trueD`k''"
            local listND = "`listND',`trueND`k''"
            }

            mat truedis=(`listD')
            mat truenondis=(`listND')

// Use the vector to replace any missing true diseased/non-diseased
with the values for that study
      forvalues z=1/`studies' {
            local zp=`z'+1
            qui replace td=truedis[1,`zp'] if missingind==1 &
`studyvar'==`z'
            qui replace tnd=truenondis[1,`zp'] if missingind==1 &
`studyvar'==`z'
            }

      // Add continuity correction (0.5) to any thresholds with zero
values
qui gen td2=.
qui gen tnd2=.
qui gen tp2=.
qui gen fn2=.
qui gen tn2=.
qui gen fp2=.
forvalues i=1/`obs' {
      if tp[`i']==0 | fn[`i']==0 | tn[`i']==0 | fp[`i']==0 {
            qui replace tp2=tp+.5 if idnum==`i'
            qui replace fn2=fn+.5 if idnum==`i'
            qui replace tn2=tn+.5 if idnum==`i'
            qui replace fp2=fp+.5 if idnum==`i'
            qui replace td2=tp2+fn2 if idnum==`i'
            qui replace tnd2=fp2+tn2 if idnum==`i'
            }
      }

qui replace td2=td if td2==.
qui replace tnd2=tnd if tnd2==.
qui replace tp2=tp if tp2==.
qui replace fn2=fn if fn2==.
qui replace tn2=tn if tn2==.
qui replace fp2=fp if fp2==.

/* Generate the sensitivity and specificity using the 2x2 table
results with any continuity
      correction values required*/

      qui gen `sensitivity'=tp2/td2
```

```
      qui gen `specificity'=tn2/tnd2


/* If the meta-analysis option is on (default) then a meta-analysis
is performed pre-imputation
      so that comparison can be made between estimated pooled sens
and spec before and after imputation*/
qui save orgdata123, replace

tempname metaAnalysis1
tempfile MA1
qui postfile `metaAnalysis1' thresh original_num_thresh pre_imp_sens
pre_sens_SE pre_imp_spec pre_spec_SE pre_logit_sens pre_logit_spec
using `MA1', replace

      forvalues b=1/`thresholds' {
            qui count if `thresholdvar'==`b' & tp!=.
            local dometa = r(N)
            if `dometa'>1 {
                  capture drop n true n1 n0 true1 true0 `studyvar'
`sensitivity' `specificity'

                  //Reshape the data from wide to long format
                  qui gen long n1=tp+fn
                  qui gen long n0=fp+tn
                  qui gen long true1=tp
                  qui gen long true0=tn
                  qui gen long id0001= _n
                  qui reshape long n true, i(id0001) j(sens0001)
                  qui sort id0001 sens0001
                  qui gen byte spec0001=1-sens0001

                  if `dometa'>5 {
                        local dometa1=5
                        }
                        else {
                              local dometa1=`dometa'
                              }

                  //Perform meta-analysis where more than one study
reports threshold results
                  qui xtmelogit true sens0001 spec0001 if
`thresholdvar'==`b', nocons || `studyvar': sens0001 spec0001, ///
                        nocons cov("`covariance'") binomial(n)
refineopts(iterate(3)) intpoints(`dometa1') variance

                  local sensB = _b[sens0001]
                  local specB = _b[spec0001]
                  local sensP = (exp(`sensB'))/(1+(exp(`sensB')))
                  local specP = (exp(`specB'))/(1+(exp(`specB')))
                  local se1 = _se[sens0001]
                  local se2 = _se[spec0001]
```

```
                }
                else {
                      qui su sens if `thresholdvar'==`b' & tp!=.
                      local sensP = r(mean)
                      qui su spec if `thresholdvar'==`b' & tp!=.
                      local specP = r(mean)

                      local sensB = ln(`sensP'/(1-`sensP'))
                      local specB = ln(`specP'/(1-`specP'))

                      qui su td if `thresholdvar'==`b' & tp!=.
                      local tdP = r(mean)
                      local se1 = (1/(`tdP'*`sensP'*(1-`sensP')))^.5

                      qui su tnd if `thresholdvar'==`b' & tp!=.
                      local tndP = r(mean)
                      local se2 = (1/(`tndP'*`specP'*(1-
`specP')))^.5
                      }

           post `metaAnalysis1' (`b') (`dometa') (`sensP') (`se1')
(`specP') (`se2') (`sensB') (`specB')
           use orgdata123, replace
           }
      postclose `metaAnalysis1'
      use `MA1', replace
      qui save MApreimpdata, replace

use orgdata123, replace


qui drop td2
qui drop tnd2
qui drop tp2
qui drop tn2
qui drop fp2
qui drop fn2
qui drop `sensitivity'
qui drop `specificity'


/* Preserve the dataset so that we can drop out study data in order
to impute missing sensitivity and
      specificity values*/
      preserve
      qui save imputation_loop_data, replace


local m = 1
while `m'<=`imputations' {
      di _n "Imputing dataset `m' " _continue
```

```
/* Loop through the studies, dropping out all but one study in each
cycle to allow analysis
      of each study individually*/

forvalues stud=1/`studies' {
      qui drop if `studyvar'!=`stud'

      local tminus=`thresholds'-1

// Cycle through the thresholds looking for those without data,
which could be imputed

forvalues i=2/`tminus' {
      if missingind[`i']==1 & impind[`i']!=1 {

            local down=`i'+1

/* When a threshold with missing data is found then the program will
cycle down the list of thresholds
      looking for the next threshold with data and store this data*/

            forvalues j=`down'/`thresholds' {
                  if missingind[`j']==0 {
                        local DthreshNM= `j'
                        local DtpNM= tp[`j']
                        local DfpNM= fp[`j']
                        continue, break
                        }
                   else if `j'==`thresholds' {
                        local DtpNM= .
                        local DfpNM= .
                        }
                  }

/* The program will then cycle up the list of thresholds from the
missing threshold, looking for
      the next threshold with data above it and store this data*/

            qui save looking_for_higher_thresh, replace
            qui drop if `thresholdvar'>`i'

            forvalues k= 1/`i' {
                  local real_thresh=(`i'-`k')
                  if missingind[`real_thresh']==0 {
                        local UthreshNM= `real_thresh'
                        local UtpNM = tp[`real_thresh']
                        local UfpNM = fp[`real_thresh']
                        continue, break
                        }
                   else if `real_thresh'==1 {
                        local UtpNM = .
                        local UfpNM = .
                        }
```

```
                }

/* Check if there are known bounding thresholds non-missing */
if `DtpNM'!=. & `UtpNM'!=. {

/* Calculate the necessary parameters*/

      local num_miss = `DthreshNM'-`UthreshNM'-1

      if `DtpNM'<=`UtpNM' {
            local pot_tp = `UtpNM'-`DtpNM'+1
            local pot_fp = `UfpNM'-`DfpNM'+1
            }
            else if `DtpNM'>=`UtpNM' {
                  local pot_tp = `DtpNM'-`UtpNM'+1
                  local pot_fp = `DfpNM'-`UfpNM'+1
                  }

      // Calculate the number of possible combinations
      local numer_tp = `pot_tp'+`num_miss'-1
      local pot_comb_tp = comb(`numer_tp',`num_miss')

      local numer_fp = `pot_fp'+`num_miss'-1
      local pot_comb_fp = comb(`numer_fp',`num_miss')


/* Create matrix of possible combinations and select one at random
to impute the missing TP */

      use looking_for_higher_thresh, replace

// calculate random number to select combination from all possible
combinations with repitition

      local rand = 1+int((`pot_comb_tp'-1+1)*runiform())
      return scalar rand_tp_`m'=`rand'

      local count=0

// run combination calculator program to get selected combination

      /*
      NT = number of missing thresholds
      KT = number of possible numbers
      ST = starting value (upper limit)
      RT = random combination selected
      */

      combswrep, nt(`num_miss') kt(`pot_tp') st(`UtpNM') rt(`rand')


      // which thresholds to impute to
      local thresh_1_id=`i'
```

```
      if `num_miss'>1 {
            local extra=2
            forvalues f=1/`num_miss' {
                  local thresh_`extra'_id = `i'+`f'
                  local extra = `extra'+1
                  }
            }

      forvalues g=1/`num_miss' {

            qui replace tp = `r(T`g')' if
`thresholdvar'==`thresh_`g'_id' & missingind==1
            di "." _continue

            qui replace fn = td - tp if
`thresholdvar'==`thresh_`g'_id' & missingind==1
            qui replace impind=1 if `thresholdvar'==`thresh_`g'_id'
            }

            qui save looking_for_higher_thresh, replace

      /* Create matrix of possible combinations and select one at
random to impute the missing fp */

      // calculate random number to select combination from all
possible combinations with repitition

      local rand = 1+int((`pot_comb_fp'-1+1)*runiform())
      return scalar rand_fp_`m'=`rand'

      local count=0

// run combination calculator program to get selected combination

      /*
      NT = number of missing thresholds
      KT = number of possible numbers
      ST = starting value (upper limit)
      RT = random combination selected
      */

      combswrep, nt(`num_miss') kt(`pot_fp') st(`UfpNM') rt(`rand')

// which thresholds to impute to
      local thresh_1_id=`i'
      if `num_miss'>1 {
            local extra=2
            forvalues f=1/`num_miss' {
                  local thresh_`extra'_id = `i'+`f'
                  local extra = `extra'+1
                  }
            }
```

```
      forvalues g=1/`num_miss' {

            qui replace fp = `r(T`g')' if
`thresholdvar'==`thresh_`g'_id' & missingind==1

            qui replace tn = tnd - fp if
`thresholdvar'==`thresh_`g'_id' & missingind==1
            qui replace impind=1 if `thresholdvar'==`thresh_`g'_id'

            }

            qui save looking_for_higher_thresh, replace

      } // from if statement to check we have non-missing bounding
thresholds

            use looking_for_higher_thresh, replace

                  } // only perform this loop if the threshold we are
on is missing

            } // cycle thresholds looking for missing thresholds

/* Save the imputed data file for this study and then restore all
studies data from memory so that the
      loop can continue by imputing data for the next study in the
dataset*/

      qui save imputed_study_data`stud', replace
      restore, preserve
      } // cycle through isolating studies to look for missing
thresholds within studies

/* When all the studies have been looped through and their missing
thresholds imputed, the dataset can
      be restored and wiped from memory*/
restore, not

/* Append all the imputed study data files into one complete meta-
analytic dataset, with
      all sensitivity and specificity imputed where possible*/

      use imputed_study_data1, replace

      if `studies'>1 {
            forvalues a=2/`studies' {
                  qui append using imputed_study_data`a'
                  qui save imputed_dataset_data, replace
                  }
            }

/* Look through the imputed 2x2 values and correct any where the
value should be equal to the above
```

```
      and below thresholds because the two surrounding thresholds are
also equal (only if the convert
      option is on (by default convert is set to on)*/

preserve
forvalues j=1/`studies' {
      qui drop if `studyvar'!=`j'

      forvalues i=1/`thresholds' {
            if missingind[`i']==1 {
                  local TPabove = tp[`i'-1]
                  local TPbelow = tp[`i'+1]
                  if `TPabove'==`TPbelow' {
                        qui replace tp = `TPabove' if
`thresholdvar'==`i' & missingind==1
                        }
                  local FPabove = fp[`i'-1]
                  local FPbelow = fp[`i'+1]
                  if `FPabove'==`FPbelow' {
                        qui replace fp = `FPabove' if
`thresholdvar'==`i' & missingind==1
                        }
                  local FNabove = fn[`i'-1]
                  local FNbelow = fn[`i'+1]
                  if `FNabove'==`FNbelow' {
                        qui replace fn = `FNabove' if
`thresholdvar'==`i' & missingind==1
                        }
                  local TNabove = tn[`i'-1]
                  local TNbelow = tn[`i'+1]
                  if `TNabove'==`TNbelow' {
                        qui replace tn = `TNabove' if
`thresholdvar'==`i' & missingind==1
                        }
            }
      }
      qui save final_study_data`j', replace
      restore, preserve
}

restore, not
use final_study_data1, replace

if `studies'>1 {
            forvalues a=2/`studies' {
                  qui append using final_study_data`a'
                  qui save final_imputed_threshold_dataset`m', replace
                  }
            }


/* Now calculate the sens/spec from the imputed (and original) data.
      Allow for continuity corrections where necessary. */
```

```
// Add continuity correction (0.5) to any thresholds with zero
values
qui gen td2=.
qui gen tnd2=.
qui gen tp2=.
qui gen fn2=.
qui gen tn2=.
qui gen fp2=.
forvalues i=1/`obs' {
     if tp[`i']==0 | fn[`i']==0 | tn[`i']==0 | fp[`i']==0 {
          qui replace tp2=tp+.5 if idnum==`i'
          qui replace fn2=fn+.5 if idnum==`i'
          qui replace tn2=tn+.5 if idnum==`i'
          qui replace fp2=fp+.5 if idnum==`i'
          qui replace td2=tp2+fn2 if idnum==`i'
          qui replace tnd2=fp2+tn2 if idnum==`i'
          }
     }

qui replace td2=td if td2==.
qui replace tnd2=tnd if tnd2==.
qui replace tp2=tp if tp2==.
qui replace fn2=fn if fn2==.
qui replace tn2=tn if tn2==.
qui replace fp2=fp if fp2==.

/* Generate the sensitivity and specificity using the 2x2 table
results with any continuity
     correction values required*/

     qui gen `sensitivity'=tp2/td2
     qui gen `specificity'=tn2/tnd2

// Finally, drop out unrequired variables from finally dataset
qui drop _merge

qui drop td2
qui drop tnd2
qui drop tp2
qui drop tn2
qui drop fp2
qui drop fn2


qui save final_imputed_threshold_dataset`m', replace

/* List out number of imputations performed*/

/* Create scalars representing the number of imputed values at each
threshold, and the total number
     of inmputed thresholds. Display these statistics to the user.
*/
```

```
local imp = 0

forvalues i=1/`thresholds' {
      qui count if `thresholdvar'==`i' & missingind==1 &
`sensitivity'!=.
      local imp_`i' = r(N)
      return scalar imp_`i' = `imp_`i''
      local imp = `imp'+`imp_`i''
      }
di _n "Total thresholds imputed = `imp'"
return scalar imp = `imp'

/* Perform meta-analysis*/

tempname metaAnalysis2
tempfile MA2
qui postfile `metaAnalysis2' imp_dataset thresh number_imputed sens
se1 lci1 uci1 spec se2 lci2 uci2 tau1 tau2 logit1 logit2 conv convFE
using `MA2', replace

      forvalues b=1/`thresholds' {
            qui count if `thresholdvar'==`b' & `sensitivity'!=.
            local dometa = r(N)
            return scalar dometa_`b' = `dometa'
            if `dometa'>1 {
                  capture drop n true n1 n0 true1 true0 `studyvar'
`sensitivity' `specificity'

                  //Reshape the data from wide to long format
                  qui gen long n1=tp+fn
                  qui gen long n0=fp+tn
                  qui gen long true1=tp
                  qui gen long true0=tn
                  qui gen long id0001= _n
                  qui reshape long n true, i(id0001) j(sens0001)
                  qui sort id0001 sens0001
                  qui gen byte spec0001=1-sens0001

                  if `dometa'>5 {
                        local dometa1=5
                        }
                        else {
                              local dometa1=`dometa'
                              }

                  //Perform meta-analysis where more than one study
reports threshold results
                  capture qui xtmelogit true sens0001 spec0001 if
`thresholdvar'==`b', nocons || `studyvar': sens0001 spec0001, ///
                        nocons cov("`covariance'") binomial(n)
refineopts(iterate(3)) intpoints(`dometa1') variance iterate(100)

                  if _rc==0 {
```

```
                        local sensB = _b[sens0001]
                        local specB = _b[spec0001]
                        local sensP =
(exp(`sensB'))/(1+(exp(`sensB')))
                        local specP =
(exp(`specB'))/(1+(exp(`specB')))
                        local se1 = _se[sens0001]
                        local se2 = _se[spec0001]

                        local conv = e(converged)
                        local convFE = .
                        return scalar conv`b' = `conv'
                        return scalar convFE`b' = `convFE'

                        qui regsave using regs1, p ci replace
                        use regs1, replace

                        local lci1= ci_lower[1]

                        local uci1= ci_upper[1]

                        local lci2= ci_lower[2]

                        local uci2= ci_upper[2]

                        local tau1 = coef[3]
                        local tau2 = coef[4]
                        local tau1 = exp(`tau1')

                        local tau2 = exp(`tau2')

                        }
                        else {
                              qui capture glm true sens0001 spec0001 if
`thresholdvar'==`b', nocons family(binomial n) link(logit)
iterate(100)

                              local sensB = _b[sens0001]
                              local specB = _b[spec0001]
                              local sensP =
(exp(`sensB'))/(1+(exp(`sensB')))
                              local specP =
(exp(`specB'))/(1+(exp(`specB')))
                              local se1 = _se[sens0001]
                              local se2 = _se[spec0001]

                              local conv = .
                              local convFE = e(converged)

                              qui regsave using regs1, p ci replace
                              use regs1, replace

                              local lci1= ci_lower[1]
```

```
                              local uci1= ci_upper[1]
                              local lci2= ci_lower[2]
                              local uci2= ci_upper[2]

                              local tau1 = .
                              local tau2 = .

                        }

                  }
                  else {

                        qui su sens if `thresholdvar'==`b' & tp!=.
                        local sensP = r(mean)
                        qui su spec if `thresholdvar'==`b' & tp!=.
                        local specP = r(mean)

                        local sensB = ln(`sensP'/(1-`sensP'))
                        local specB = ln(`specP'/(1-`specP'))

                        qui su td if `thresholdvar'==`b' & tp!=.
                        local tdP = r(mean)
                        local se1 = (1/(`tdP'*`sensP'*(1-`sensP')))^.5

                        local lci1 = `sensB'-(1.96*`se1')

                        local uci1 = `sensB'+(1.96*`se1')

                        qui su tnd if `thresholdvar'==`b' & tp!=.
                        local tndP = r(mean)
                        local se2 = (1/(`tndP'*`specP'*(1-
`specP')))^.5

                        local lci2 = `specB'-(1.96*`se2')

                        local uci2 = `specB'+(1.96*`se2')

                        local tau1 = .
                        local tau2 = .

                        local conv = .
                        local convFE = .
                        }

            post `metaAnalysis2' (`m') (`b') (`dometa') (`sensP')
(`se1') (`lci1') (`uci1') (`specP') (`se2') (`lci2') (`uci2')
(`tau1') (`tau2') (`sensB') (`specB') (`conv') (`convFE')
            use final_imputed_threshold_dataset`m', replace
            }

      if _rc==0 {
            postclose `metaAnalysis2'
            use `MA2', replace
```

```
            qui save MApostimpdata_`m', replace
            }


      qui use imputation_loop_data, replace
      preserve

      local m = `m'+1


} // loop for multiple imputations

restore, not

/* Combine the pre-imputed results with the post-imputation combined
results*/

      use MApostimpdata_1, replace

      if `imputations'>1 {
            forvalues m1=2/`imputations' {
                  qui append using MApostimpdata_`m1'
                  qui save all_MApostimpdata, replace
                  }
            }


/* Combine using rubins rules */

qui gen diff_sens = .
qui gen diffsq_sens = .
qui gen diff_spec = .
qui gen diffsq_spec = .

      forvalues h2=1/`thresholds' {
            qui su sens if thresh==`h2'
            local pooled_sens_`h2' = (1/`imputations')*(r(sum))

            qui su se1 if thresh==`h2'
            local within_sens_var_`h2' = (1/`imputations')*(r(sum))



            qui su spec if thresh==`h2'
            local pooled_spec_`h2' = (1/`imputations')*(r(sum))

            qui su se2 if thresh==`h2'
            local within_spec_var_`h2' = (1/`imputations')*(r(sum))



            qui replace diff_sens = sens-`pooled_sens_`h2'' if
thresh==`h2'
```

```
            qui replace diffsq_sens = diff_sens^2 if thresh==`h2'

            qui su diffsq_sens if thresh==`h2'
            local btwn_sens_var_`h2' = (1/(`imputations'-1))*(r(sum))



            qui replace diff_spec = spec-`pooled_spec_`h2'' if
thresh==`h2'
            qui replace diffsq_spec = diff_spec^2 if thresh==`h2'

            qui su diffsq_spec if thresh==`h2'
            local btwn_spec_var_`h2' = (1/(`imputations'-1))*(r(sum))



            local pooled_sens_var_`h2' =
`within_sens_var_`h2''+((1+(1/`imputations'))*`btwn_sens_var_`h2'')
            local pooled_spec_var_`h2' =
`within_spec_var_`h2''+((1+(1/`imputations'))*`btwn_spec_var_`h2'')

      ************************************************************

            qui su logit1 if thresh==`h2'
            local pooled_logit1_`h2' = (1/`imputations')*(r(sum))

            qui su logit2 if thresh==`h2'
            local pooled_logit2_`h2' = (1/`imputations')*(r(sum))

            qui su tau1 if thresh==`h2'
            local pooled_tau1_`h2' = (1/`imputations')*(r(sum))

            qui su tau2 if thresh==`h2'
            local pooled_tau2_`h2' = (1/`imputations')*(r(sum))

            qui su lci1 if thresh==`h2'
            local pooled_lci1_`h2' = (1/`imputations')*(r(sum))

            qui su uci1 if thresh==`h2'
            local pooled_uci1_`h2' = (1/`imputations')*(r(sum))

            qui su lci2 if thresh==`h2'
            local pooled_lci2_`h2' = (1/`imputations')*(r(sum))

            qui su uci2 if thresh==`h2'
            local pooled_uci2_`h2' = (1/`imputations')*(r(sum))

            qui su number_imputed if thresh==`h2'
            local pooled_studies_`h2' = (1/`imputations')*(r(sum))

            qui su conv if thresh==`h2'
            local pooled_conv_`h2' = (1/`imputations')*(r(sum))
```

```
          qui su convFE if thresh==`h2'
          local pooled_convFE_`h2' = (1/`imputations')*(r(sum))
          }

     qui save rubins_data, replace

tempname metaAnalysis3
tempfile MA3
qui postfile `metaAnalysis3' thresh number_imputed psens pspec pse1
pse2 plci1 puci1 plci2 puci2 plogit1 plogit2 ptau1 ptau2 pconv
pconvFE using `MA3', replace

     forvalues b=1/`thresholds' {
          if   sens>=0 {

               local sensP = `pooled_sens_`b''
               return scalar sens`b' = `sensP'

               local specP = `pooled_spec_`b''
               return scalar spec`b' = `specP'

               local sensB_se = `pooled_sens_var_`b''
               return scalar se1`b' = `sensB_se'

               local specB_se = `pooled_spec_var_`b''
               return scalar se2`b' = `specB_se'

               local sensB = `pooled_logit1_`b''
               return scalar logit1`b' = `sensB'

               local specB = `pooled_logit2_`b''
               return scalar logit2`b' = `specB'

               local tau1 = `pooled_tau1_`b''
               return scalar tau1`b' = `tau1'

               local tau2 = `pooled_tau2_`b''
               return scalar tau2`b' = `tau2'

               local lci1 =  `pooled_lci1_`b''
               return scalar lci1`b' = `lci1'

               local uci1 =  `pooled_uci1_`b''
               return scalar uci1`b' = `uci1'

               local lci2 =  `pooled_lci2_`b''
               return scalar lci2`b' = `lci2'

               local uci2 =  `pooled_uci2_`b''
               return scalar uci2`b' = `uci2'

               local pnumber_imputed = `pooled_studies_`b''
               return scalar studies`b' = `pnumber_imputed'
```

```
                local conv = `pooled_conv_`b''
                return scalar conv`b' = `conv'

                local convFE = `pooled_convFE_`b''
                return scalar convFE`b' = `convFE'

                }
                else {
                      local sensP = .
                      local specP = .
                      local sensB_se = .
                      local specB_se = .
                      local sensB = .
                      local specB = .
                      local tau1 = .
                      local tau2 = .
                      local lci1 = .
                      local uci1 = .
                      local lci2 = .
                      local uci2 = .
                      local pnumber_imputed = .
                      local conv = .
                      local convFE = .

                return scalar sens`b' = `sensP'
                return scalar spec`b' = `specP'
                return scalar se1`b' = `sensB_se'
                return scalar se2`b' = `specB_se'
                return scalar logit1`b' = `sensB'
                return scalar logit2`b' = `specB'
                return scalar tau1`b' = `tau1'
                return scalar tau2`b' = `tau2'
                return scalar lci1`b' = `lci1'
                return scalar uci1`b' = `uci1'
                return scalar lci2`b' = `lci2'
                return scalar uci2`b' = `uci2'
                return scalar studies`b' = `pnumber_imputed'
                return scalar conv`b' = `conv'
                return scalar convFE`b' = `convFE'
                      }

           qui post `metaAnalysis3' (`b') (`pnumber_imputed')
(`sensP') (`specP') (`sensB_se') (`specB_se') (`lci1') (`uci1')
(`lci2') (`uci2') (`sensB') (`specB') (`tau1') (`tau2') (`conv')
(`convFE')
           use rubins_data, replace
           }
     postclose `metaAnalysis3'
     use `MA3', replace
     qui save MAcombinedimpdata, replace
```

```
      /* Combine the pre-imputed results with the post-imputation
combined results*/

      // Merge user data into the template
      /*qui merge 1:1 thresh using MApreimpdata.dta
      sort thresh
      drop _merge
      order thresh original_num_thresh number_imputed pre_imp_sens
pre_imp_spec post_imp_sens post_imp_spec post_sens_se post_spec_se
      */
      qui save pooled_ma_results, replace

forvalues e=1/`studies' {
      erase final_study_data`e'.dta
      erase imputed_study_data`e'.dta
      }
erase orgdata123.dta
/*
if "`MA'"=="yes" {
      erase MApreimpdata.dta
            forvalues e1=1/`imputations' {
                  erase MApostimpdata_`e1'.dta
                  *erase final_imputed_threshold_dataset`e1'.dta
                  }
}
*/
erase template.dta
erase userdata.dta
erase looking_for_higher_thresh.dta
erase imputed_dataset_data.dta

timer off 1

qui timer list
local prog_time = r(t1)
local onoffs = r(nt1)
return scalar prog_time = `prog_time'
return scalar timer_switch = `onoffs'

di _n "Program took `prog_time' seconds, for each imputed dataset"


qui save pooled_ma_results, replace

di "Results are stored in memory"

end
```

APPENDIX E: Chapter 6 Appendices

APPENDIX E1: Base case scenarios

Scenario 1

Table 0.14 - Results for summary sensitivity for scenario 1

Method	Threshold	No. Studies*	Sensitivity	Bias	SE	MSE	LCI	UCI	Coverage
NI	1	5.12	0.95	-0.007	0.68	0.46	0.83	0.98	95.87
NI	2	5.08	0.93	-0.008	0.61	0.37	0.82	0.97	95.07
NI	3	4.92	0.92	-0.007	0.55	0.30	0.80	0.97	96.01
NI	4	5.01	0.90	-0.005	0.50	0.25	0.78	0.95	95.87
NI	5	5.03	0.87	-0.005	0.44	0.20	0.75	0.94	95.74
NI	6	5.00	0.84	-0.003	0.39	0.15	0.71	0.91	96.40
NI	7	5.03	0.80	-0.003	0.36	0.13	0.67	0.88	96.41
NI	8	5.03	0.75	-0.003	0.33	0.11	0.62	0.85	95.73
NI	9	5.00	0.70	-0.002	0.31	0.10	0.56	0.81	95.35
NI	10	4.92	0.64	0.000	0.30	0.09	0.50	0.76	95.08
NI	11	4.99	0.58	0.004	0.29	0.08	0.44	0.71	94.81
SI	1	5.12	0.95	-0.007	0.68	0.46	0.83	0.98	95.87
SI	2	7.60	0.93	-0.012	0.45	0.20	0.85	0.97	92.69
SI	3	8.79	0.91	-0.014	0.36	0.13	0.83	0.95	92.15
SI	4	9.38	0.89	-0.013	0.32	0.10	0.81	0.94	93.22
SI	5	9.61	0.86	-0.012	0.29	0.08	0.78	0.92	94.15
SI	6	9.70	0.83	-0.011	0.26	0.07	0.75	0.89	95.48
SI	7	9.60	0.79	-0.010	0.24	0.06	0.71	0.86	94.68
SI	8	9.32	0.75	-0.010	0.23	0.05	0.65	0.82	94.95
SI	9	8.70	0.69	-0.009	0.22	0.05	0.60	0.78	94.41
SI	10	7.46	0.63	-0.008	0.23	0.05	0.53	0.73	94.68
SI	11	4.99	0.58	0.004	0.29	0.08	0.44	0.71	94.81
MIDC	1	5.12	0.95	-0.009	0.68	0.46	0.83	0.98	95.74
MIDC	2	7.60	0.93	-0.007	0.47	0.22	0.86	0.97	94.68
MIDC	3	8.79	0.92	-0.008	0.38	0.14	0.84	0.96	95.61
MIDC	4	9.38	0.89	-0.008	0.32	0.10	0.82	0.94	94.95
MIDC	5	9.61	0.87	-0.008	0.29	0.08	0.79	0.92	96.14
MIDC	6	9.70	0.83	-0.007	0.26	0.07	0.75	0.89	95.61
MIDC	7	9.60	0.80	-0.006	0.24	0.06	0.71	0.86	95.88
MIDC	8	9.32	0.75	-0.005	0.23	0.05	0.66	0.82	96.01
MIDC	9	8.70	0.70	-0.003	0.22	0.05	0.60	0.78	95.61
MIDC	10	7.46	0.64	-0.001	0.23	0.05	0.53	0.74	95.48
MIDC	11	4.99	0.58	0.003	0.29	0.08	0.44	0.71	94.68

Table 0.15 - Results for summary specificity for scenario 1

Method	Threshold	No. Studies*	Specificity	Bias	SE	MSE	LCI	UCI	Coverage
NI	1	5.12	0.57	0.000	0.09	0.01	0.52	0.61	95.47
NI	2	5.08	0.64	0.001	0.10	0.01	0.59	0.68	94.14
NI	3	4.92	0.70	0.001	0.10	0.01	0.66	0.74	94.41
NI	4	5.01	0.76	0.001	0.11	0.01	0.71	0.79	94.80
NI	5	5.03	0.80	0.001	0.12	0.01	0.77	0.84	94.94
NI	6	5.00	0.85	0.001	0.13	0.02	0.81	0.88	94.41
NI	7	5.03	0.88	-0.001	0.14	0.02	0.85	0.90	93.35
NI	8	5.03	0.91	0.000	0.16	0.03	0.88	0.93	94.27
NI	9	5.00	0.93	0.000	0.18	0.03	0.90	0.95	94.81
NI	10	4.92	0.95	0.001	0.21	0.04	0.92	0.96	94.81
NI	11	4.99	0.96	0.001	0.24	0.06	0.94	0.97	95.74
SI	1	5.12	0.57	0.000	0.09	0.01	0.52	0.61	95.47
SI	2	7.60	0.64	-0.001	0.08	0.01	0.60	0.67	95.08
SI	3	8.79	0.70	-0.001	0.07	0.01	0.67	0.73	95.35
SI	4	9.38	0.75	-0.001	0.08	0.01	0.73	0.78	94.41
SI	5	9.61	0.80	-0.001	0.08	0.01	0.78	0.83	94.81
SI	6	9.70	0.84	-0.001	0.09	0.01	0.82	0.87	94.41
SI	7	9.60	0.88	-0.001	0.10	0.01	0.86	0.90	92.95
SI	8	9.32	0.91	0.000	0.11	0.01	0.89	0.92	94.41
SI	9	8.70	0.93	0.000	0.13	0.02	0.91	0.94	94.68
SI	10	7.46	0.95	0.001	0.16	0.03	0.93	0.96	94.55
SI	11	4.99	0.96	0.001	0.24	0.06	0.94	0.97	95.74
MIDC	1	5.12	0.57	-0.001	0.09	0.01	0.52	0.61	95.35
MIDC	2	7.60	0.63	-0.001	0.08	0.01	0.60	0.67	94.55
MIDC	3	8.79	0.70	-0.004	0.07	0.01	0.66	0.72	92.69
MIDC	4	9.38	0.75	-0.005	0.07	0.01	0.72	0.78	92.02
MIDC	5	9.61	0.80	-0.006	0.08	0.01	0.77	0.82	92.42
MIDC	6	9.70	0.84	-0.006	0.09	0.01	0.82	0.86	92.29
MIDC	7	9.60	0.87	-0.005	0.10	0.01	0.85	0.89	91.76
MIDC	8	9.32	0.90	-0.004	0.11	0.01	0.88	0.92	92.55
MIDC	9	8.70	0.93	-0.003	0.13	0.02	0.91	0.94	92.95
MIDC	10	7.46	0.94	-0.001	0.16	0.03	0.93	0.96	93.62
MIDC	11	4.99	0.96	0.000	0.24	0.06	0.94	0.97	95.61

Scenario 2

Table 0.16 - Results for summary sensitivity for scenario 2

Method	Threshold	No. Studies*	Sensitivity	Bias	SE	MSE	LCI	UCI	Coverage	$\hat{\tau}$	$\hat{\tau}$ bias
NI	1	5.15	0.95	-0.005	0.83	0.68	0.80	0.98	96.09	0.25	0.00
NI	2	5.13	0.94	-0.002	0.73	0.54	0.79	0.98	97.27	0.21	0.04
NI	3	5.08	0.92	-0.001	0.66	0.43	0.78	0.97	96.61	0.22	0.03
NI	4	4.97	0.90	-0.004	0.59	0.34	0.75	0.96	95.96	0.18	0.07
NI	5	5.02	0.88	0.000	0.51	0.26	0.73	0.94	97.39	0.19	0.06
NI	6	5.03	0.84	-0.004	0.45	0.20	0.69	0.92	96.61	0.18	0.07
NI	7	4.97	0.80	-0.004	0.41	0.17	0.64	0.89	96.35	0.18	0.07
NI	8	4.95	0.75	-0.002	0.39	0.15	0.60	0.86	97.13	0.17	0.08
NI	9	4.93	0.70	-0.004	0.36	0.13	0.54	0.82	95.31	0.17	0.08
NI	10	4.96	0.64	-0.005	0.34	0.11	0.48	0.76	95.18	0.15	0.10
NI	11	4.90	0.57	-0.003	0.33	0.11	0.42	0.71	96.74	0.16	0.09
SI	1	5.15	0.95	-0.005	0.83	0.68	0.80	0.98	96.09	0.25	0.00
SI	2	7.62	0.93	-0.008	0.52	0.27	0.84	0.97	95.44	0.20	0.05
SI	3	8.80	0.91	-0.010	0.42	0.17	0.83	0.96	94.27	0.18	0.07
SI	4	9.37	0.89	-0.011	0.36	0.13	0.80	0.94	95.05	0.18	0.07
SI	5	9.63	0.86	-0.011	0.32	0.10	0.77	0.92	95.57	0.19	0.06
SI	6	9.67	0.83	-0.012	0.29	0.08	0.74	0.89	95.44	0.17	0.08
SI	7	9.57	0.79	-0.013	0.27	0.07	0.69	0.86	95.18	0.18	0.07
SI	8	9.27	0.74	-0.012	0.25	0.06	0.64	0.82	95.18	0.18	0.07
SI	9	8.64	0.69	-0.013	0.25	0.06	0.58	0.78	94.01	0.17	0.08
SI	10	7.44	0.63	-0.013	0.26	0.07	0.51	0.74	95.44	0.17	0.08
SI	11	4.90	0.57	-0.003	0.33	0.11	0.42	0.71	96.74	0.16	0.09
MIDC	1	5.15	0.95	-0.006	0.83	0.68	0.80	0.98	95.96	0.24	0.01
MIDC	2	7.62	0.94	-0.003	0.55	0.30	0.84	0.97	96.88	0.22	0.03
MIDC	3	8.80	0.92	-0.003	0.45	0.20	0.83	0.96	96.61	0.22	0.03
MIDC	4	9.37	0.90	-0.005	0.38	0.14	0.81	0.95	96.61	0.22	0.03
MIDC	5	9.63	0.87	-0.005	0.33	0.11	0.78	0.93	97.27	0.22	0.03
MIDC	6	9.67	0.84	-0.006	0.30	0.09	0.74	0.90	97.40	0.22	0.03
MIDC	7	9.57	0.80	-0.007	0.27	0.08	0.70	0.87	97.27	0.21	0.04
MIDC	8	9.27	0.75	-0.006	0.26	0.07	0.64	0.83	97.14	0.20	0.05
MIDC	9	8.64	0.70	-0.006	0.25	0.06	0.58	0.79	96.48	0.19	0.06
MIDC	10	7.44	0.64	-0.006	0.26	0.07	0.51	0.74	96.09	0.17	0.08
MIDC	11	4.90	0.57	-0.004	0.33	0.11	0.42	0.71	96.48	0.16	0.09

Table 0.17 - Results for summary specificity for scenario 2

Method	Threshold	No. Studies*	Specificity	Bias	SE	MSE	LCI	UCI	Coverage	$\hat{\tau}$	$\hat{\tau}$ bias
NI	1	5.15	0.57	-0.001	0.13	0.02	0.50	0.63	86.18	0.16	0.09
NI	2	5.13	0.63	-0.001	0.13	0.02	0.57	0.69	88.80	0.16	0.09
NI	3	5.08	0.70	0.000	0.14	0.02	0.64	0.75	90.49	0.16	0.09
NI	4	4.97	0.76	0.000	0.15	0.02	0.70	0.80	89.19	0.15	0.10
NI	5	5.02	0.80	-0.001	0.15	0.02	0.75	0.84	89.31	0.15	0.10
NI	6	5.03	0.84	-0.002	0.16	0.03	0.80	0.88	89.18	0.14	0.11
NI	7	4.97	0.88	0.000	0.18	0.03	0.84	0.91	91.13	0.14	0.11
NI	8	4.95	0.91	0.000	0.19	0.04	0.87	0.93	92.70	0.13	0.12
NI	9	4.93	0.93	0.000	0.22	0.05	0.89	0.95	93.10	0.13	0.12
NI	10	4.96	0.94	0.000	0.24	0.06	0.91	0.96	93.36	0.14	0.11
NI	11	4.90	0.96	0.000	0.28	0.08	0.93	0.97	93.47	0.14	0.11
SI	1	5.15	0.57	-0.001	0.13	0.02	0.50	0.63	86.18	0.16	0.09
SI	2	7.62	0.63	-0.002	0.11	0.01	0.58	0.68	90.76	0.18	0.07
SI	3	8.80	0.70	-0.002	0.10	0.01	0.65	0.74	92.32	0.19	0.06
SI	4	9.37	0.75	-0.002	0.10	0.01	0.71	0.79	92.32	0.19	0.06
SI	5	9.63	0.80	-0.002	0.11	0.01	0.77	0.83	92.06	0.18	0.07
SI	6	9.67	0.84	-0.001	0.11	0.01	0.81	0.87	93.10	0.17	0.08
SI	7	9.57	0.88	-0.001	0.12	0.02	0.85	0.90	93.10	0.17	0.08
SI	8	9.27	0.91	-0.001	0.14	0.02	0.88	0.93	93.36	0.16	0.09
SI	9	8.64	0.93	0.000	0.16	0.02	0.90	0.95	93.23	0.16	0.09
SI	10	7.44	0.94	0.000	0.19	0.04	0.92	0.96	93.62	0.16	0.09
SI	11	4.90	0.96	0.000	0.28	0.08	0.93	0.97	93.47	0.14	0.11
MIDC	1	5.15	0.57	-0.002	0.13	0.02	0.50	0.63	86.07	0.16	0.09
MIDC	2	7.62	0.63	-0.003	0.12	0.01	0.58	0.68	92.19	0.20	0.05
MIDC	3	8.80	0.70	-0.004	0.11	0.01	0.65	0.74	93.75	0.23	0.02
MIDC	4	9.37	0.75	-0.005	0.11	0.01	0.71	0.79	93.75	0.23	0.02
MIDC	5	9.63	0.80	-0.005	0.12	0.01	0.76	0.83	92.84	0.23	0.02
MIDC	6	9.67	0.84	-0.005	0.12	0.02	0.81	0.87	92.06	0.23	0.02
MIDC	7	9.57	0.87	-0.004	0.13	0.02	0.84	0.90	92.58	0.22	0.03
MIDC	8	9.27	0.90	-0.004	0.14	0.02	0.88	0.92	92.97	0.21	0.04
MIDC	9	8.64	0.93	-0.002	0.16	0.03	0.90	0.94	93.62	0.20	0.05
MIDC	10	7.44	0.94	-0.001	0.19	0.04	0.92	0.96	93.36	0.18	0.07
MIDC	11	4.90	0.96	-0.002	0.28	0.08	0.93	0.97	93.23	0.14	0.11

Scenario 3

Table 0.18 - Results for summary sensitivity for scenario 3

Method	Threshold	No. Studies*	Sensitivity	Bias	SE	MSE	LCI	UCI	Coverage	$\hat{\tau}$	$\hat{\tau}$ bias
NI	1	5.12	0.95	-0.006	0.84	0.70	0.79	0.98	96.16	0.26	0.24
NI	2	5.09	0.93	-0.008	0.76	0.57	0.77	0.98	93.63	0.27	0.23
NI	3	5.12	0.92	-0.007	0.66	0.44	0.76	0.97	95.76	0.27	0.23
NI	4	5.00	0.89	-0.010	0.60	0.35	0.74	0.95	92.56	0.26	0.24
NI	5	5.05	0.87	-0.009	0.54	0.29	0.71	0.94	92.97	0.28	0.22
NI	6	4.99	0.83	-0.008	0.48	0.24	0.67	0.92	94.42	0.28	0.22
NI	7	5.06	0.79	-0.007	0.46	0.21	0.62	0.89	92.72	0.32	0.18
NI	8	5.03	0.75	-0.009	0.42	0.18	0.58	0.86	92.19	0.31	0.19
NI	9	4.97	0.70	-0.007	0.40	0.16	0.52	0.82	92.58	0.31	0.19
NI	10	4.90	0.63	-0.009	0.39	0.15	0.46	0.78	93.24	0.32	0.18
NI	11	4.94	0.58	-0.002	0.37	0.14	0.41	0.73	91.11	0.29	0.21
SI	1	5.12	0.95	-0.006	0.84	0.70	0.79	0.98	96.16	0.26	0.24
SI	2	7.60	0.93	-0.012	0.53	0.28	0.83	0.97	92.85	0.24	0.26
SI	3	8.81	0.91	-0.014	0.43	0.18	0.82	0.96	91.26	0.26	0.24
SI	4	9.36	0.89	-0.017	0.37	0.14	0.79	0.94	89.14	0.27	0.23
SI	5	9.60	0.86	-0.017	0.34	0.11	0.76	0.92	92.32	0.29	0.21
SI	6	9.66	0.82	-0.017	0.31	0.10	0.72	0.89	90.86	0.32	0.18
SI	7	9.59	0.79	-0.017	0.30	0.09	0.67	0.86	91.92	0.34	0.16
SI	8	9.32	0.74	-0.018	0.28	0.08	0.62	0.83	91.92	0.34	0.16
SI	9	8.68	0.68	-0.017	0.28	0.08	0.56	0.79	92.32	0.35	0.15
SI	10	7.41	0.63	-0.014	0.30	0.09	0.49	0.75	92.05	0.34	0.16
SI	11	4.94	0.58	-0.002	0.37	0.14	0.41	0.73	91.11	0.29	0.21
MIDC	1	5.12	0.95	-0.006	0.84	0.70	0.79	0.98	96.16	0.26	0.24
MIDC	2	7.60	0.93	-0.006	0.57	0.32	0.83	0.97	96.29	0.28	0.22
MIDC	3	8.81	0.92	-0.007	0.46	0.21	0.82	0.96	94.57	0.31	0.19
MIDC	4	9.36	0.89	-0.009	0.40	0.16	0.80	0.94	94.44	0.32	0.18
MIDC	5	9.60	0.87	-0.009	0.36	0.13	0.77	0.92	93.64	0.33	0.17
MIDC	6	9.66	0.83	-0.010	0.33	0.11	0.73	0.90	94.17	0.35	0.15
MIDC	7	9.59	0.79	-0.010	0.31	0.09	0.68	0.87	94.44	0.36	0.14
MIDC	8	9.32	0.75	-0.010	0.29	0.08	0.63	0.83	94.30	0.36	0.14
MIDC	9	8.68	0.69	-0.009	0.29	0.08	0.57	0.79	94.44	0.36	0.14
MIDC	10	7.41	0.63	-0.007	0.30	0.09	0.49	0.75	93.51	0.35	0.15
MIDC	11	4.94	0.58	-0.002	0.37	0.14	0.41	0.73	90.99	0.29	0.21

Table 0.19 - Results for summary specificity for scenario 3

Method	Threshold	No. Studies*	Specificity	Bias	SE	MSE	LCI	UCI	Coverage	$\hat{\tau}$	$\hat{\tau}$ bias
NI	1	5.12	0.57	-0.001	0.21	0.04	0.47	0.66	83.84	0.39	0.11
NI	2	5.09	0.63	-0.002	0.21	0.04	0.54	0.72	83.80	0.38	0.12
NI	3	5.12	0.69	-0.004	0.21	0.04	0.60	0.77	83.82	0.37	0.13
NI	4	5.00	0.75	-0.004	0.21	0.05	0.67	0.82	83.67	0.36	0.14
NI	5	5.05	0.80	-0.003	0.22	0.05	0.72	0.86	86.47	0.38	0.12
NI	6	4.99	0.84	-0.003	0.23	0.05	0.77	0.89	84.86	0.35	0.15
NI	7	5.06	0.88	-0.002	0.24	0.06	0.82	0.92	86.62	0.35	0.15
NI	8	5.03	0.90	-0.002	0.25	0.06	0.85	0.94	85.83	0.33	0.17
NI	9	4.97	0.93	-0.002	0.27	0.07	0.88	0.95	85.83	0.32	0.18
NI	10	4.90	0.94	-0.002	0.29	0.09	0.90	0.96	88.33	0.30	0.20
NI	11	4.94	0.96	-0.002	0.35	0.12	0.92	0.97	90.58	0.45	0.05
SI	1	5.12	0.57	-0.001	0.21	0.04	0.47	0.66	83.84	0.39	0.11
SI	2	7.60	0.63	-0.002	0.18	0.03	0.55	0.71	87.81	0.43	0.07
SI	3	8.81	0.70	-0.003	0.17	0.03	0.62	0.76	88.21	0.44	0.06
SI	4	9.36	0.75	-0.003	0.17	0.03	0.69	0.81	88.61	0.44	0.06
SI	5	9.60	0.80	-0.003	0.17	0.03	0.74	0.85	89.01	0.44	0.06
SI	6	9.66	0.84	-0.002	0.17	0.03	0.79	0.88	89.01	0.43	0.07
SI	7	9.59	0.88	-0.002	0.18	0.03	0.83	0.91	88.34	0.42	0.08
SI	8	9.32	0.90	-0.002	0.19	0.04	0.87	0.93	90.20	0.41	0.09
SI	9	8.68	0.93	-0.002	0.21	0.04	0.89	0.95	88.34	0.39	0.11
SI	10	7.41	0.94	-0.002	0.24	0.06	0.91	0.96	90.07	0.36	0.14
SI	11	4.94	0.96	-0.002	0.35	0.12	0.92	0.97	90.58	0.45	0.05
MIDC	1	5.12	0.57	-0.001	0.21	0.04	0.47	0.66	83.84	0.38	0.12
MIDC	2	7.60	0.63	-0.002	0.18	0.03	0.55	0.71	88.87	0.43	0.07
MIDC	3	8.81	0.69	-0.005	0.17	0.03	0.62	0.76	89.27	0.45	0.05
MIDC	4	9.36	0.75	-0.006	0.17	0.03	0.68	0.81	90.07	0.45	0.05
MIDC	5	9.60	0.80	-0.006	0.17	0.03	0.74	0.85	90.60	0.45	0.05
MIDC	6	9.66	0.84	-0.006	0.18	0.03	0.79	0.88	89.93	0.46	0.04
MIDC	7	9.59	0.87	-0.005	0.19	0.03	0.83	0.91	89.93	0.45	0.05
MIDC	8	9.32	0.90	-0.005	0.19	0.04	0.86	0.93	90.46	0.44	0.06
MIDC	9	8.68	0.92	-0.004	0.21	0.04	0.89	0.95	89.93	0.42	0.08
MIDC	10	7.41	0.94	-0.003	0.24	0.06	0.91	0.96	90.73	0.37	0.13
MIDC	11	4.94	0.96	-0.003	0.35	0.12	0.92	0.97	90.46	0.44	0.06

APPENDIX E2: Missing not at random

Scenario 7

Table 0.20 - Results for summary sensitivity for scenario 7

Method	Threshold	No. Studies*	Sensitivity	Bias	SE	MSE	LCI	UCI	Coverage
NI	1	5.20	0.95	-0.005	0.67	0.46	0.84	0.98	96.66
NI	2	5.19	0.94	-0.005	0.61	0.37	0.83	0.98	96.27
NI	3	5.44	0.92	0.001	0.54	0.29	0.82	0.97	97.69
NI	4	6.26	0.91	0.012	0.46	0.22	0.82	0.96	96.53
NI	5	6.95	0.90	0.022	0.40	0.16	0.81	0.95	94.86
NI	6	7.30	0.87	0.023	0.34	0.11	0.77	0.92	93.06
NI	7	7.27	0.83	0.028	0.31	0.09	0.73	0.90	90.62
NI	8	6.96	0.79	0.033	0.29	0.08	0.68	0.86	88.95
NI	9	6.52	0.74	0.034	0.28	0.08	0.62	0.82	89.33
NI	10	6.06	0.67	0.029	0.27	0.08	0.55	0.77	90.73
NI	11	5.67	0.60	0.026	0.27	0.07	0.48	0.72	93.06
SI	1	5.20	0.95	-0.005	0.67	0.46	0.84	0.98	96.66
SI	2	7.69	0.93	-0.009	0.45	0.20	0.85	0.97	93.96
SI	3	8.97	0.92	-0.007	0.37	0.14	0.84	0.96	94.60
SI	4	9.61	0.90	-0.003	0.33	0.11	0.83	0.94	95.12
SI	5	9.82	0.87	-0.001	0.29	0.08	0.80	0.92	94.73
SI	6	9.87	0.84	-0.001	0.26	0.07	0.76	0.90	93.32
SI	7	9.83	0.80	0.000	0.24	0.06	0.72	0.87	93.06
SI	8	9.60	0.76	0.002	0.23	0.05	0.67	0.83	92.42
SI	9	9.13	0.70	0.003	0.22	0.05	0.61	0.78	93.83
SI	10	8.05	0.65	0.007	0.22	0.05	0.55	0.74	93.44
SI	11	5.67	0.60	0.026	0.27	0.07	0.48	0.72	93.06
MIDC	1	5.20	0.95	-0.005	0.67	0.46	0.84	0.98	96.66
MIDC	2	7.69	0.94	-0.004	0.47	0.22	0.86	0.97	97.30
MIDC	3	8.97	0.92	-0.003	0.38	0.15	0.85	0.96	96.66
MIDC	4	9.61	0.90	-0.001	0.33	0.11	0.83	0.94	96.14
MIDC	5	9.82	0.88	0.001	0.29	0.09	0.80	0.92	94.47
MIDC	6	9.87	0.84	0.001	0.26	0.07	0.76	0.90	94.09
MIDC	7	9.83	0.80	0.002	0.24	0.06	0.72	0.87	93.83
MIDC	8	9.60	0.76	0.004	0.23	0.05	0.67	0.83	93.70
MIDC	9	9.13	0.71	0.007	0.22	0.05	0.61	0.79	93.96
MIDC	10	8.05	0.65	0.011	0.22	0.05	0.55	0.74	93.83
MIDC	11	5.67	0.60	0.026	0.27	0.07	0.48	0.72	93.06

Table 0.21 - Results for summary specificity for scenario 7

Method	Threshold	No. Studies*	Specificity	Bias	SE	MSE	LCI	UCI	Coverage
NI	1	5.20	0.57	0.001	0.09	0.01	0.52	0.61	94.73
NI	2	5.19	0.64	0.000	0.10	0.01	0.59	0.68	94.86
NI	3	5.44	0.70	0.005	0.10	0.01	0.66	0.74	93.06
NI	4	6.26	0.76	0.007	0.10	0.01	0.73	0.79	92.93
NI	5	6.95	0.81	0.005	0.10	0.01	0.78	0.84	93.96
NI	6	7.30	0.85	0.002	0.10	0.01	0.82	0.87	94.60
NI	7	7.27	0.88	0.001	0.12	0.01	0.85	0.90	95.12
NI	8	6.96	0.91	0.002	0.13	0.02	0.88	0.93	94.99
NI	9	6.52	0.93	0.001	0.16	0.02	0.91	0.95	95.24
NI	10	6.06	0.95	0.001	0.19	0.03	0.92	0.96	95.50
NI	11	5.67	0.96	0.000	0.22	0.05	0.94	0.97	95.63
SI	1	5.20	0.57	0.001	0.09	0.01	0.52	0.61	94.73
SI	2	7.69	0.63	-0.001	0.07	0.01	0.60	0.67	95.12
SI	3	8.97	0.70	0.000	0.07	0.01	0.67	0.73	95.37
SI	4	9.61	0.76	0.000	0.07	0.01	0.73	0.78	95.37
SI	5	9.82	0.80	0.000	0.08	0.01	0.78	0.83	94.86
SI	6	9.87	0.85	0.000	0.09	0.01	0.82	0.87	95.63
SI	7	9.83	0.88	0.000	0.10	0.01	0.86	0.90	95.50
SI	8	9.60	0.91	0.000	0.11	0.01	0.89	0.92	95.76
SI	9	9.13	0.93	0.000	0.13	0.02	0.91	0.94	95.37
SI	10	8.05	0.94	0.000	0.16	0.02	0.93	0.96	95.12
SI	11	5.67	0.96	0.000	0.22	0.05	0.94	0.97	95.63
MIDC	1	5.20	0.57	0.001	0.09	0.01	0.52	0.61	94.73
MIDC	2	7.69	0.63	-0.001	0.07	0.01	0.60	0.67	94.34
MIDC	3	8.97	0.70	-0.001	0.07	0.01	0.67	0.73	93.70
MIDC	4	9.61	0.75	-0.002	0.07	0.01	0.73	0.78	93.32
MIDC	5	9.82	0.80	-0.002	0.08	0.01	0.78	0.83	93.44
MIDC	6	9.87	0.84	-0.002	0.09	0.01	0.82	0.86	94.99
MIDC	7	9.83	0.88	-0.002	0.10	0.01	0.86	0.90	94.60
MIDC	8	9.60	0.90	-0.002	0.11	0.01	0.88	0.92	94.34
MIDC	9	9.13	0.93	-0.002	0.13	0.02	0.91	0.94	94.73
MIDC	10	8.05	0.94	-0.001	0.15	0.02	0.93	0.96	94.86
MIDC	11	5.67	0.96	0.000	0.22	0.05	0.94	0.97	95.63

Scenario 8

Table 0.22 - Results for summary sensitivity for scenario 8

Method	Threshold	No. Studies*	Sensitivity	Bias	SE	MSE	LCI	UCI	Coverage	$\hat{\tau}$	$\hat{\tau}$ bias
NI	1	5.20	0.95	-0.002	0.86	0.75	0.80	0.99	97.69	0.27	-0.02
NI	2	5.28	0.94	0.001	0.74	0.55	0.80	0.98	97.05	0.22	0.03
NI	3	5.74	0.93	0.006	0.65	0.42	0.80	0.97	97.30	0.24	0.01
NI	4	6.28	0.92	0.015	0.57	0.33	0.79	0.96	97.56	0.28	-0.03
NI	5	6.84	0.90	0.022	0.47	0.22	0.78	0.95	97.05	0.24	0.01
NI	6	7.20	0.87	0.026	0.40	0.16	0.75	0.93	96.02	0.22	0.03
NI	7	7.16	0.83	0.031	0.36	0.13	0.71	0.91	93.58	0.22	0.03
NI	8	6.94	0.79	0.035	0.34	0.12	0.66	0.87	92.94	0.21	0.04
NI	9	6.59	0.74	0.035	0.33	0.11	0.60	0.84	92.04	0.22	0.03
NI	10	6.06	0.68	0.037	0.33	0.11	0.53	0.79	91.91	0.22	0.03
NI	11	5.72	0.61	0.031	0.32	0.10	0.46	0.74	92.17	0.22	0.03
SI	1	5.20	0.95	-0.002	0.86	0.75	0.80	0.99	97.69	0.27	-0.02
SI	2	7.73	0.94	-0.005	0.54	0.29	0.84	0.97	95.25	0.23	0.02
SI	3	9.00	0.92	-0.004	0.44	0.19	0.83	0.96	95.38	0.23	0.02
SI	4	9.54	0.90	-0.001	0.39	0.15	0.81	0.95	95.64	0.25	0.00
SI	5	9.78	0.87	0.000	0.34	0.12	0.78	0.93	96.28	0.24	0.01
SI	6	9.84	0.84	0.000	0.31	0.09	0.75	0.90	95.51	0.24	0.01
SI	7	9.79	0.80	0.001	0.28	0.08	0.70	0.87	96.28	0.23	0.02
SI	8	9.59	0.76	0.001	0.26	0.07	0.65	0.84	95.64	0.22	0.03
SI	9	9.09	0.71	0.003	0.25	0.06	0.59	0.79	94.61	0.21	0.04
SI	10	7.97	0.65	0.010	0.26	0.07	0.53	0.76	94.09	0.21	0.04
SI	11	5.72	0.61	0.031	0.32	0.10	0.46	0.74	92.17	0.22	0.03
MIDC	1	5.20	0.95	-0.003	0.86	0.74	0.80	0.99	97.56	0.27	-0.02
MIDC	2	7.73	0.94	0.000	0.57	0.32	0.84	0.98	97.56	0.24	0.01
MIDC	3	9.00	0.92	0.001	0.46	0.21	0.84	0.96	96.79	0.25	0.00
MIDC	4	9.54	0.90	0.002	0.40	0.16	0.82	0.95	97.30	0.27	-0.02
MIDC	5	9.78	0.88	0.002	0.35	0.12	0.79	0.93	96.92	0.24	0.01
MIDC	6	9.84	0.84	0.002	0.31	0.09	0.75	0.91	96.66	0.23	0.02
MIDC	7	9.79	0.81	0.004	0.28	0.08	0.71	0.88	96.28	0.22	0.03
MIDC	8	9.59	0.76	0.005	0.26	0.07	0.66	0.84	96.15	0.21	0.04
MIDC	9	9.09	0.71	0.007	0.25	0.06	0.60	0.80	95.51	0.21	0.04
MIDC	10	7.97	0.66	0.014	0.26	0.07	0.53	0.76	94.87	0.21	0.04
MIDC	11	5.72	0.61	0.031	0.32	0.10	0.46	0.74	92.17	0.22	0.03

Table 0.23 - Results for summary specificity for scenario 8

Method	Threshold	No. Studies*	Specificity	Bias	SE	MSE	LCI	UCI	Coverage	$\hat{\tau}$	$\hat{\tau}$ bias
NI	1	5.20	0.57	0.002	0.13	0.02	0.51	0.63	88.43	0.18	0.07
NI	2	5.28	0.64	0.007	0.14	0.02	0.58	0.70	90.24	0.19	0.06
NI	3	5.74	0.71	0.010	0.14	0.02	0.65	0.76	86.65	0.19	0.06
NI	4	6.28	0.77	0.012	0.13	0.02	0.72	0.81	88.45	0.18	0.07
NI	5	6.84	0.81	0.010	0.13	0.02	0.77	0.85	88.58	0.16	0.09
NI	6	7.20	0.85	0.007	0.14	0.02	0.81	0.88	90.50	0.17	0.08
NI	7	7.16	0.88	0.004	0.15	0.02	0.85	0.91	91.66	0.17	0.08
NI	8	6.94	0.91	0.003	0.16	0.03	0.88	0.93	93.32	0.16	0.09
NI	9	6.59	0.93	0.002	0.19	0.03	0.90	0.95	92.94	0.16	0.09
NI	10	6.06	0.95	0.001	0.22	0.05	0.92	0.96	93.84	0.15	0.10
NI	11	5.72	0.96	0.001	0.26	0.07	0.93	0.97	95.12	0.15	0.10
SI	1	5.20	0.57	0.002	0.13	0.02	0.51	0.63	88.43	0.18	0.07
SI	2	7.73	0.64	0.002	0.11	0.01	0.59	0.69	91.01	0.20	0.05
SI	3	9.00	0.70	0.001	0.11	0.01	0.65	0.74	90.76	0.20	0.05
SI	4	9.54	0.76	0.001	0.11	0.01	0.72	0.79	91.78	0.20	0.05
SI	5	9.78	0.81	0.001	0.11	0.01	0.77	0.84	92.81	0.19	0.06
SI	6	9.84	0.85	0.001	0.12	0.01	0.81	0.87	93.58	0.19	0.06
SI	7	9.79	0.88	0.000	0.13	0.02	0.85	0.90	94.09	0.19	0.06
SI	8	9.59	0.91	0.000	0.14	0.02	0.88	0.93	93.58	0.18	0.07
SI	9	9.09	0.93	0.000	0.15	0.02	0.91	0.95	94.99	0.17	0.08
SI	10	7.97	0.95	0.000	0.19	0.03	0.92	0.96	94.74	0.16	0.09
SI	11	5.72	0.96	0.001	0.26	0.07	0.93	0.97	95.12	0.15	0.10
MIDC	1	5.20	0.57	0.002	0.13	0.02	0.51	0.63	88.32	0.17	0.08
MIDC	2	7.73	0.64	0.002	0.12	0.01	0.58	0.69	92.30	0.22	0.03
MIDC	3	9.00	0.70	0.001	0.11	0.01	0.65	0.74	92.43	0.23	0.02
MIDC	4	9.54	0.76	0.000	0.11	0.01	0.71	0.79	93.32	0.23	0.02
MIDC	5	9.78	0.80	0.000	0.12	0.01	0.77	0.84	93.45	0.22	0.03
MIDC	6	9.84	0.84	-0.001	0.12	0.01	0.81	0.87	93.84	0.22	0.03
MIDC	7	9.79	0.88	-0.001	0.13	0.02	0.85	0.90	94.61	0.22	0.03
MIDC	8	9.59	0.91	-0.001	0.14	0.02	0.88	0.93	94.61	0.21	0.04
MIDC	9	9.09	0.93	-0.001	0.16	0.03	0.90	0.95	95.25	0.20	0.05
MIDC	10	7.97	0.94	0.000	0.19	0.03	0.92	0.96	95.12	0.18	0.07
MIDC	11	5.72	0.96	0.001	0.26	0.07	0.93	0.97	95.12	0.15	0.10

Scenario 9

Table 0.24 - Results for summary sensitivity for scenario 9

Method	Threshold	No. Studies*	Sensitivity	Bias	SE	MSE	LCI	UCI	Coverage	$\hat{\tau}$	$\hat{\tau}$ bias
NI	1	5.32	0.95	-0.005	0.84	0.71	0.79	0.98	95.07	0.28	0.22
NI	2	5.53	0.94	-0.003	0.76	0.57	0.79	0.98	95.20	0.32	0.18
NI	3	5.96	0.93	0.002	0.66	0.44	0.79	0.97	96.37	0.32	0.18
NI	4	6.35	0.91	0.007	0.57	0.32	0.78	0.96	96.37	0.33	0.17
NI	5	6.71	0.89	0.014	0.49	0.24	0.76	0.95	95.73	0.34	0.16
NI	6	6.95	0.86	0.023	0.45	0.20	0.73	0.93	95.08	0.35	0.15
NI	7	7.00	0.83	0.026	0.40	0.16	0.69	0.91	94.43	0.36	0.14
NI	8	6.86	0.79	0.035	0.38	0.15	0.65	0.88	93.52	0.37	0.13
NI	9	6.62	0.74	0.038	0.37	0.14	0.59	0.85	90.03	0.38	0.12
NI	10	6.24	0.68	0.038	0.36	0.13	0.52	0.80	88.47	0.39	0.11
NI	11	5.85	0.62	0.038	0.38	0.14	0.44	0.76	90.28	0.42	0.08
SI	1	5.32	0.95	-0.005	0.84	0.71	0.79	0.98	95.07	0.28	0.22
SI	2	7.86	0.93	-0.009	0.55	0.30	0.83	0.97	92.75	0.29	0.21
SI	3	9.01	0.91	-0.010	0.45	0.20	0.82	0.96	91.19	0.32	0.18
SI	4	9.54	0.89	-0.010	0.40	0.16	0.80	0.94	92.62	0.35	0.15
SI	5	9.77	0.87	-0.009	0.36	0.13	0.77	0.92	91.97	0.35	0.15
SI	6	9.84	0.83	-0.008	0.33	0.11	0.73	0.90	92.23	0.38	0.12
SI	7	9.78	0.80	-0.007	0.31	0.10	0.68	0.87	93.01	0.40	0.10
SI	8	9.59	0.75	-0.005	0.30	0.09	0.63	0.84	91.71	0.40	0.10
SI	9	9.18	0.70	-0.003	0.29	0.08	0.57	0.80	92.75	0.40	0.10
SI	10	8.14	0.65	0.006	0.30	0.09	0.51	0.76	92.36	0.41	0.09
SI	11	5.85	0.62	0.038	0.38	0.14	0.44	0.76	90.28	0.42	0.08
MIDC	1	5.32	0.95	-0.006	0.84	0.71	0.79	0.98	94.95	0.28	0.22
MIDC	2	7.86	0.94	-0.004	0.58	0.34	0.83	0.97	96.11	0.32	0.18
MIDC	3	9.01	0.92	-0.005	0.48	0.23	0.82	0.96	95.21	0.35	0.15
MIDC	4	9.54	0.90	-0.005	0.41	0.17	0.80	0.95	94.82	0.36	0.14
MIDC	5	9.77	0.87	-0.006	0.36	0.13	0.77	0.93	93.78	0.37	0.13
MIDC	6	9.84	0.84	-0.006	0.33	0.11	0.73	0.90	93.78	0.38	0.12
MIDC	7	9.78	0.80	-0.004	0.31	0.10	0.69	0.88	94.43	0.39	0.11
MIDC	8	9.59	0.75	-0.001	0.30	0.09	0.64	0.84	93.91	0.39	0.11
MIDC	9	9.18	0.70	0.003	0.29	0.08	0.58	0.80	94.30	0.39	0.11
MIDC	10	8.14	0.65	0.011	0.30	0.09	0.51	0.77	93.26	0.40	0.10
MIDC	11	5.85	0.62	0.038	0.38	0.14	0.44	0.76	90.28	0.42	0.08

Table 0.25 - Results for summary specificity for scenario 9

Method	Threshold	No. Studies*	Specificity	Bias	SE	MSE	LCI	UCI	Coverage	$\hat{\tau}$	$\hat{\tau}$ bias
NI	1	5.32	0.57	0.007	0.22	0.05	0.47	0.67	84.05	0.42	0.08
NI	2	5.53	0.65	0.017	0.22	0.05	0.55	0.74	85.08	0.43	0.07
NI	3	5.96	0.72	0.018	0.21	0.05	0.63	0.79	85.49	0.43	0.07
NI	4	6.35	0.77	0.018	0.20	0.04	0.70	0.83	83.94	0.41	0.09
NI	5	6.71	0.82	0.013	0.20	0.04	0.75	0.87	85.23	0.40	0.10
NI	6	6.95	0.85	0.009	0.20	0.04	0.80	0.90	87.05	0.39	0.11
NI	7	7.00	0.88	0.005	0.21	0.04	0.83	0.92	89.25	0.39	0.11
NI	8	6.86	0.91	0.003	0.22	0.05	0.87	0.94	88.34	0.36	0.14
NI	9	6.62	0.93	0.001	0.24	0.06	0.89	0.95	89.64	0.36	0.14
NI	10	6.24	0.94	0.000	0.26	0.07	0.91	0.97	91.58	0.34	0.16
NI	11	5.85	0.96	0.000	0.30	0.09	0.93	0.97	91.19	0.32	0.18
SI	1	5.32	0.57	0.007	0.22	0.05	0.47	0.67	84.05	0.42	0.08
SI	2	7.86	0.64	0.003	0.18	0.03	0.55	0.72	89.25	0.45	0.05
SI	3	9.01	0.70	-0.001	0.17	0.03	0.62	0.76	90.41	0.45	0.05
SI	4	9.54	0.75	-0.002	0.17	0.03	0.69	0.81	90.54	0.45	0.05
SI	5	9.77	0.80	-0.002	0.17	0.03	0.74	0.85	89.77	0.44	0.06
SI	6	9.84	0.84	-0.002	0.17	0.03	0.79	0.88	90.80	0.44	0.06
SI	7	9.78	0.88	-0.002	0.18	0.03	0.83	0.91	90.16	0.43	0.07
SI	8	9.59	0.90	-0.002	0.19	0.04	0.87	0.93	89.12	0.42	0.08
SI	9	9.18	0.93	-0.001	0.20	0.04	0.89	0.95	89.12	0.40	0.10
SI	10	8.14	0.94	-0.001	0.23	0.05	0.92	0.96	89.12	0.37	0.13
SI	11	5.85	0.96	0.000	0.30	0.09	0.93	0.97	91.19	0.32	0.18
MIDC	1	5.32	0.57	0.006	0.22	0.05	0.47	0.67	83.94	0.41	0.09
MIDC	2	7.86	0.64	0.003	0.19	0.03	0.55	0.72	89.64	0.45	0.05
MIDC	3	9.01	0.70	-0.001	0.18	0.03	0.62	0.77	90.54	0.46	0.04
MIDC	4	9.54	0.75	-0.003	0.17	0.03	0.68	0.81	91.32	0.46	0.04
MIDC	5	9.77	0.80	-0.003	0.18	0.03	0.74	0.85	90.54	0.46	0.04
MIDC	6	9.84	0.84	-0.004	0.18	0.03	0.79	0.88	90.16	0.46	0.04
MIDC	7	9.78	0.88	-0.004	0.18	0.03	0.83	0.91	90.54	0.45	0.05
MIDC	8	9.59	0.90	-0.003	0.19	0.04	0.86	0.93	90.28	0.44	0.06
MIDC	9	9.18	0.93	-0.002	0.21	0.04	0.89	0.95	90.03	0.42	0.08
MIDC	10	8.14	0.94	-0.001	0.23	0.05	0.91	0.96	89.51	0.39	0.11
MIDC	11	5.85	0.96	0.000	0.30	0.09	0.93	0.97	91.19	0.32	0.18

APPENDIX E3: Unequal threshold spacing

Scenario 12

Table 0.26 - Results for summary sensitivity for scenario 12

Method	Threshold	No. Studies*	Sensitivity	Bias	SE	MSE	LCI	UCI	Coverage	$\hat{\tau}$	$\hat{\tau}$ bias
NI	1	5.03	0.95	-0.011	0.86	0.73	0.79	0.99	93.58	0.22	0.28
NI	2	5.03	0.95	-0.013	0.78	0.61	0.80	0.98	93.23	0.19	0.31
NI	3	5.09	0.94	-0.012	0.80	0.64	0.79	0.98	92.69	0.26	0.24
NI	4	4.97	0.93	-0.013	0.73	0.53	0.79	0.98	93.93	0.27	0.23
NI	5	5.01	0.93	-0.010	0.70	0.49	0.77	0.97	95.01	0.26	0.24
NI	6	5.03	0.91	-0.010	0.66	0.43	0.76	0.97	94.83	0.26	0.24
NI	7	5.09	0.90	-0.009	0.61	0.38	0.75	0.96	93.40	0.28	0.22
NI	8	5.00	0.88	-0.011	0.56	0.32	0.72	0.95	93.76	0.30	0.20
NI	9	5.07	0.85	-0.009	0.51	0.26	0.70	0.93	91.62	0.27	0.23
NI	10	5.09	0.82	-0.014	0.46	0.21	0.66	0.91	93.40	0.25	0.25
NI	11	4.93	0.78	-0.008	0.47	0.22	0.61	0.88	92.86	0.28	0.22
SI	1	5.03	0.95	-0.011	0.86	0.73	0.79	0.99	93.58	0.22	0.28
SI	2	7.50	0.94	-0.016	0.57	0.33	0.85	0.98	88.24	0.18	0.32
SI	3	8.72	0.94	-0.018	0.49	0.24	0.85	0.97	86.10	0.20	0.30
SI	4	9.30	0.93	-0.019	0.45	0.20	0.85	0.97	85.92	0.22	0.28
SI	5	9.57	0.92	-0.018	0.42	0.18	0.84	0.96	86.63	0.23	0.27
SI	6	9.66	0.91	-0.018	0.40	0.16	0.82	0.95	90.37	0.26	0.24
SI	7	9.61	0.89	-0.019	0.37	0.14	0.80	0.94	90.37	0.29	0.21
SI	8	9.32	0.87	-0.019	0.35	0.12	0.77	0.93	91.09	0.30	0.20
SI	9	8.75	0.84	-0.018	0.34	0.12	0.74	0.91	91.98	0.29	0.21
SI	10	7.50	0.81	-0.015	0.36	0.13	0.69	0.89	92.51	0.29	0.21
SI	11	4.93	0.78	-0.008	0.47	0.22	0.61	0.88	92.86	0.28	0.22
MIDC	1	5.03	0.95	-0.011	0.86	0.73	0.79	0.99	93.58	0.22	0.28
MIDC	2	7.50	0.95	-0.011	0.62	0.39	0.85	0.98	91.98	0.22	0.28
MIDC	3	8.72	0.94	-0.011	0.54	0.29	0.86	0.98	91.98	0.26	0.24
MIDC	4	9.30	0.94	-0.011	0.50	0.24	0.85	0.97	93.05	0.27	0.23
MIDC	5	9.57	0.93	-0.011	0.46	0.21	0.84	0.97	93.58	0.29	0.21
MIDC	6	9.66	0.91	-0.011	0.43	0.18	0.83	0.96	94.65	0.31	0.19
MIDC	7	9.61	0.90	-0.012	0.40	0.16	0.80	0.95	94.12	0.33	0.17
MIDC	8	9.32	0.88	-0.013	0.37	0.14	0.78	0.93	94.12	0.33	0.17
MIDC	9	8.75	0.85	-0.012	0.35	0.12	0.74	0.92	93.58	0.31	0.19
MIDC	10	7.50	0.82	-0.010	0.36	0.13	0.70	0.90	93.58	0.30	0.20
MIDC	11	4.93	0.78	-0.010	0.47	0.22	0.61	0.88	92.69	0.28	0.22

Table 0.27 - Results for summary specificity for scenario 12

Method	Threshold	No. Studies*	Specificity	Bias	SE	MSE	LCI	UCI	Coverage	$\hat{\tau}$	$\hat{\tau}$ bias
NI	1	5.03	0.51	0.001	0.21	0.04	0.41	0.61	86.45	0.38	0.12
NI	2	5.03	0.54	-0.001	0.21	0.04	0.44	0.63	82.17	0.38	0.12
NI	3	5.09	0.57	-0.002	0.21	0.04	0.47	0.66	84.85	0.39	0.11
NI	4	4.97	0.61	0.003	0.21	0.04	0.51	0.70	84.82	0.37	0.13
NI	5	5.01	0.65	-0.004	0.21	0.05	0.55	0.73	83.24	0.38	0.12
NI	6	5.03	0.69	-0.003	0.21	0.05	0.60	0.77	83.24	0.38	0.12
NI	7	5.09	0.74	0.000	0.22	0.05	0.65	0.81	86.45	0.39	0.11
NI	8	5.00	0.78	-0.003	0.22	0.05	0.70	0.84	85.74	0.37	0.13
NI	9	5.07	0.82	-0.003	0.23	0.05	0.74	0.87	85.20	0.38	0.12
NI	10	5.09	0.86	-0.002	0.24	0.06	0.79	0.90	84.67	0.35	0.15
NI	11	4.93	0.89	-0.002	0.25	0.06	0.83	0.92	86.43	0.34	0.16
SI	1	5.03	0.51	0.001	0.21	0.04	0.41	0.61	86.45	0.38	0.12
SI	2	7.50	0.54	0.001	0.18	0.03	0.45	0.62	88.24	0.43	0.07
SI	3	8.72	0.57	0.000	0.17	0.03	0.49	0.65	91.80	0.45	0.05
SI	4	9.30	0.61	0.000	0.17	0.03	0.53	0.68	90.73	0.45	0.05
SI	5	9.57	0.65	0.000	0.16	0.03	0.58	0.72	91.44	0.45	0.05
SI	6	9.66	0.69	0.000	0.17	0.03	0.62	0.76	91.09	0.45	0.05
SI	7	9.61	0.74	-0.001	0.17	0.03	0.67	0.80	91.62	0.44	0.06
SI	8	9.32	0.78	-0.001	0.17	0.03	0.72	0.83	91.62	0.44	0.06
SI	9	8.75	0.82	-0.001	0.18	0.03	0.76	0.87	91.27	0.43	0.07
SI	10	7.50	0.86	-0.001	0.20	0.04	0.80	0.90	88.24	0.40	0.10
SI	11	4.93	0.89	-0.002	0.25	0.06	0.83	0.92	86.43	0.34	0.16
MIDC	1	5.03	0.51	0.001	0.21	0.04	0.41	0.61	86.45	0.38	0.12
MIDC	2	7.50	0.54	0.003	0.18	0.03	0.45	0.63	87.88	0.43	0.07
MIDC	3	8.72	0.57	0.002	0.17	0.03	0.49	0.65	91.62	0.45	0.05
MIDC	4	9.30	0.61	0.002	0.17	0.03	0.53	0.68	91.09	0.45	0.05
MIDC	5	9.57	0.65	0.001	0.17	0.03	0.58	0.72	91.27	0.45	0.05
MIDC	6	9.66	0.69	0.000	0.17	0.03	0.62	0.76	92.51	0.45	0.05
MIDC	7	9.61	0.74	-0.002	0.17	0.03	0.67	0.80	93.94	0.45	0.05
MIDC	8	9.32	0.78	-0.003	0.17	0.03	0.72	0.83	92.34	0.45	0.05
MIDC	9	8.75	0.82	-0.003	0.18	0.03	0.76	0.87	91.44	0.44	0.06
MIDC	10	7.50	0.86	-0.002	0.20	0.04	0.80	0.90	89.66	0.41	0.09
MIDC	11	4.93	0.88	-0.003	0.24	0.06	0.83	0.92	86.27	0.34	0.16

APPENDIX E4: Extreme threshold spacing

Scenario 15

Table 0.28 - Results for summary sensitivity for scenario 15

Method	Threshold	No. Studies*	Sensitivity	Bias	SE	MSE	LCI	UCI	Coverage	$\hat{\tau}$	$\hat{\tau}$ bias
NI	1	5.11	0.95	-0.009	0.85	0.72	0.81	0.99	94.92	0.24	0.26
NI	2	4.97	0.95	-0.011	0.80	0.64	0.80	0.98	93.45	0.23	0.27
NI	3	5.14	0.94	-0.009	0.81	0.65	0.79	0.98	94.60	0.25	0.25
NI	4	5.00	0.93	-0.008	0.73	0.54	0.77	0.97	93.78	0.29	0.21
NI	5	5.13	0.92	-0.009	0.68	0.46	0.76	0.97	96.40	0.32	0.18
NI	6	4.95	0.73	-0.010	0.42	0.18	0.55	0.84	92.79	0.31	0.19
NI	7	5.00	0.71	-0.010	0.39	0.15	0.54	0.83	91.64	0.28	0.22
NI	8	4.94	0.68	-0.013	0.40	0.16	0.50	0.81	90.66	0.30	0.20
NI	9	4.99	0.66	-0.012	0.39	0.15	0.48	0.79	92.62	0.32	0.18
NI	10	4.91	0.62	-0.013	0.37	0.14	0.45	0.76	90.83	0.28	0.22
NI	11	4.97	0.59	-0.008	0.37	0.14	0.42	0.73	91.49	0.31	0.19
SI	1	5.11	0.95	-0.009	0.85	0.72	0.81	0.99	94.92	0.24	0.26
SI	2	7.50	0.94	-0.014	0.59	0.35	0.85	0.98	89.85	0.23	0.27
SI	3	8.76	0.93	-0.019	0.49	0.24	0.85	0.97	83.80	0.23	0.27
SI	4	9.29	0.91	-0.023	0.42	0.18	0.82	0.96	82.98	0.28	0.22
SI	5	9.56	0.88	-0.048	0.37	0.13	0.78	0.93	63.01	0.36	0.14
SI	6	9.65	0.77	0.029	0.29	0.08	0.65	0.85	92.31	0.36	0.14
SI	7	9.57	0.72	0.000	0.27	0.07	0.60	0.81	92.80	0.35	0.15
SI	8	9.32	0.68	-0.011	0.27	0.07	0.56	0.78	92.80	0.36	0.14
SI	9	8.74	0.65	-0.016	0.28	0.08	0.52	0.76	91.98	0.36	0.14
SI	10	7.47	0.62	-0.016	0.29	0.09	0.48	0.74	90.67	0.34	0.16
SI	11	4.97	0.59	-0.008	0.37	0.14	0.42	0.73	91.49	0.31	0.19
MIDC	1	5.11	0.95	-0.011	0.85	0.72	0.81	0.99	94.76	0.23	0.27
MIDC	2	7.50	0.95	-0.010	0.62	0.39	0.85	0.98	94.44	0.24	0.26
MIDC	3	8.76	0.94	-0.013	0.54	0.29	0.85	0.97	90.02	0.30	0.20
MIDC	4	9.29	0.92	-0.017	0.47	0.22	0.82	0.96	89.85	0.41	0.09
MIDC	5	9.56	0.88	-0.042	0.41	0.17	0.78	0.94	73.81	0.49	0.01
MIDC	6	9.65	0.77	0.034	0.31	0.10	0.65	0.86	93.45	0.43	0.07
MIDC	7	9.57	0.72	0.006	0.28	0.08	0.60	0.81	94.93	0.37	0.13
MIDC	8	9.32	0.69	-0.005	0.27	0.07	0.57	0.79	93.78	0.36	0.14
MIDC	9	8.74	0.66	-0.009	0.27	0.08	0.53	0.76	93.13	0.35	0.15
MIDC	10	7.47	0.62	-0.011	0.29	0.09	0.49	0.74	91.16	0.34	0.16
MIDC	11	4.97	0.59	-0.008	0.37	0.14	0.42	0.73	91.49	0.31	0.19

Table 0.29 - Results for summary specificity for scenario 15

Method	Threshold	No. Studies*	Specificity	Bias	SE	MSE	LCI	UCI	Coverage	$\hat{\tau}$	$\hat{\tau}$ bias
NI	1	5.11	0.50	-0.005	0.20	0.04	0.41	0.60	82.62	0.37	0.13
NI	2	4.97	0.54	0.000	0.21	0.05	0.44	0.64	85.92	0.39	0.11
NI	3	5.14	0.58	-0.001	0.20	0.04	0.48	0.67	81.67	0.37	0.13
NI	4	5.00	0.66	0.000	0.21	0.04	0.57	0.74	81.18	0.36	0.14
NI	5	5.13	0.70	-0.001	0.21	0.04	0.60	0.77	83.80	0.37	0.13
NI	6	4.95	0.91	-0.001	0.26	0.07	0.86	0.94	85.25	0.32	0.18
NI	7	5.00	0.92	-0.003	0.27	0.07	0.87	0.95	87.21	0.34	0.16
NI	8	4.94	0.93	-0.004	0.27	0.07	0.88	0.95	87.87	0.32	0.18
NI	9	4.99	0.94	-0.003	0.28	0.08	0.90	0.96	86.89	0.30	0.20
NI	10	4.91	0.94	-0.003	0.30	0.09	0.90	0.97	89.20	0.32	0.18
NI	11	4.97	0.95	-0.002	0.31	0.10	0.92	0.97	89.03	0.31	0.19
SI	1	5.11	0.50	-0.005	0.20	0.04	0.41	0.60	82.62	0.37	0.13
SI	2	7.50	0.54	0.003	0.18	0.03	0.46	0.63	88.05	0.42	0.08
SI	3	8.76	0.59	0.014	0.17	0.03	0.51	0.67	86.91	0.45	0.05
SI	4	9.29	0.67	0.010	0.17	0.03	0.59	0.74	88.22	0.47	0.03
SI	5	9.56	0.75	0.054	0.20	0.04	0.67	0.81	65.79	0.54	-0.04
SI	6	9.65	0.88	-0.036	0.22	0.05	0.83	0.92	57.94	0.57	-0.07
SI	7	9.57	0.91	-0.012	0.20	0.04	0.87	0.94	82.32	0.47	0.03
SI	8	9.32	0.93	-0.005	0.20	0.04	0.89	0.95	88.05	0.42	0.08
SI	9	8.74	0.94	-0.002	0.21	0.05	0.91	0.96	88.05	0.39	0.11
SI	10	7.47	0.95	-0.002	0.24	0.06	0.92	0.96	91.65	0.37	0.13
SI	11	4.97	0.95	-0.002	0.31	0.10	0.92	0.97	89.03	0.31	0.19
MIDC	1	5.11	0.50	-0.006	0.20	0.04	0.41	0.60	82.49	0.36	0.14
MIDC	2	7.50	0.54	0.003	0.18	0.03	0.46	0.63	89.20	0.43	0.07
MIDC	3	8.76	0.59	0.011	0.17	0.03	0.51	0.67	89.53	0.45	0.05
MIDC	4	9.29	0.66	0.001	0.18	0.03	0.58	0.73	91.16	0.47	0.03
MIDC	5	9.56	0.73	0.038	0.20	0.04	0.65	0.80	81.34	0.54	-0.04
MIDC	6	9.65	0.87	-0.045	0.25	0.06	0.80	0.92	56.63	0.69	-0.19
MIDC	7	9.57	0.91	-0.016	0.23	0.05	0.86	0.94	81.51	0.56	-0.06
MIDC	8	9.32	0.92	-0.007	0.21	0.05	0.89	0.95	88.38	0.46	0.04
MIDC	9	8.74	0.94	-0.003	0.22	0.05	0.90	0.96	89.85	0.41	0.09
MIDC	10	7.47	0.94	-0.002	0.24	0.06	0.91	0.96	91.98	0.38	0.12
MIDC	11	4.97	0.95	-0.002	0.31	0.10	0.92	0.97	89.03	0.31	0.19

APPENDIX E5: Extensions

Fewer studies - Scenario 1

Table 0.30 - Results for summary sensitivity - scenario 1 with 5 studies

Method	Threshold	No. Studies*	Sensitivity	Bias	SE	MSE	LCI	UCI	Coverage
NI	1	2.65	0.93	-0.023	0.86	0.74	0.73	0.98	0.94
NI	2	2.67	0.92	-0.017	0.81	0.66	0.73	0.98	0.95
NI	3	2.53	0.90	-0.025	0.77	0.59	0.69	0.97	0.93
NI	4	2.60	0.89	-0.016	0.69	0.48	0.69	0.96	0.95
NI	5	2.62	0.86	-0.015	0.65	0.42	0.65	0.95	0.96
NI	6	2.59	0.83	-0.008	0.58	0.34	0.63	0.93	0.96
NI	7	2.55	0.79	-0.010	0.55	0.30	0.59	0.90	0.97
NI	8	2.63	0.75	-0.011	0.50	0.25	0.54	0.87	0.96
NI	9	2.54	0.69	-0.008	0.49	0.24	0.49	0.84	0.94
NI	10	2.57	0.63	-0.010	0.45	0.20	0.43	0.79	0.95
NI	11	2.56	0.58	0.005	0.44	0.19	0.39	0.75	0.98
SI	1	2.65	0.93	-0.023	0.86	0.74	0.73	0.98	0.94
SI	2	3.87	0.92	-0.018	0.62	0.38	0.79	0.97	0.94
SI	3	4.40	0.90	-0.019	0.51	0.26	0.78	0.96	0.92
SI	4	4.70	0.88	-0.017	0.45	0.20	0.77	0.95	0.94
SI	5	4.81	0.86	-0.018	0.40	0.16	0.74	0.93	0.93
SI	6	4.86	0.82	-0.017	0.36	0.13	0.70	0.90	0.94
SI	7	4.84	0.79	-0.016	0.34	0.11	0.66	0.87	0.95
SI	8	4.74	0.74	-0.017	0.32	0.10	0.61	0.84	0.95
SI	9	4.40	0.69	-0.015	0.31	0.10	0.55	0.80	0.95
SI	10	3.80	0.63	-0.013	0.33	0.11	0.48	0.76	0.96
SI	11	2.56	0.58	0.005	0.44	0.19	0.39	0.75	0.98
MIDC	1	2.65	0.93	-0.023	0.86	0.74	0.73	0.98	0.94
MIDC	2	3.87	0.93	-0.014	0.63	0.40	0.80	0.97	0.95
MIDC	3	4.40	0.91	-0.014	0.53	0.28	0.79	0.96	0.94
MIDC	4	4.70	0.89	-0.014	0.45	0.21	0.77	0.95	0.96
MIDC	5	4.81	0.86	-0.013	0.40	0.16	0.74	0.93	0.96
MIDC	6	4.86	0.83	-0.014	0.36	0.13	0.71	0.91	0.96
MIDC	7	4.84	0.79	-0.013	0.34	0.11	0.66	0.88	0.97
MIDC	8	4.74	0.74	-0.012	0.32	0.10	0.61	0.84	0.95
MIDC	9	4.40	0.69	-0.008	0.31	0.10	0.55	0.80	0.95
MIDC	10	3.80	0.63	-0.008	0.33	0.11	0.48	0.76	0.96
MIDC	11	2.56	0.58	0.005	0.44	0.19	0.39	0.75	0.98

Table 0.31 - Results for summary specificity - scenario 1 with 5 studies

Method	Threshold	No. Studies*	Specificity	Bias	SE	MSE	LCI	UCI	Coverage
NI	1	2.65	0.57	0.001	0.14	0.02	0.50	0.63	0.96
NI	2	2.67	0.63	-0.001	0.14	0.02	0.57	0.69	0.96
NI	3	2.53	0.70	0.002	0.15	0.02	0.64	0.76	0.95
NI	4	2.60	0.76	0.000	0.16	0.02	0.69	0.81	0.95
NI	5	2.62	0.80	-0.002	0.17	0.03	0.74	0.85	0.97
NI	6	2.59	0.84	-0.002	0.19	0.03	0.79	0.88	0.94
NI	7	2.55	0.88	-0.001	0.21	0.04	0.83	0.91	0.94
NI	8	2.63	0.90	-0.001	0.24	0.06	0.86	0.94	0.96
NI	9	2.54	0.93	-0.002	0.27	0.07	0.88	0.95	0.94
NI	10	2.57	0.94	0.000	0.31	0.09	0.90	0.97	0.96
NI	11	2.56	0.96	-0.001	0.35	0.13	0.92	0.98	0.95
SI	1	2.65	0.57	0.001	0.14	0.02	0.50	0.63	0.96
SI	2	3.87	0.64	-0.001	0.11	0.01	0.59	0.68	0.94
SI	3	4.40	0.70	0.000	0.10	0.01	0.65	0.74	0.95
SI	4	4.70	0.75	-0.001	0.11	0.01	0.71	0.79	0.93
SI	5	4.81	0.80	-0.001	0.11	0.01	0.77	0.84	0.96
SI	6	4.86	0.84	-0.001	0.12	0.02	0.81	0.87	0.95
SI	7	4.84	0.88	-0.001	0.14	0.02	0.85	0.90	0.94
SI	8	4.74	0.91	-0.001	0.16	0.02	0.88	0.93	0.95
SI	9	4.40	0.93	0.000	0.19	0.03	0.90	0.95	0.95
SI	10	3.80	0.94	0.000	0.23	0.06	0.92	0.96	0.95
SI	11	2.56	0.96	-0.001	0.35	0.13	0.92	0.98	0.95
MIDC	1	2.65	0.57	0.001	0.14	0.02	0.50	0.63	0.96
MIDC	2	3.87	0.64	-0.001	0.11	0.01	0.59	0.68	0.95
MIDC	3	4.40	0.70	-0.003	0.10	0.01	0.65	0.74	0.93
MIDC	4	4.70	0.75	-0.005	0.11	0.01	0.71	0.79	0.91
MIDC	5	4.81	0.80	-0.006	0.11	0.01	0.76	0.83	0.92
MIDC	6	4.86	0.84	-0.006	0.12	0.01	0.81	0.87	0.92
MIDC	7	4.84	0.87	-0.005	0.14	0.02	0.84	0.90	0.91
MIDC	8	4.74	0.90	-0.005	0.15	0.02	0.87	0.93	0.91
MIDC	9	4.40	0.92	-0.003	0.18	0.03	0.90	0.95	0.94
MIDC	10	3.80	0.94	-0.001	0.23	0.05	0.91	0.96	0.95
MIDC	11	2.56	0.96	-0.001	0.35	0.13	0.92	0.98	0.95

Fewer studies - Scenario 2

Table 0.32 - Results for summary sensitivity - scenario 2 with 5 studies

Method	Threshold	No. Studies*	Sensitivity	Bias	SE	MSE	LCI	UCI	Coverage	Tau	Tau bias
NI	1	2.78	0.94	-0.017	0.89	0.80	0.75	0.98	0.85	0.46	-0.21
NI	2	2.85	0.93	-0.010	0.81	0.65	0.75	0.98	0.93	0.11	0.14
NI	3	3.00	0.91	-0.013	0.72	0.51	0.73	0.97	0.93	0.13	0.12
NI	4	3.07	0.89	-0.012	0.66	0.43	0.71	0.96	0.90	0.10	0.15
NI	5	2.93	0.85	-0.021	0.62	0.38	0.65	0.94	0.92	0.17	0.08
NI	6	3.08	0.83	-0.009	0.52	0.27	0.65	0.92	0.95	0.11	0.14
NI	7	2.98	0.79	-0.012	0.53	0.28	0.59	0.90	0.95	0.17	0.08
NI	8	2.86	0.73	-0.027	0.46	0.21	0.53	0.86	0.97	0.14	0.11
NI	9	2.85	0.68	-0.024	0.42	0.18	0.49	0.82	0.97	0.11	0.14
NI	10	2.90	0.62	-0.019	0.43	0.18	0.43	0.78	0.98	0.17	0.08
NI	11	2.66	0.57	-0.009	0.42	0.18	0.37	0.74	0.95	0.12	0.13
SI	1	2.78	0.94	-0.017	0.89	0.80	0.75	0.98	0.85	0.46	-0.21
SI	2	3.98	0.93	-0.010	0.69	0.48	0.79	0.98	0.92	0.14	0.11
SI	3	4.58	0.91	-0.014	0.58	0.34	0.78	0.96	0.92	0.16	0.09
SI	4	4.83	0.89	-0.016	0.47	0.22	0.76	0.95	0.85	0.13	0.12
SI	5	4.88	0.86	-0.018	0.43	0.18	0.73	0.93	0.92	0.15	0.10
SI	6	4.97	0.83	-0.014	0.40	0.16	0.69	0.91	0.90	0.15	0.10
SI	7	4.93	0.79	-0.016	0.37	0.14	0.65	0.88	0.92	0.15	0.10
SI	8	4.83	0.74	-0.018	0.34	0.12	0.59	0.84	0.93	0.16	0.09
SI	9	4.63	0.68	-0.019	0.34	0.11	0.53	0.80	0.93	0.16	0.09
SI	10	4.00	0.62	-0.017	0.34	0.12	0.47	0.76	0.93	0.13	0.12
SI	11	2.66	0.57	-0.009	0.42	0.18	0.37	0.74	0.95	0.12	0.13
MIDC	1	2.78	0.94	-0.017	0.89	0.80	0.75	0.98	0.85	0.46	-0.21
MIDC	2	3.98	0.93	-0.009	0.67	0.44	0.80	0.98	0.95	0.12	0.13
MIDC	3	4.58	0.91	-0.009	0.56	0.31	0.79	0.97	0.93	0.13	0.12
MIDC	4	4.83	0.89	-0.011	0.49	0.24	0.76	0.95	0.90	0.16	0.09
MIDC	5	4.88	0.86	-0.014	0.44	0.19	0.73	0.93	0.93	0.16	0.09
MIDC	6	4.97	0.83	-0.012	0.40	0.16	0.70	0.91	0.93	0.16	0.09
MIDC	7	4.93	0.79	-0.010	0.37	0.14	0.65	0.88	0.95	0.16	0.09
MIDC	8	4.83	0.74	-0.015	0.34	0.12	0.60	0.84	0.97	0.16	0.09
MIDC	9	4.63	0.69	-0.016	0.34	0.11	0.53	0.80	0.97	0.17	0.08
MIDC	10	4.00	0.63	-0.013	0.34	0.11	0.47	0.76	0.93	0.13	0.12
MIDC	11	2.66	0.57	-0.009	0.42	0.18	0.37	0.74	0.95	0.12	0.13

Table 0.33 - Results for summary specificity - scenario 2 with 5 studies

Method	Threshold	No. Studies*	Specificity	Bias	SE	MSE	LCI	UCI	Coverage	Tau	Tau bias
NI	1	2.78	0.57	0.005	0.16	0.03	0.50	0.64	0.83	0.11	0.14
NI	2	2.85	0.63	-0.004	0.15	0.02	0.56	0.69	0.86	0.07	0.18
NI	3	3.00	0.70	0.003	0.16	0.03	0.63	0.76	0.86	0.11	0.14
NI	4	3.07	0.76	0.006	0.17	0.03	0.70	0.82	0.92	0.12	0.13
NI	5	2.93	0.80	0.000	0.19	0.04	0.74	0.85	0.90	0.12	0.13
NI	6	3.08	0.84	-0.002	0.21	0.04	0.78	0.89	0.88	0.15	0.10
NI	7	2.98	0.88	-0.001	0.23	0.05	0.82	0.92	0.92	0.13	0.12
NI	8	2.86	0.90	-0.003	0.24	0.06	0.85	0.93	0.83	0.09	0.16
NI	9	2.85	0.93	0.000	0.27	0.07	0.88	0.95	0.88	0.08	0.17
NI	10	2.90	0.94	-0.001	0.29	0.08	0.90	0.97	0.92	0.09	0.16
NI	11	2.66	0.96	0.002	0.38	0.15	0.92	0.98	0.93	0.13	0.12
SI	1	2.78	0.57	0.005	0.16	0.03	0.50	0.64	0.83	0.11	0.14
SI	2	3.98	0.64	0.002	0.13	0.02	0.58	0.69	0.92	0.12	0.13
SI	3	4.58	0.70	0.000	0.13	0.02	0.64	0.75	0.85	0.14	0.11
SI	4	4.83	0.76	0.000	0.14	0.02	0.70	0.80	0.92	0.17	0.08
SI	5	4.88	0.80	-0.001	0.15	0.02	0.75	0.84	0.92	0.16	0.09
SI	6	4.97	0.84	-0.001	0.16	0.02	0.80	0.88	0.86	0.16	0.09
SI	7	4.93	0.88	0.000	0.17	0.03	0.84	0.91	0.92	0.16	0.09
SI	8	4.83	0.90	-0.001	0.18	0.03	0.87	0.93	0.90	0.14	0.11
SI	9	4.63	0.93	-0.001	0.21	0.04	0.90	0.95	0.95	0.14	0.11
SI	10	4.00	0.94	-0.001	0.24	0.06	0.91	0.96	0.95	0.11	0.14
SI	11	2.66	0.96	0.002	0.38	0.15	0.92	0.98	0.93	0.13	0.12
MIDC	1	2.78	0.57	0.005	0.16	0.03	0.50	0.64	0.83	0.11	0.14
MIDC	2	3.98	0.64	0.001	0.14	0.02	0.57	0.70	0.92	0.13	0.12
MIDC	3	4.58	0.70	-0.001	0.14	0.02	0.64	0.75	0.88	0.15	0.10
MIDC	4	4.83	0.75	-0.002	0.14	0.02	0.70	0.80	0.92	0.18	0.07
MIDC	5	4.88	0.80	-0.004	0.15	0.02	0.75	0.84	0.93	0.18	0.07
MIDC	6	4.97	0.84	-0.005	0.16	0.03	0.79	0.88	0.83	0.17	0.08
MIDC	7	4.93	0.88	-0.003	0.17	0.03	0.83	0.91	0.90	0.17	0.08
MIDC	8	4.83	0.90	-0.004	0.19	0.04	0.87	0.93	0.92	0.17	0.08
MIDC	9	4.63	0.93	-0.002	0.21	0.05	0.89	0.95	0.95	0.17	0.08
MIDC	10	4.00	0.94	-0.002	0.24	0.06	0.91	0.96	0.93	0.12	0.13
MIDC	11	2.66	0.96	0.002	0.38	0.15	0.92	0.98	0.93	0.13	0.12

Fewer studies - Scenario 3

Table 0.34 - Results for summary sensitivity - scenario 3 with 5 studies

Method	Threshold	No. Studies*	Sensitivity	Bias	SE	MSE	LCI	UCI	Coverage	Tau	Tau bias
NI	1	3.15	0.92	-0.029	0.86	0.74	0.73	0.98	0.91	0.15	0.35
NI	2	2.93	0.91	-0.028	0.79	0.62	0.71	0.97	0.93	0.13	0.37
NI	3	2.89	0.90	-0.020	0.87	0.75	0.66	0.97	0.98	0.25	0.25
NI	4	3.19	0.87	-0.029	0.62	0.38	0.68	0.95	0.89	0.13	0.37
NI	5	3.20	0.84	-0.037	0.57	0.33	0.65	0.93	0.91	0.13	0.37
NI	6	2.98	0.81	-0.033	0.55	0.30	0.60	0.92	0.93	0.14	0.36
NI	7	2.94	0.77	-0.028	0.57	0.33	0.55	0.90	0.91	0.23	0.27
NI	8	2.76	0.71	-0.049	0.58	0.33	0.48	0.85	0.83	0.24	0.26
NI	9	3.13	0.67	-0.034	0.45	0.20	0.47	0.82	0.91	0.14	0.36
NI	10	3.04	0.62	-0.024	0.46	0.21	0.41	0.79	0.94	0.21	0.29
NI	11	2.85	0.53	-0.045	0.50	0.24	0.32	0.73	0.85	0.25	0.25
SI	1	3.15	0.92	-0.029	0.86	0.74	0.73	0.98	0.91	0.15	0.35
SI	2	4.17	0.91	-0.028	0.67	0.44	0.76	0.97	0.89	0.15	0.35
SI	3	4.69	0.89	-0.029	0.56	0.31	0.75	0.96	0.96	0.17	0.33
SI	4	4.91	0.87	-0.033	0.47	0.22	0.73	0.94	0.81	0.14	0.36
SI	5	4.94	0.84	-0.036	0.43	0.18	0.70	0.92	0.87	0.14	0.36
SI	6	4.98	0.80	-0.043	0.39	0.15	0.65	0.89	0.81	0.16	0.34
SI	7	4.96	0.76	-0.046	0.38	0.14	0.60	0.86	0.89	0.22	0.28
SI	8	4.83	0.72	-0.040	0.40	0.16	0.54	0.83	0.89	0.31	0.19
SI	9	4.70	0.65	-0.049	0.37	0.13	0.48	0.79	0.91	0.28	0.22
SI	10	4.20	0.60	-0.043	0.38	0.15	0.42	0.75	0.89	0.26	0.24
SI	11	2.85	0.53	-0.045	0.50	0.24	0.32	0.73	0.85	0.25	0.25
MIDC	1	3.15	0.92	-0.029	0.86	0.74	0.73	0.98	0.91	0.15	0.35
MIDC	2	4.17	0.91	-0.026	0.66	0.44	0.76	0.97	0.89	0.14	0.36
MIDC	3	4.69	0.90	-0.028	0.55	0.30	0.76	0.96	0.94	0.17	0.33
MIDC	4	4.91	0.87	-0.031	0.48	0.23	0.73	0.94	0.87	0.16	0.34
MIDC	5	4.94	0.84	-0.032	0.45	0.20	0.70	0.92	0.89	0.19	0.31
MIDC	6	4.98	0.81	-0.036	0.41	0.17	0.66	0.90	0.87	0.19	0.31
MIDC	7	4.96	0.77	-0.035	0.40	0.16	0.61	0.87	0.93	0.26	0.24
MIDC	8	4.83	0.72	-0.033	0.40	0.16	0.55	0.84	0.93	0.30	0.20
MIDC	9	4.70	0.66	-0.041	0.37	0.14	0.49	0.80	0.91	0.27	0.23
MIDC	10	4.20	0.61	-0.036	0.39	0.15	0.43	0.76	0.93	0.27	0.23
MIDC	11	2.85	0.53	-0.045	0.50	0.24	0.32	0.73	0.85	0.25	0.25

Table 0.35 - Results for summary specificity - scenario 3 with 5 studies

Method	Threshold	No. Studies*	Specificity	Bias	SE	MSE	LCI	UCI	Coverage	Tau	Tau bias
NI	1	3.15	0.56	-0.013	0.23	0.05	0.45	0.66	0.72	0.30	0.20
NI	2	2.93	0.62	-0.013	0.23	0.05	0.52	0.71	0.57	0.25	0.25
NI	3	2.89	0.68	-0.021	0.25	0.06	0.57	0.76	0.72	0.28	0.22
NI	4	3.19	0.74	-0.011	0.24	0.06	0.65	0.82	0.74	0.28	0.22
NI	5	3.20	0.79	-0.018	0.25	0.06	0.69	0.85	0.74	0.29	0.21
NI	6	2.98	0.83	-0.018	0.26	0.07	0.74	0.88	0.72	0.25	0.25
NI	7	2.94	0.86	-0.015	0.27	0.07	0.79	0.91	0.74	0.23	0.27
NI	8	2.76	0.90	-0.007	0.34	0.12	0.83	0.94	0.74	0.27	0.23
NI	9	3.13	0.92	-0.012	0.31	0.10	0.86	0.95	0.76	0.22	0.28
NI	10	3.04	0.94	-0.008	0.37	0.13	0.88	0.96	0.78	0.21	0.29
NI	11	2.85	0.96	-0.001	0.46	0.21	0.90	0.98	0.91	0.31	0.19
SI	1	3.15	0.56	-0.013	0.23	0.05	0.45	0.66	0.72	0.30	0.20
SI	2	4.17	0.62	-0.017	0.22	0.05	0.52	0.71	0.63	0.35	0.15
SI	3	4.69	0.68	-0.020	0.21	0.04	0.58	0.76	0.78	0.37	0.13
SI	4	4.91	0.74	-0.020	0.21	0.04	0.65	0.81	0.80	0.37	0.13
SI	5	4.94	0.79	-0.016	0.22	0.05	0.71	0.85	0.80	0.36	0.14
SI	6	4.98	0.83	-0.013	0.23	0.05	0.76	0.88	0.80	0.37	0.13
SI	7	4.96	0.87	-0.011	0.24	0.06	0.81	0.91	0.78	0.36	0.14
SI	8	4.83	0.90	-0.008	0.26	0.07	0.84	0.93	0.83	0.35	0.15
SI	9	4.70	0.92	-0.008	0.27	0.08	0.87	0.95	0.80	0.32	0.18
SI	10	4.20	0.94	-0.006	0.32	0.10	0.89	0.96	0.83	0.29	0.21
SI	11	2.85	0.96	-0.001	0.46	0.21	0.90	0.98	0.91	0.31	0.19
MIDC	1	3.15	0.56	-0.013	0.23	0.05	0.45	0.66	0.72	0.30	0.20
MIDC	2	4.17	0.62	-0.015	0.22	0.05	0.52	0.71	0.67	0.35	0.15
MIDC	3	4.69	0.68	-0.021	0.22	0.05	0.58	0.76	0.81	0.37	0.13
MIDC	4	4.91	0.73	-0.021	0.22	0.05	0.64	0.81	0.78	0.38	0.12
MIDC	5	4.94	0.79	-0.019	0.22	0.05	0.70	0.85	0.80	0.38	0.12
MIDC	6	4.98	0.83	-0.017	0.23	0.05	0.76	0.88	0.78	0.39	0.11
MIDC	7	4.96	0.87	-0.014	0.24	0.06	0.80	0.91	0.83	0.37	0.13
MIDC	8	4.83	0.90	-0.010	0.27	0.07	0.84	0.93	0.80	0.38	0.12
MIDC	9	4.70	0.92	-0.009	0.28	0.08	0.87	0.95	0.78	0.34	0.16
MIDC	10	4.20	0.94	-0.007	0.33	0.11	0.89	0.96	0.81	0.32	0.18
MIDC	11	2.85	0.96	-0.001	0.46	0.21	0.90	0.98	0.91	0.31	0.19

Increased number of MIDC imputations - Scenario 1

Table 0.36 - Results for summary sensitivity - scenario 1 with 10 MIDC imputations

Method	Threshold	No. Studies*	Sensitivity	Bias	SE	MSE	LCI	UCI	Coverage
NI	1	5.23	0.95	-0.007	0.67	0.45	0.84	0.98	0.96
NI	2	5.03	0.93	-0.006	0.61	0.38	0.82	0.98	0.96
NI	3	5.17	0.92	-0.006	0.53	0.28	0.81	0.97	0.96
NI	4	5.01	0.90	-0.004	0.49	0.24	0.78	0.95	0.97
NI	5	5.04	0.87	-0.001	0.45	0.20	0.75	0.94	0.97
NI	6	4.96	0.84	-0.002	0.40	0.16	0.71	0.92	0.96
NI	7	5.01	0.80	-0.003	0.36	0.13	0.67	0.88	0.96
NI	8	4.94	0.75	-0.002	0.34	0.12	0.62	0.85	0.95
NI	9	4.93	0.70	-0.005	0.32	0.10	0.56	0.81	0.95
NI	10	4.97	0.64	-0.003	0.30	0.09	0.50	0.76	0.95
NI	11	4.94	0.57	-0.004	0.29	0.08	0.44	0.70	0.95
SI	1	5.23	0.95	-0.007	0.67	0.45	0.84	0.98	0.96
SI	2	7.59	0.93	-0.010	0.45	0.20	0.85	0.97	0.93
SI	3	8.83	0.91	-0.012	0.37	0.13	0.84	0.95	0.93
SI	4	9.35	0.89	-0.012	0.32	0.10	0.82	0.94	0.92
SI	5	9.62	0.86	-0.010	0.29	0.08	0.79	0.92	0.94
SI	6	9.69	0.83	-0.012	0.26	0.07	0.75	0.89	0.93
SI	7	9.56	0.79	-0.012	0.24	0.06	0.70	0.86	0.95
SI	8	9.27	0.74	-0.013	0.23	0.05	0.65	0.82	0.94
SI	9	8.66	0.69	-0.013	0.22	0.05	0.59	0.77	0.94
SI	10	7.45	0.63	-0.010	0.23	0.05	0.52	0.73	0.94
SI	11	4.94	0.57	-0.004	0.29	0.08	0.44	0.70	0.95
MIDC	1	5.23	0.95	-0.007	0.67	0.45	0.84	0.98	0.96
MIDC	2	7.59	0.93	-0.006	0.46	0.22	0.86	0.97	0.95
MIDC	3	8.83	0.92	-0.007	0.38	0.14	0.84	0.96	0.95
MIDC	4	9.35	0.89	-0.007	0.32	0.10	0.82	0.94	0.96
MIDC	5	9.62	0.87	-0.007	0.29	0.08	0.79	0.92	0.96
MIDC	6	9.69	0.83	-0.008	0.26	0.07	0.75	0.89	0.96
MIDC	7	9.56	0.79	-0.008	0.24	0.06	0.71	0.86	0.97
MIDC	8	9.27	0.75	-0.008	0.23	0.05	0.66	0.82	0.96
MIDC	9	8.66	0.70	-0.007	0.22	0.05	0.60	0.78	0.95
MIDC	10	7.45	0.64	-0.004	0.23	0.05	0.53	0.73	0.95
MIDC	11	4.94	0.57	-0.004	0.29	0.08	0.44	0.70	0.95

Table 0.37 - Results for summary specificity - scenario 1 with 10 MIDC imputations

Method	Threshold	No. Studies*	Specificity	Bias	SE	MSE	LCI	UCI	Coverage
NI	1	5.23	0.57	-0.001	0.09	0.01	0.52	0.61	0.95
NI	2	5.03	0.63	-0.002	0.10	0.01	0.59	0.68	0.96
NI	3	5.17	0.70	-0.001	0.10	0.01	0.66	0.74	0.95
NI	4	5.01	0.75	-0.001	0.11	0.01	0.71	0.79	0.95
NI	5	5.04	0.80	-0.002	0.12	0.01	0.76	0.84	0.94
NI	6	4.96	0.85	0.000	0.13	0.02	0.81	0.88	0.96
NI	7	5.01	0.88	0.000	0.14	0.02	0.85	0.91	0.94
NI	8	4.94	0.91	0.000	0.16	0.03	0.88	0.93	0.94
NI	9	4.93	0.93	0.000	0.18	0.03	0.90	0.95	0.94
NI	10	4.97	0.94	0.000	0.21	0.04	0.92	0.96	0.95
NI	11	4.94	0.96	0.000	0.24	0.06	0.93	0.97	0.96
SI	1	5.23	0.57	-0.001	0.09	0.01	0.52	0.61	0.95
SI	2	7.59	0.63	-0.003	0.07	0.01	0.60	0.67	0.97
SI	3	8.83	0.70	-0.003	0.07	0.01	0.67	0.73	0.96
SI	4	9.35	0.75	-0.002	0.07	0.01	0.73	0.78	0.96
SI	5	9.62	0.80	-0.002	0.08	0.01	0.78	0.83	0.95
SI	6	9.69	0.84	-0.002	0.09	0.01	0.82	0.86	0.95
SI	7	9.56	0.88	-0.001	0.10	0.01	0.86	0.90	0.94
SI	8	9.27	0.91	0.000	0.11	0.01	0.89	0.92	0.94
SI	9	8.66	0.93	0.000	0.13	0.02	0.91	0.94	0.94
SI	10	7.45	0.94	0.000	0.16	0.03	0.93	0.96	0.95
SI	11	4.94	0.96	0.000	0.24	0.06	0.93	0.97	0.96
MIDC	1	5.23	0.57	-0.001	0.09	0.01	0.52	0.61	0.95
MIDC	2	7.59	0.63	-0.003	0.07	0.01	0.60	0.67	0.96
MIDC	3	8.83	0.69	-0.005	0.07	0.01	0.66	0.72	0.95
MIDC	4	9.35	0.75	-0.006	0.07	0.01	0.72	0.78	0.94
MIDC	5	9.62	0.80	-0.007	0.08	0.01	0.77	0.82	0.91
MIDC	6	9.69	0.84	-0.007	0.09	0.01	0.81	0.86	0.90
MIDC	7	9.56	0.87	-0.006	0.10	0.01	0.85	0.89	0.91
MIDC	8	9.27	0.90	-0.004	0.11	0.01	0.88	0.92	0.91
MIDC	9	8.66	0.93	-0.003	0.13	0.02	0.91	0.94	0.93
MIDC	10	7.45	0.94	-0.001	0.16	0.03	0.92	0.96	0.95
MIDC	11	4.94	0.96	0.000	0.24	0.06	0.93	0.97	0.96

Increased number of MIDC imputations - Scenario 2

Table 0.38 - Results for summary sensitivity - scenario 2 with 10 MIDC imputations

Method	Threshold	No. Studies*	Sensitivity	Bias	SE	MSE	LCI	UCI	Coverage	Tau	Tau bias
NI	1	5.23	0.95	-0.006	0.81	0.65	0.80	0.98	0.96	0.23	0.02
NI	2	5.21	0.94	-0.003	0.74	0.54	0.79	0.98	0.97	0.22	0.03
NI	3	5.11	0.92	-0.002	0.64	0.42	0.78	0.97	0.96	0.20	0.05
NI	4	5.07	0.90	-0.005	0.57	0.33	0.75	0.96	0.94	0.20	0.05
NI	5	5.03	0.87	-0.002	0.53	0.28	0.72	0.94	0.97	0.21	0.04
NI	6	4.99	0.84	-0.003	0.45	0.21	0.69	0.92	0.96	0.17	0.08
NI	7	5.02	0.80	-0.001	0.42	0.18	0.65	0.89	0.94	0.18	0.07
NI	8	5.01	0.75	-0.001	0.38	0.14	0.60	0.86	0.95	0.16	0.09
NI	9	5.05	0.70	-0.003	0.36	0.13	0.54	0.82	0.95	0.17	0.08
NI	10	5.06	0.64	-0.004	0.33	0.11	0.49	0.76	0.97	0.15	0.10
NI	11	4.95	0.58	-0.001	0.33	0.11	0.42	0.72	0.95	0.18	0.07
SI	1	5.23	0.95	-0.006	0.81	0.65	0.80	0.98	0.96	0.23	0.02
SI	2	7.71	0.93	-0.009	0.54	0.29	0.83	0.97	0.93	0.21	0.04
SI	3	8.84	0.91	-0.011	0.41	0.17	0.83	0.96	0.94	0.17	0.08
SI	4	9.38	0.89	-0.012	0.36	0.13	0.80	0.94	0.93	0.18	0.07
SI	5	9.64	0.86	-0.013	0.32	0.10	0.77	0.92	0.93	0.18	0.07
SI	6	9.70	0.83	-0.013	0.29	0.08	0.74	0.89	0.95	0.18	0.07
SI	7	9.61	0.79	-0.012	0.27	0.07	0.69	0.86	0.93	0.18	0.07
SI	8	9.36	0.74	-0.013	0.25	0.06	0.64	0.82	0.92	0.17	0.08
SI	9	8.78	0.69	-0.012	0.25	0.06	0.58	0.78	0.94	0.17	0.08
SI	10	7.48	0.63	-0.010	0.26	0.07	0.51	0.74	0.96	0.16	0.09
SI	11	4.95	0.58	-0.001	0.33	0.11	0.42	0.72	0.95	0.18	0.07
MIDC	1	5.23	0.95	-0.006	0.81	0.65	0.80	0.98	0.96	0.23	0.02
MIDC	2	7.71	0.94	-0.004	0.56	0.31	0.84	0.97	0.97	0.23	0.02
MIDC	3	8.84	0.92	-0.005	0.44	0.19	0.83	0.96	0.96	0.21	0.04
MIDC	4	9.38	0.90	-0.006	0.38	0.14	0.81	0.94	0.95	0.21	0.04
MIDC	5	9.64	0.87	-0.007	0.33	0.11	0.78	0.92	0.95	0.21	0.04
MIDC	6	9.70	0.83	-0.007	0.30	0.09	0.74	0.90	0.96	0.21	0.04
MIDC	7	9.61	0.80	-0.007	0.28	0.08	0.70	0.87	0.96	0.20	0.05
MIDC	8	9.36	0.75	-0.006	0.26	0.07	0.65	0.83	0.96	0.19	0.06
MIDC	9	8.78	0.70	-0.005	0.25	0.06	0.59	0.79	0.95	0.18	0.07
MIDC	10	7.48	0.64	-0.004	0.26	0.07	0.52	0.74	0.96	0.17	0.08
MIDC	11	4.95	0.58	-0.001	0.33	0.11	0.42	0.72	0.95	0.18	0.07

Table 0.39 - Results for summary specificity - scenario 2 with 10 MIDC imputations

Method	Threshold	No. Studies*	Specificity	Bias	SE	MSE	LCI	UCI	Coverage	Tau	Tau bias
NI	1	5.23	0.57	-0.001	0.13	0.02	0.50	0.63	0.90	0.17	0.08
NI	2	5.21	0.64	-0.001	0.13	0.02	0.57	0.69	0.88	0.16	0.09
NI	3	5.11	0.70	-0.001	0.14	0.02	0.64	0.75	0.88	0.16	0.09
NI	4	5.07	0.76	0.001	0.14	0.02	0.70	0.80	0.89	0.16	0.09
NI	5	5.03	0.80	-0.001	0.15	0.02	0.75	0.84	0.86	0.14	0.11
NI	6	4.99	0.85	0.000	0.17	0.03	0.80	0.88	0.88	0.15	0.10
NI	7	5.02	0.88	-0.001	0.18	0.03	0.84	0.91	0.90	0.15	0.10
NI	8	5.01	0.91	-0.001	0.19	0.04	0.87	0.93	0.91	0.14	0.11
NI	9	5.05	0.93	0.000	0.22	0.05	0.89	0.95	0.94	0.15	0.10
NI	10	5.06	0.94	0.000	0.24	0.06	0.91	0.96	0.93	0.14	0.11
NI	11	4.95	0.96	0.000	0.28	0.08	0.93	0.97	0.96	0.16	0.09
SI	1	5.23	0.57	-0.001	0.13	0.02	0.50	0.63	0.90	0.17	0.08
SI	2	7.71	0.63	-0.002	0.11	0.01	0.58	0.68	0.90	0.19	0.06
SI	3	8.84	0.70	-0.001	0.11	0.01	0.65	0.74	0.89	0.19	0.06
SI	4	9.38	0.75	-0.001	0.11	0.01	0.71	0.79	0.90	0.19	0.06
SI	5	9.64	0.80	-0.001	0.11	0.01	0.77	0.83	0.88	0.19	0.06
SI	6	9.70	0.84	-0.001	0.12	0.01	0.81	0.87	0.91	0.19	0.06
SI	7	9.61	0.88	-0.001	0.12	0.02	0.85	0.90	0.91	0.18	0.07
SI	8	9.36	0.91	-0.001	0.14	0.02	0.88	0.93	0.92	0.17	0.08
SI	9	8.78	0.93	0.000	0.16	0.02	0.90	0.95	0.93	0.17	0.08
SI	10	7.48	0.94	0.000	0.19	0.04	0.92	0.96	0.94	0.16	0.09
SI	11	4.95	0.96	0.000	0.28	0.08	0.93	0.97	0.96	0.16	0.09
MIDC	1	5.23	0.57	-0.001	0.13	0.02	0.50	0.63	0.90	0.17	0.08
MIDC	2	7.71	0.63	-0.002	0.12	0.01	0.58	0.68	0.92	0.21	0.04
MIDC	3	8.84	0.70	-0.003	0.11	0.01	0.65	0.74	0.92	0.23	0.02
MIDC	4	9.38	0.75	-0.004	0.12	0.01	0.71	0.79	0.92	0.23	0.02
MIDC	5	9.64	0.80	-0.005	0.12	0.01	0.76	0.83	0.91	0.23	0.02
MIDC	6	9.70	0.84	-0.005	0.13	0.02	0.80	0.87	0.91	0.24	0.01
MIDC	7	9.61	0.87	-0.005	0.13	0.02	0.84	0.90	0.93	0.23	0.02
MIDC	8	9.36	0.90	-0.004	0.14	0.02	0.88	0.92	0.92	0.22	0.03
MIDC	9	8.78	0.93	-0.002	0.16	0.03	0.90	0.94	0.94	0.21	0.04
MIDC	10	7.48	0.94	-0.001	0.19	0.04	0.92	0.96	0.94	0.18	0.07
MIDC	11	4.95	0.96	0.000	0.28	0.08	0.93	0.97	0.96	0.16	0.09

Increased number of MIDC imputations - Scenario 3

Table 0.40 - Results for summary sensitivity - scenario 3 with 10 MIDC imputations

Method	Threshold	No. Studies*	Sensitivity	Bias	SE	MSE	LCI	UCI	Coverage	Tau	Tau bias
NI	1	5.21	0.95	-0.007	0.84	0.70	0.79	0.98	0.93	0.30	0.20
NI	2	5.04	0.93	-0.006	0.76	0.58	0.78	0.98	0.95	0.26	0.24
NI	3	5.05	0.92	-0.007	0.66	0.43	0.77	0.97	0.95	0.25	0.25
NI	4	5.09	0.89	-0.008	0.62	0.38	0.73	0.96	0.94	0.30	0.20
NI	5	5.01	0.87	-0.009	0.54	0.29	0.71	0.94	0.95	0.28	0.22
NI	6	4.95	0.84	-0.007	0.51	0.26	0.67	0.92	0.92	0.30	0.20
NI	7	4.95	0.79	-0.008	0.45	0.20	0.63	0.89	0.93	0.30	0.20
NI	8	5.00	0.75	-0.006	0.43	0.18	0.57	0.86	0.92	0.30	0.20
NI	9	5.08	0.70	-0.007	0.39	0.15	0.53	0.82	0.93	0.30	0.20
NI	10	4.99	0.63	-0.010	0.38	0.14	0.46	0.77	0.90	0.30	0.20
NI	11	4.85	0.57	-0.007	0.37	0.13	0.40	0.72	0.90	0.28	0.22
SI	1	5.21	0.95	-0.007	0.84	0.70	0.79	0.98	0.93	0.30	0.20
SI	2	7.61	0.93	-0.012	0.53	0.28	0.83	0.97	0.91	0.25	0.25
SI	3	8.82	0.91	-0.015	0.43	0.18	0.82	0.95	0.91	0.25	0.25
SI	4	9.39	0.89	-0.015	0.38	0.14	0.79	0.94	0.93	0.29	0.21
SI	5	9.65	0.86	-0.016	0.34	0.12	0.76	0.92	0.91	0.32	0.18
SI	6	9.73	0.82	-0.018	0.32	0.10	0.72	0.89	0.90	0.35	0.15
SI	7	9.62	0.78	-0.018	0.30	0.09	0.67	0.86	0.91	0.35	0.15
SI	8	9.33	0.74	-0.018	0.29	0.08	0.62	0.83	0.91	0.36	0.14
SI	9	8.69	0.68	-0.017	0.29	0.08	0.56	0.79	0.91	0.37	0.13
SI	10	7.41	0.62	-0.018	0.30	0.09	0.49	0.74	0.91	0.34	0.16
SI	11	4.85	0.57	-0.007	0.37	0.13	0.40	0.72	0.90	0.28	0.22
MIDC	1	5.21	0.95	-0.007	0.84	0.70	0.79	0.98	0.93	0.30	0.20
MIDC	2	7.61	0.93	-0.007	0.57	0.33	0.83	0.97	0.95	0.29	0.21
MIDC	3	8.82	0.92	-0.008	0.46	0.21	0.82	0.96	0.94	0.30	0.20
MIDC	4	9.39	0.89	-0.009	0.40	0.16	0.80	0.94	0.96	0.33	0.17
MIDC	5	9.65	0.86	-0.010	0.36	0.13	0.77	0.92	0.95	0.35	0.15
MIDC	6	9.73	0.83	-0.011	0.33	0.11	0.73	0.90	0.93	0.37	0.13
MIDC	7	9.62	0.79	-0.011	0.30	0.09	0.68	0.87	0.93	0.37	0.13
MIDC	8	9.33	0.75	-0.011	0.29	0.08	0.63	0.83	0.92	0.37	0.13
MIDC	9	8.69	0.69	-0.010	0.29	0.08	0.56	0.79	0.92	0.37	0.13
MIDC	10	7.41	0.63	-0.011	0.30	0.09	0.49	0.75	0.92	0.34	0.16
MIDC	11	4.85	0.57	-0.007	0.37	0.13	0.40	0.72	0.90	0.28	0.22

Table 0.41 - Results for summary specificity - scenario 3 with 10 MIDC imputations

Method	Threshold	No. Studies*	Specificity	Bias	SE	MSE	LCI	UCI	Coverage	Tau	Tau bias
NI	1	5.21	0.57	0.001	0.21	0.04	0.47	0.66	0.84	0.38	0.12
NI	2	5.04	0.63	-0.003	0.21	0.04	0.53	0.72	0.82	0.38	0.12
NI	3	5.05	0.69	-0.005	0.21	0.04	0.60	0.77	0.83	0.37	0.13
NI	4	5.09	0.75	-0.004	0.22	0.05	0.67	0.82	0.83	0.37	0.13
NI	5	5.01	0.80	-0.003	0.22	0.05	0.72	0.86	0.84	0.36	0.14
NI	6	4.95	0.84	-0.003	0.23	0.05	0.77	0.89	0.85	0.35	0.15
NI	7	4.95	0.88	-0.003	0.24	0.06	0.82	0.92	0.85	0.35	0.15
NI	8	5.00	0.90	-0.001	0.25	0.06	0.85	0.94	0.86	0.33	0.17
NI	9	5.08	0.93	-0.001	0.27	0.07	0.88	0.95	0.88	0.32	0.18
NI	10	4.99	0.94	-0.002	0.29	0.08	0.90	0.96	0.86	0.30	0.20
NI	11	4.85	0.96	-0.002	0.31	0.10	0.92	0.97	0.89	0.28	0.22
SI	1	5.21	0.57	0.001	0.21	0.04	0.47	0.66	0.84	0.38	0.12
SI	2	7.61	0.63	-0.002	0.18	0.03	0.55	0.71	0.86	0.42	0.08
SI	3	8.82	0.70	-0.004	0.17	0.03	0.62	0.76	0.88	0.43	0.07
SI	4	9.39	0.75	-0.003	0.17	0.03	0.69	0.81	0.89	0.43	0.07
SI	5	9.65	0.80	-0.003	0.17	0.03	0.74	0.85	0.90	0.43	0.07
SI	6	9.73	0.84	-0.003	0.17	0.03	0.79	0.88	0.90	0.43	0.07
SI	7	9.62	0.88	-0.002	0.18	0.03	0.83	0.91	0.90	0.41	0.09
SI	8	9.33	0.90	-0.002	0.19	0.04	0.87	0.93	0.89	0.40	0.10
SI	9	8.69	0.93	-0.002	0.20	0.04	0.89	0.95	0.87	0.38	0.12
SI	10	7.41	0.94	-0.001	0.24	0.06	0.91	0.96	0.90	0.36	0.14
SI	11	4.85	0.96	-0.002	0.31	0.10	0.92	0.97	0.89	0.28	0.22
MIDC	1	5.21	0.57	0.001	0.21	0.04	0.47	0.66	0.84	0.38	0.12
MIDC	2	7.61	0.63	-0.002	0.18	0.03	0.55	0.71	0.88	0.43	0.07
MIDC	3	8.82	0.69	-0.005	0.17	0.03	0.62	0.76	0.89	0.44	0.06
MIDC	4	9.39	0.75	-0.006	0.17	0.03	0.68	0.81	0.91	0.45	0.05
MIDC	5	9.65	0.80	-0.006	0.17	0.03	0.74	0.85	0.90	0.45	0.05
MIDC	6	9.73	0.84	-0.006	0.18	0.03	0.79	0.88	0.90	0.45	0.05
MIDC	7	9.62	0.87	-0.006	0.18	0.03	0.83	0.91	0.90	0.44	0.06
MIDC	8	9.33	0.90	-0.005	0.19	0.04	0.86	0.93	0.89	0.43	0.07
MIDC	9	8.69	0.92	-0.004	0.21	0.04	0.89	0.95	0.89	0.40	0.10
MIDC	10	7.41	0.94	-0.002	0.24	0.06	0.91	0.96	0.90	0.37	0.13
MIDC	11	4.85	0.96	-0.002	0.31	0.10	0.92	0.97	0.89	0.28	0.22

REFERENCE LIST

1. Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? 2009;338.
2. Mathers CD, Loncar D. Projections of global mortality and burden of disease from 2002 to 2030. PLoS Med. 2006;3(11):e442.
3. 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 : British Medical Journal. 2013;346.
4. Kent DM, Hayward RA. Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. Jama. 2007;298(10):1209-12.
5. Roozenbeek B, Maas AI, Lingsma HF, Butcher I, Lu J, Marmarou A, et al. Baseline characteristics and statistical power in randomized controlled trials: selection, prognostic targeting, or covariate adjustment? Critical care medicine. 2009;37(10):2683-90.
6. Hernandez AV, Steyerberg EW, Habbema JD. Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. Journal of clinical epidemiology. 2004;57(5):454-60.
7. Lingsma HF, Roozenbeek B, Li B, Lu J, Weir J, Butcher I, et al. Large between-center differences in outcome after moderate and severe traumatic brain injury in the international mission on prognosis and clinical trial design in traumatic brain injury (IMPACT) study. Neurosurgery. 2011;68(3):601-7; discussion 7-8.
8. Riley RD, Hayden JA, Steyerberg EW, Moons KG, Abrams K, Kyzas PA, et al. Prognosis Research Strategy (PROGRESS) 2: prognostic factor research. 2013;2013/02/09:e1001380.
9. Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. 2013;2013/02/09:e1001381.
10. Hingorani AD, Windt DAvd, Riley RD, Abrams K, Moons KGM, Steyerberg EW, et al. Prognosis research strategy (PROGRESS) 4: Stratified medicine research. BMJ : British Medical Journal. 2013;346.
11. Adams ST, Leveson SH. Clinical prediction rules. BMJ (Clinical research ed). 2012;344:d8312.
12. 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;14:3-.
13. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, May M, Brindle P. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. BMJ (Clinical research ed). 2007;335(7611):136.
14. Steyerberg EW, Mushkudiani N, Perel P, Butcher I, Lu J, McHugh GS, et al. Predicting Outcome after Traumatic Brain Injury: Development and International Validation of Prognostic Scores Based on Admission Characteristics. PLOS Medicine. 2008;5(8):e165.
15. Douketis J, Tosetto A, Marcucci M, Baglin T, Cushman M, Eichinger S, et al. Patient-level meta-analysis: effect of measurement timing, threshold, and patient age on ability of D-dimer testing to assess recurrence risk after unprovoked venous thromboembolism. [Review]. Annals of internal medicine. 2010;153(8):523-31.
16. Abo-Zaid G, Guo B, Deeks JJ, Debray TP, Steyerberg EW, Moons KG, et al. Individual participant data meta-analyses should not ignore clustering. 2013;2013/05/09:865-73.
17. Wynants L, Vergouwe Y, Van Huffel S, Timmerman D, Van Calster B. Does ignoring clustering in multicenter data influence the performance of prediction models? A simulation study. Statistical methods in medical research. 2016:0962280216668555.

18. Bouwmeester W, Twisk JW, Kappen TH, van Klei WA, Moons KG, Vergouwe Y. Prediction models for clustered data: comparison of a random intercept and standard regression model. BMC medical research methodology. 2013;13:19.
19. 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;69:40-50.
20. Debray TP, Moons KG, 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;2013/01/12:3158-80.
21. 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 (Clinical research ed). 2016;353.
22. Debray TPA, Koffijberg H, Lu D, Vergouwe Y, Steyerberg EW, M Moons KG. Incorporating published univariable associations in diagnostic and prognostic modeling. BMC medical research methodology. 2012;12(1):121.
23. Debray TPA, Koffijberg H, Vergouwe Y, Moons KGM, Steyerberg EW. Aggregating published prediction models with individual participant data: a comparison of different approaches. Statistics in medicine. 2012;31(23):2697-712.
24. van Klaveren D, Steyerberg E, Perel P, Vergouwe Y. Assessing discriminative ability of risk models in clustered data. BMC medical research methodology. 2014;14(1):5.
25. Hemingway H, Riley RD, Altman DG. Ten steps towards improving prognosis research. BMJ (Clinical research ed). 2009;339.
26. Moons KGM, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ (Clinical research ed). 2009;338.
27. Royston P, Moons KG, Altman DG, Vergouwe Y. Prognosis and prognostic research: Developing a prognostic model. BMJ (Clinical research ed). 2009;338:b604.
28. Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. 2009;2009/05/30:b605.
29. Brown MA, Lindheimer MD, de Swiet M, Van Assche A, Moutquin JM. The classification and diagnosis of the hypertensive disorders of pregnancy: statement from the International Society for the Study of Hypertension in Pregnancy (ISSHP). Hypertension in pregnancy. 2001;20(1):IX-XIV.
30. Morris RK, Riley RD, Doug M, Deeks JJ, Kilby MD. Diagnostic accuracy of spot urinary protein and albumin to creatinine ratios for detection of significant proteinuria or adverse pregnancy outcome in patients with suspected pre-eclampsia: systematic review and meta-analysis. BMJ (Clinical research ed). 2012;345.
31. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). European journal of cancer. 2009;45(2):228-47.
32. Group EBCTC. Tamoxifen for early breast cancer: an overview of the randomised trials. . Lancet (London, England). 1998;351(9114):1451-67.
33. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. BMC medicine. 2015;13:1.
34. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. 2009.
35. Hippisley-Cox J, Coupland C. Predicting risk of osteoporotic fracture in men and women in England and Wales: prospective derivation and validation of QFractureScores. BMJ (Clinical research ed). 2009;339:b4229.

36. Hippisley-Cox J, Coupland C. Derivation and validation of updated QFracture algorithm to predict risk of osteoporotic fracture in primary care in the United Kingdom: prospective open cohort study. BMJ (Clinical research ed). 2012;344:e3427.
37. Tosetto A, Iorio A, Marcucci M, Baglin T, Cushman M, Eichinger S, et al. Predicting disease recurrence in patients with previous unprovoked venous thromboembolism: A proposed prediction score (DASH). Journal of Thrombosis and Haemostasis. 2012;10(6):1019-25.
38. Harrell FE. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis: Springer; 2001.
39. Rodger MA, Kahn SR, Wells PS, Anderson DA, Chagnon I, Le GG, et al. Identifying unprovoked thromboembolism patients at low risk for recurrence who can discontinue anticoagulant therapy. CMAJ Canadian Medical Association Journal. 2008;179(5):417-26.
40. Eichinger S, Heinze G, Jandeck LM, Kyrle PA. Risk assessment of recurrence in patients with unprovoked deep vein thrombosis or pulmonary embolism: the Vienna prediction model. Circulation. 2010;121(14):1630-6.
41. Cox DR. Regression models and life tables. JR stat soc B. 1972;34(2):187-220.
42. Klein JP, Van Houwelingen HC, Ibrahim JG, Scheike TH. Handbook of survival analysis: Chapman and Hall/CRC; 2013.
43. Royston P, Parmar MK. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Statistics in medicine. 2002;21(15):2175-97.
44. Royston P. Flexible parametric alternatives to the Cox model, and more. Stata Journal. 2001;1(1):1-28.
45. Royston P. Flexible parametric alternatives to the Cox model: update. The Stata Journal. 2004;4(1):98-101.
46. Royston P, Lambert PC. Flexible parametric survival analysis using Stata: beyond the Cox model. 2006.
47. Royston P. Flexible parametric alternatives to the Cox model, and more. 2001;1:1-28.
48. Steyerberg EW, Eijkemans MJ, Harrell FE, Jr., Habbema JD. Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. Med Decis Making. 2001;21(1):45-56.
49. Sun GW, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. Journal of clinical epidemiology. 1996;49(8):907-16.
50. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Statistics in medicine. 2006;25(1):127-41.
51. Altman DG, Lausen B, Sauerbrei W, Schumacher M. Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. Journal of the National Cancer Institute. 1994;86(11):829-35.
52. Altman DG, Royston P. The cost of dichotomising continuous variables. BMJ (Clinical research ed). 2006;332(7549):1080.
53. Collins GS, Ogundimu EO, Cook JA, Manach YL, Altman DG. Quantifying the impact of different approaches for handling continuous predictors on the performance of a prognostic model. Statistics in medicine. 2016;35(23):4124-35.
54. Royston P, Sauerbrei W. Multivariable model-building: a pragmatic approach to regression anaylsis based on fractional polynomials for modelling continuous variables: Wiley. com; 2008.
55. Royston P, Ambler G, Sauerbrei W. The use of fractional polynomials to model continuous risk variables in epidemiology. International journal of epidemiology. 1999;28(5):964-74.
56. Rubin DB. Multiple Imputation for Nonresponse in Surveys. Wiley: New York. 1987.
57. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Statistics in medicine. 2011;30(4):377-99.

58. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ (Clinical research ed). 2009;338:b2393.
59. Sauerbrei W, Royston P, Binder H. Selection of important variables and determination of functional form for continuous predictors in multivariable model building. 2007;26:5512-28.
60. Mantel N. Why stepdown procedures in variable selection. Technometrics. 1970;12(3):621-5.
61. Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KG. Internal and external validation of predictive models: a simulation study of bias and precision in small samples. Journal of clinical epidemiology. 2003;56(5):441-7.
62. Steyerberg EW, Harrell FE, Jr., Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. Journal of clinical epidemiology. 2001;54(8):774-81.
63. Altman DG, Royston P. What do we mean by validating a prognostic model? Statistics in medicine. 2000;19(4):453-73.
64. Moons KG, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart (British Cardiac Society). 2012;98(9):691-8.
65. Hickey GL, Grant SW, Murphy GJ, Bhabra M, Pagano D, McAllister K, et al. Dynamic trends in cardiac surgery: why the logistic EuroSCORE is no longer suitable for contemporary cardiac surgery and implications for future risk models. European journal of cardio-thoracic surgery : official journal of the European Association for Cardio-thoracic Surgery. 2013;43(6):1146-52.
66. Royston P, Parmar MK, Sylvester R. Construction and validation of a prognostic model across several studies, with an application in superficial bladder cancer. 2004;2004/03/18:907-26.
67. Harrell FE, Jr., 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):361-87.
68. Moons KM, Altman DG, Reitsma JB, 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.
69. Toll DB, Janssen KJ, Vergouwe Y, Moons KG. Validation, updating and impact of clinical prediction rules: a review. Journal of clinical epidemiology. 2008;61(11):1085-94.
70. 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;68(3):279-89.
71. 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 (Clinical research ed). 2017;356.
72. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology (Cambridge, Mass). 2010;21(1):128-38.
73. Vergouwe Y, Moons KG, 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):971-80.
74. van Klaveren D, Gonen M, Steyerberg EW, Vergouwe Y. A new concordance measure for risk prediction models in external validation settings. Statistics in medicine. 2016;35(23):4136-52.
75. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. Statistics in medicine. 2004;23(5):723-48.
76. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. Journal of clinical epidemiology. 2016.

77. Collins GS, de Groot JA, 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;14:40.
78. Wyatt JC, Altman DG. Commentary: Prognostic models: clinically useful or quickly forgotten? BMJ (Clinical research ed). 1995;311(7019):1539-41.
79. Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. Annals of internal medicine. 2006;144(3):201-9.
80. James BC. Making it easy to do it right. The New England journal of medicine. 2001;345(13):991-3.
81. Hill JC, Whitehurst DG, Lewis M, Bryan S, Dunn KM, Foster NE, et al. Comparison of stratified primary care management for low back pain with current best practice (STarT Back): a randomised controlled trial. Lancet (London, England). 2011;378(9802):1560-71.
82. 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 (Clinical research ed). 2016;353.
83. Echouffo-Tcheugui JB, Batty GD, Kivimäki M, Kengne AP. Risk Models to Predict Hypertension: A Systematic Review. PLOS ONE. 2013;8(7):e67370.
84. Perel P, Edwards P, Wentz R, Roberts I. Systematic review of prognostic models in traumatic brain injury. BMC Medical Informatics and Decision Making. 2006;6(1):38.
85. Plint AC, Moher D, Morrison A, Schulz K, Altman DG, Hill C, et al. Does the CONSORT checklist improve the quality of reports of randomised controlled trials? A systematic review. The Medical journal of Australia. 2006;185(5):263-7.
86. Bossuyt PM. STARD statement: still room for improvement in the reporting of diagnostic accuracy studies. Radiology. 2008;248(3):713-4.
87. Smidt N, Rutjes AW, van der Windt DA, Ostelo RW, Bossuyt PM, Reitsma JB, et al. The quality of diagnostic accuracy studies since the STARD statement: has it improved? Neurology. 2006;67(5):792-7.
88. Ingui BJ, Rogers MAM. Searching for Clinical Prediction Rules in Medline. Journal of the American Medical Informatics Association. 2001;8(4):391-7.
89. Geersing G-J, Bouwmeester W, Zuithoff P, Spijker R, Leeflang M, Moons K. Search Filters for Finding Prognostic and Diagnostic Prediction Studies in Medline to Enhance Systematic Reviews. PLOS ONE. 2012;7(2):e32844.
90. Wong SS, Wilczynski NL, Haynes RB, Ramkissoonsingh R. Developing optimal search strategies for detecting sound clinical prediction studies in MEDLINE. AMIA Annual Symposium proceedings AMIA Symposium. 2003:728-32.
91. 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.
92. Wolff R WP, Mallett S, et al. PROBAST: a risk of bias tool for prediction modelling studies. Cochrane Colloquium Vienna. 2015.
93. Hayden JA, Cote P, Bombardier C. Evaluation of the quality of prognosis studies in systematic reviews. Annals of internal medicine. 2006;2006/03/22:427-37.
94. Hayden JA, van der Windt DA, Cartwright JL, Cote P, Bombardier C. Assessing bias in studies of prognostic factors. Annals of internal medicine. 2013;2013/02/20:280-6.
95. Ensor J, Riley RD, Moore D, Snell KI, Bayliss S, Fitzmaurice D. Systematic review of prognostic models for recurrent venous thromboembolism (VTE) post-treatment of first unprovoked VTE. BMJ open. 2016;6(5):e011190.
96. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. 2009;2009/07/28:e1-34.

97. Stewart LA, Clarke M, Rovers M, et al. Preferred reporting items for a systematic review and meta-analysis of individual participant data: The prisma-ipd statement. Jama. 2015;313(16):1657-65.
98. Ensor J, Riley RD, Jowett S, Monahan M, Snell KIE, Bayliss S, et al. Prediction of risk of recurrence of venous thromboembolism following treatment for a first unprovoked venous thromboembolism: systematic review, prognostic model and clinical decision rule, and economic evaluation. Health Technol Assess. 2016;20(12).
99. Whitehead A, Whitehead J. A general parametric approach to the meta-analysis of randomized clinical trials. Statistics in medicine. 1991;10(11):1665-77.
100. Higgins JPT, Green S. Cochrane handbook for systematic reviews of interventions: John Wiley & Sons; 2011.
101. 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. Res Synth Methods. 2016;7(1):55-79.
102. IntHout J, Ioannidis JP, 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;14:25.
103. Langan D, Higgins JP, Simmonds M. An empirical comparison of heterogeneity variance estimators in 12 894 meta-analyses. Res Synth Methods. 2015;6(2):195-205.
104. Cornell JE, Mulrow CD, Localio R, Stack CB, Meibohm AR, Guallar E, et al. Random-effects meta-analysis of inconsistent effects: a time for change. Annals of internal medicine. 2014;160(4):267-70.
105. Broström G, Holmberg H. Generalized linear models with clustered data: Fixed and random effects models. Computational Statistics & Data Analysis. 2011;55(12):3123-34.
106. Noh M, Lee Y. REML estimation for binary data in GLMMs. Journal of Multivariate Analysis. 2007;98(5):896-915.
107. Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. Statistics in medicine. 1999;18(20):2693-708.
108. Berkey CS, Hoaglin DC, Mosteller F, Colditz GA. A random-effects regression model for meta-analysis. Statistics in medicine. 1995;14(4):395-411.
109. Austin PC. Estimating multilevel logistic regression models when the number of clusters is low: a comparison of different statistical software procedures. Int J Biostat. 2010;6(1):Article 16.
110. Sidik K, Jonkman JN. A comparison of heterogeneity variance estimators in combining results of studies. Statistics in medicine. 2007;26(9):1964-81.
111. Snell KIE. Development and application of statistical methods for prognosis research. Birmingham, UK: University of Birmingham; 2015.
112. Higgins JP, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. J R Stat Soc Ser A Stat Soc. 2009;172(1):137-59.
113. Riley RD, Higgins JP, Deeks JJ. Interpretation of random effects meta-analyses. 2011;2011/02/12:d549.
114. Partlett C, Riley RD. Random effects meta-analysis: Coverage performance of 95% confidence and prediction intervals following REML estimation. Statistics in medicine. 2017;36(2):301-17.
115. Debray TP, Moons KG, van Valkenhoef G, Efthimiou O, Hummel N, Groenwold RH, et al. Get real in individual participant data (IPD) meta-analysis: a review of the methodology. Res Synth Methods. 2015;6(4):293-309.
116. Simmonds M, Stewart G, Stewart L. A decade of individual participant data meta-analyses: A review of current practice. Contemp Clin Trials. 2015;45(Pt A):76-83.
117. Stewart GB, Altman DG, Askie LM, Duley L, Simmonds MC, Stewart LA. Statistical analysis of individual participant data meta-analyses: a comparison of methods and recommendations for practice. PLoS One. 2012;7(10):e46042.

118. 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. 2002;2002/01/29:371-87.
119. Simmonds MC, Higgins JP, Stewart LA, Tierney JF, Clarke MJ, Thompson SG. Meta-analysis of individual patient data from randomized trials: a review of methods used in practice. Clin Trials. 2005;2(3):209-17.
120. 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. 2016:n/a-n/a.
121. Debray TP, Moons KG, Abo-Zaid GM, Koffijberg H, Riley RD. Individual participant data meta-analysis for a binary outcome: one-stage or two-stage? PLoS One. 2013;8(4):e60650.
122. Ahmed I, Sutton AJ, Riley RD. Assessment of publication bias, selection bias, and unavailable data in meta-analyses using individual participant data: a database survey. BMJ (Clinical research ed). 2012;344:d7762.
123. Audigier V, White IR, Jolani S, Debray TPA, Quartagno M, Carpenter J, et al. Multiple imputation for multilevel data with continuous and binary variables. eprint arXiv:170200971. 2017.
124. Jolani S, Debray TP, Koffijberg H, van Buuren S, Moons KG. Imputation of systematically missing predictors in an individual participant data meta-analysis: a generalized approach using MICE. Statistics in medicine. 2015;34(11):1841-63.
125. Fibrinogen Studies C, Jackson D, White I, Kostis JB, Wilson AC, Folsom AR, et al. Systematically missing confounders in individual participant data meta-analysis of observational cohort studies. Statistics in medicine. 2009;28(8):1218-37.
126. Kovacic J, Varnai VM. A graphical model approach to systematically missing data in meta-analysis of observational studies. Statistics in medicine. 2016;35(24):4443-58.
127. Kline D, Andridge R, Kaizar E. Comparing multiple imputation methods for systematically missing subject-level data. Res Synth Methods. 2015.
128. Resche-Rigon M, White IR, Bartlett JW, Peters SA, Thompson SG, Group P-IS. Multiple imputation for handling systematically missing confounders in meta-analysis of individual participant data. Statistics in medicine. 2013;32(28):4890-905.
129. Resche-Rigon M, White IR. Multiple imputation by chained equations for systematically and sporadically missing multilevel data. Statistical methods in medical research. 2016.
130. Quartagno M, Carpenter JR. Multiple imputation for IPD meta-analysis: allowing for heterogeneity and studies with missing covariates. Statistics in medicine. 2016;35(17):2938-54.
131. Kleinrouweler CE, Cheong-See FM, Collins GS, Kwee A, Thangaratinam S, Khan KS, et al. Prognostic models in obstetrics: available, but far from applicable. American journal of obstetrics and gynecology. 2016;214(1):79-90 e36.
132. Altman DG. Prognostic models: a methodological framework and review of models for breast cancer. Cancer Invest. 2009;27(3):235-43.
133. Collins GS, Mallett S, Omar O, Yu LM. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. BMC medicine. 2011;9:103.
134. Bouwmeester W, Zuithoff NP, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, et al. Reporting and methods in clinical prediction research: a systematic review. PLoS Med. 2012;9(5):1-12.
135. Collins GS, Omar O, Shanyinde M, Yu LM. A systematic review finds prediction models for chronic kidney disease were poorly reported and often developed using inappropriate methods. Journal of clinical epidemiology. 2013;66(3):268-77.
136. Mallett S, Royston P, Dutton S, Waters R, Altman DG. Reporting methods in studies developing prognostic models in cancer: a review. BMC medicine. 2010;8:20.
137. van Es N, Kraaijpoel N, Klok FA, Huisman MV, Den Exter PL, Mos IC, et al. The original and simplified Wells rules and age-adjusted D-dimer testing to rule out pulmonary embolism: an individual patient data meta-analysis. Journal of thrombosis and haemostasis : JTH. 2017.

138. Shen Y, Liu YM, Wang B, Zhu YG, Wang YY, Wang XL, et al. External validation and comparison of six prognostic models in a prospective cohort of HBV-ACLF in China. Ann Hepatol. 2016;15(2):236-45.
139. Masconi K, Matsha TE, Erasmus RT, Kengne AP. Independent external validation and comparison of prevalent diabetes risk prediction models in a mixed-ancestry population of South Africa. Diabetol Metab Syndr. 2015;7:42.
140. Heng DY, Xie W, Regan MM, Harshman LC, Bjarnason GA, Vaishampayan UN, et al. External validation and comparison with other models of the International Metastatic Renal-Cell Carcinoma Database Consortium prognostic model: a population-based study. Lancet Oncol. 2013;14(2):141-8.
141. Willis BH, Hyde CJ. What is the test's accuracy in my practice population? Tailored meta-analysis provides a plausible estimate. Journal of clinical epidemiology. 2015;68(8):847-54.
142. Janssen KJ, Moons KG, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. Journal of clinical epidemiology. 2008;61(1):76-86.
143. Steyerberg EW, Borsboom GJ, van Houwelingen HC, Eijkemans MJ, Habbema JD. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. Statistics in medicine. 2004;23(16):2567-86.
144. Van Houwelingen HC, Thorogood J. Construction, validation and updating of a prognostic model for kidney graft survival. Statistics in medicine. 1995;14(18):1999-2008.
145. van Houwelingen HC. Validation, calibration, revision and combination of prognostic survival models. Statistics in medicine. 2000;19(24):3401-15.
146. Janssen KJ, Vergouwe Y, Kalkman CJ, Grobbee DE, Moons KG. A simple method to adjust clinical prediction models to local circumstances. Canadian journal of anaesthesia = Journal canadien d'anesthesie. 2009;56(3):194-201.
147. Royston P, Altman DG. External validation of a Cox prognostic model: principles and methods. BMC medical research methodology. 2013;13:33.
148. Proust-Lima C, Taylor JM. Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment PSA: a joint modeling approach. Biostatistics. 2009;10(3):535-49.
149. Akbarov A, Williams R, Brown B, Mamas M, Peek N, Buchan I, et al. A Two-stage Dynamic Model to Enable Updating of Clinical Risk Prediction from Longitudinal Health Record Data: Illustrated with Kidney Function. Stud Health Technol Inform. 2015;216:696-700.
150. 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;17(1):1.
151. Naess IA, Christiansen SC, Romundstad P, Cannegieter SC, Rosendaal FR, Hammerstrom J. Incidence and mortality of venous thrombosis: a population-based study. Journal of thrombosis and haemostasis : JTH. 2007;5(4):692-9.
152. Kyrle PA, Eichinger S. Deep vein thrombosis. Lancet (London, England). 2005;365(9465):1163-74.
153. Kyrle PA, Rosendaal FR, Eichinger S. Risk assessment for recurrent venous thrombosis. Lancet (London, England). 2010;376(9757):2032-9.
154. Baglin T, Luddington R, Brown K, Baglin C. Incidence of recurrent venous thromboembolism in relation to clinical and thrombophilic risk factors: prospective cohort study. Lancet (London, England). 2003;362(9383):523-6.
155. Rosendaal FR. Venous thrombosis: a multicausal disease. Lancet (London, England). 1999;353(9159):1167-73.
156. Kearon C, Akl EA, Comerota AJ, Prandoni P, Bounameaux H, Goldhaber SZ, et al. Antithrombotic therapy for VTE disease: Antithrombotic Therapy and Prevention of Thrombosis, 9th

ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. Chest. 2012;141(2 Suppl):e419S-94S.
157. Keeling D, Baglin T, Tait C, Watson H, Perry D, Baglin C, et al. Guidelines on oral anticoagulation with warfarin – fourth edition. British Journal of Haematology. 2011;154(3):311-24.
158. Ensor J, Riley RD, Moore D, Bayliss S, Jowett S, Fitzmaurice DA. Protocol for a systematic review of prognostic models for the recurrence of venous thromboembolism (VTE) following treatment for a first unprovoked VTE. 2013;2013/10/05:91.
159. DerSimonian R, Laird N. Meta-analysis in clinical trials. 1986;1986/09/01:177-88.
160. Marcucci M, Iorio A, Douketis JD, Eichinger S, Tosetto A, Baglin T, et al. Risk of recurrence after a first unprovoked venous thromboembolism: External validation of the Vienna Prediction Model using pooled individual patient data. Journal of thrombosis and haemostasis : JTH. 2015.
161. Tritschler T, Mean M, Limacher A, Rodondi N, Aujesky D. Predicting recurrence after unprovoked venous thromboembolism: prospective validation of the updated Vienna Prediction Model. Blood. 2015;126(16):1949-51.
162. Emmerich J. [Risk factors of the recurrence of venous thromboembolism]. [French]. Revue du Praticien. 2007;57(7):717-8.
163. Meyer G. [Pulmonary embolism. Significant diagnostic and therapeutic advances]. [French]. Revue du Praticien. 2007;57(7):709-10.
164. Ramalle-Gomara E, Javier Ochoa-Gomez F. Low risk of pulmonary embolism after discontinuing anticoagulant treatment for deep venous thrombosis?. [Spanish]. FMC Formacion Medica Continuada en Atencion Primaria. 2008;15(7):480.
165. Man M, Bugalho A. [Update in pulmonary thromboembolic disease]. [Review] [90 refs] [Portuguese]. Revista Portuguesa de Pneumologia. 2009;15(3):483-505.
166. Vorob'eva NM, Panchenko EP, Dobrovol'skii AB, Titaeva EV, Fedotkina I, Kirienko AI. [Risk factors for venous thromboembolic complications and their association with D-dimer level]. [Russian]. Terapevticheskii Arkhiv. 2010;82(8):30-4.
167. Vorob'eva NM, Panchenko EP, Dobrovol'skii AB, Titaeva EV, Khasanova ZB, Konovalova NV, et al. [Independent predictors of deep vein thrombosis (results of prospective 18 months study)]. [Russian]. Kardiologiia. 2010;50(12):52-8.
168. Cost-effectiveness of tailoring anticoagulant therapy by a VTE recurrence prediction model in patients with venous thrombo-embolism as compared to care-as-usual: The VISTA study. - VISTA. 2013.
169. Rodger M, KovacsC MJ, Kahn S, Wells P, Anderson D, Gregoire LG, et al. Extended follow-up of the multi-center multi-national prospective cohort study that derived the "men continue and HERDOO2" clinical decision rule which identifies low risk patients who may be able to discontinue oral anticoagulants (Oac) 5-7 months after treatment for unprovoked venous thromboembolism (VTE). Blood. 2009;Conference: 51st Annual Meeting of the American Society of Hematology, ASH New Orleans, LA United States. Conference Start: 20091205 Conference End: 20091208. Conference Publication:(var.pagings).
170. Rodger MA. Clinical Decision Rule Validation Study to Predict Low Recurrent Risk in Patients With Unprovoked Venous Thromboembolism. clinicaltrialsgov. 2011.
171. Rodger MA, Rodger M, Kovacs M, Le GG, Kahn S, Anderson D, et al. Extended follow-up of the multi-center prospective cohort that derived the 'men continue and HERDOO2' clinical decision rule identifying low risk unprovoked patients. Journal of Thrombosis and Haemostasis. 2011;Conference: 23rd Congress of the International Society on Thrombosis and Haemostasis 57th Annual SSC Meeting Kyoto Japan. Conference Start: 20110723 Conference End: 20110728. Conference Publication:(var.pagings):39-40.
172. Lazo-Langner A, Abdulrehman J, Taylor EJ, Sharma S, Kovacs MJ. The use of the REVERSE study clinical prediction rule for risk stratification after initial anticoagulation results in decreased recurrences in patients with idiopathic venous thromboembolism. Journal of Thrombosis and

HaemostasisConference: 24th Congress of the International Society on Thrombosis and Haemostasis Amsterdam NetherlandsConference Start: 20130629 Conference End: 20130704Conference Publication: (varpagings)11 (pp 879), 201. 2013(var.pagings):879.
173. Marcucci M, Eichinger S, Iorio A, Douketis JD, Tosetto A, Baglin TPT, et al. External validation and updating of the Vienna Prediction Model for recurrent venous thromboembolism using a pooled individual patient data database. Journal of Thrombosis and HaemostasisConference: 24th Congress of the International Society on Thrombosis and Haemostasis Amsterdam NetherlandsConference Start: 20130629 Conference End: 20130704Conference Publication: (varpagings)11 (pp 879-880). 2013(var.pagings):879-80.
174. Rodger M, Kovacs M, Le GG, Anderson D, Righini M, Beaudoin T. The REVERSE I and II studies: Impact of using. Men continue and HERDOO2 clinical decision rule to guide anticoagulant therapy in patients with first unprovoked venous thromboembolism. Journal of Thrombosis and HaemostasisConference: 24th Congress of the International Society on Thrombosis and Haemostasis Amsterdam NetherlandsConference Start: 20130629 Conference End: 20130704Conference Publication: (varpagings)11 (pp 114), 201. 2013(var.pagings):114.
175. Eichinger S, Heinze G, Kyrle PA. Risk assessment model to predict recurrence in patients with unprovoked deep vein thrombosis or pulmonary embolism. Blood. 2009;Conference: 51st Annual Meeting of the American Society of Hematology, ASH New Orleans, LA United States. Conference Start: 20091205 Conference End: 20091208. Conference Publication:(var.pagings).
176. Raskob GE, Anthonie LWA, Prins MH, Schellong S, Buller HR. Risk assessment for recurrent venous thromboembolism (VTE) after 6-14 months of anticoagulant treatment. Journal of Thrombosis and Haemostasis. 2011;Conference: 23rd Congress of the International Society on Thrombosis and Haemostasis 57th Annual SSC Meeting Kyoto Japan. Conference Start: 20110723 Conference End: 20110728. Conference Publication:(var.pagings):857-8.
177. Tosetto A, Iorio A, Marcucci M, Baglin T, Cushman M, Eichinger S, et al. Predicting disease recurrence in patients with previous unprovoked venous thromboembolism: The DASH prediction score. Blood. 2011;Conference: 53rd Annual Meeting of the American Society of Hematology, ASH 2011 San Diego, CA United States. Conference Start: 20111210 Conference End: 20111213. Conference Publication:(var.pagings).
178. Tosetto A, Iorio A, Marcucci M, Baglin T, Cushman M, Eichinger S, et al. Clinical prediction of VTE recurrence in patients with previous unprovoked venous thromboembolism. Results from an individual-level meta-analysis. Pathophysiology of Haemostasis and Thrombosis. 2010;Conference: 21st International Congress on Thrombosis - The Start of a New Era Antithrombotic Agents Milan Italy. Conference Start: 20100706 Conference End: 20100709. Conference Publication:(var.pagings):A29.
179. Tosetto A, Lorio A, Marcucci M, Baglin T, Cushman M, Eichinger S, et al. Clinical prediction guide to predict thrombosis recurrence after a first unprovoked venous thromboembolism. Journal of Thrombosis and Haemostasis. 2009;Conference: 22nd Congress of the International Society of Thrombosis and Haemostasis Boston, MA United States. Conference Start: 20090711 Conference End: 20090716. Conference Publication:(var.pagings):266.
180. Eichinger S, Heinze G, Kyrle PA. D-Dimer levels over time and the risk of recurrent venous thromboembolism: An update of the Vienna Prediction Model. Journal of Thrombosis and HaemostasisConference: 24th Congress of the International Society on Thrombosis and Haemostasis Amsterdam NetherlandsConference Start: 20130629 Conference End: 20130704Conference Publication: (varpagings)11 (pp 115), 201. 2013(var.pagings):115.
181. Eichinger S, Heinze G, Kyrle PA. D-dimer levels over time and the risk of recurrent venous thromboembolism: An update of the vienna prediction model. BloodConference: 55th Annual Meeting of the American Society of Hematology, ASH 2013 New Orleans, LA United StatesConference Start: 20131207 Conference End: 20131210Conference Publication: (varpagings)122 (21) , 2013Date of Publication: 21 Oct. 2014(var.pagings).

182. Eichinger S, Heinze G, Kyrle PA. D-dimer levels over time and the risk of recurrent venous thromboembolism: An update of the vienna prediction model. Vasa - Journal of Vascular DiseasesConference: 16th Tri-Country Congress of the Austrian, German and Swiss Society of Angiology 2013 Graz AustriaConference Start: 20130915 Conference End: 20130918Conference Publication: (varpagings)42 (pp 36), 20. 2014(var.pagings):36-Journal.
183. The development and evaluation of a prognostic model and clinical decision rule to help decide on cessation of anticoagulant therapy in patients with idiopathic venous thromboembolism (VTE) (Project record). Health Technology Assessment Database. 2014.
184. Eichinger S, Heinze G, Kyrle PA. D-dimer levels over time and the risk of recurrent venous thromboembolism: an update of the Vienna prediction model. Journal of the American Heart Association. 2014;3(1):e000467.
185. Romualdi E, Donadini MP, Ageno W. Oral rivaroxaban after symptomatic venous thromboembolism: the continued treatment study (EINSTEIN-extension study). Expert review of cardiovascular therapy. 2011;9(7):841-4.
186. Douketis JD, Ginsberg JS, Holbrook A, Crowther M, Duku EK, Burrows RF. A reevaluation of the risk for venous thromboembolism with the use of oral contraceptives and hormone replacement therapy. Arch Intern Med. 1997;157(14):1522-30.
187. Altman DG. Prognostic models: a methodological framework and review of models for breast cancer. 2009;2009/03/18:235-43.
188. Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. Journal of clinical epidemiology. 1995;48(12):1503-10.
189. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. 1996;49:1373-9.
190. van Smeden M, de Groot JA, Moons KG, Collins GS, Altman DG, Eijkemans MJ, et al. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. BMC medical research methodology. 2016;16(1):163.
191. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. Statistics in medicine. 2016;35(2):214-26.
192. McRae S, Tran H, Schulman S, Ginsberg J, Kearon C. Effect of patient's sex on risk of recurrent venous thromboembolism: a meta-analysis. Lancet (London, England). 2006;368(9533):371-8.
193. Douketis J, Tosetto A, Marcucci M, Baglin T, Cosmi B, Cushman M, et al. Risk of recurrence after venous thromboembolism in men and women: patient level meta-analysis. [Review]. BMJ (Clinical research ed). 2011;342:d813.
194. Thachil J, Fitzmaurice DA, Toh CH. Appropriate use of D-dimer in hospital patients. The American journal of medicine. 2010;123(1):17-9.
195. Schulman S, Lindmarker P, Holmstrom M, Larfars G, Carlsson A, Nicol P, et al. Post-thrombotic syndrome, recurrence, and death 10 years after the first episode of venous thromboembolism treated with warfarin for 6 weeks or 6 months. Journal of Thrombosis & Haemostasis. 2006;4(4):734-42.
196. Douketis J, Tosetto A, Marcucci M, Baglin T, Cushman M, Eichinger S, et al. Are men at higher risk for disease recurrence than women. Pathophysiology of Haemostasis and Thrombosis. 2010;Conference: 21st International Congress on Thrombosis - The Start of a New Era Antithrombotic Agents Milan Italy. Conference Start: 20100706 Conference End: 20100709. Conference Publication:(var.pagings):A30.
197. Palareti G, Legnani C, Cosmi B, Guazzaloca G, Pancani C, Coccheri S. Risk of venous thromboembolism recurrence: high negative predictive value of D-dimer performed after oral anticoagulation is stopped. Thrombosis & Haemostasis. 2002;87(1):7-12.

198. Baglin T, Palmer CR, Luddington R, Baglin C. Unprovoked recurrent venous thrombosis: prediction by D-dimer and clinical risk factors. Journal of Thrombosis & Haemostasis. 2008;6(4):577-82.
199. Cosmi B. Value of D-dimer testing to decide duration of anticoagulation after deep vein thrombosis: yes. Journal of Thrombosis & Haemostasis. 2006;4(12):2527-9.
200. Cosmi B, Legnani C, Pengo V, Tosetto A, Ghirarduzzi A, Alatri A, et al. D-dimer and sex as risk factors for recurrence after a first episode of venous thromboembolism in the extended follow-up of the prolong study. Journal of Thrombosis and Haemostasis. 2009;Conference: 22nd Congress of the International Society of Thrombosis and Haemostasis Boston, MA United States. Conference Start: 20090711 Conference End: 20090716. Conference Publication:(var.pagings):416.
201. Douketis J. D-dimer can predict risk of recurrent venous thromboembolism regardless of patient age, timing of testing, or characteristics of assay. Journal of Clinical Outcomes Management. 2011;18(6):246-8.
202. Douketis J, Tosetto A, Marcucci M, Baglin T, Cushman M, Eichinger S, et al. D-dimer to determine risk for disease recurrence after unprovoked venous thromboembolism: Addressing unanswered questions with a large individual patient meta-analysis. Pathophysiology of Haemostasis and Thrombosis. 2010;Conference: 21st International Congress on Thrombosis - The Start of a New Era Antithrombotic Agents Milan Italy. Conference Start: 20100706 Conference End: 20100709. Conference Publication:(var.pagings):A46.
203. Eichinger S, Minar E, Bialonczyk C, Hirschl M, Quehenberger P, Schneider B, et al. D-dimer levels and risk of recurrent venous thromboembolism. 2003;2003/08/28:1071-4.
204. Palareti G, Cosmi B, Legnani C, Tosetto A, Brusi C, Iorio A, et al. D-dimer testing to determine the duration of anticoagulation therapy. 2006;2006/10/27:1780-9.
205. Lambert PC, Royston P. Further development of flexible parametric models for survival analysis. Stata Journal. 2009;9(2):265-90.
206. 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;33(22):3844-58.
207. Burnham KP, Anderson DR. Multimodel inference understanding AIC and BIC in model selection. Sociological methods & research. 2004;33(2):261-304.
208. Spencer FA, Gore JM, Lessard D, Emery C, Pacifico L, Reed G, et al. Venous thromboembolism in the elderly. A community-based perspective. Thromb Haemost. 2008;100(5):780-8.
209. Cushman M, Folsom AR, Wang L, Aleksic N, Rosamond WD, Tracy RP, et al. Fibrin fragment D-dimer and the risk of future venous thrombosis. Blood. 2003;101(4):1243-8.
210. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. 2009;2009/07/01:b2393.
211. Koopman L, van der Heijden GJ, Grobbee DE, Rovers MM. Comparison of methods of handling missing data in individual patient data meta-analyses: an empirical example on antibiotics in children with acute otitis media. 2008;2008/01/11:540-5.
212. Harrell FE, Jr., Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. Jama. 1982;247(18):2543-6.
213. Harrell FE, Jr., Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. Statistics in medicine. 1984;3(2):143-52.
214. Poli D, Antonucci E, Ciuti G, Abbate R, Prisco D. Combination of D-dimer, F1+2 and residual vein obstruction as predictors of VTE recurrence in patients with first VTE episode after OAT withdrawal. Journal of Thrombosis & Haemostasis. 2008;6(4):708-10.
215. Vergouwe Y, Steyerberg EW, Eijkemans MJ, Habbema JD. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. Journal of clinical epidemiology. 2005;58(5):475-83.

216. Shrivastava S, Ridker PM, Glynn RJ, Goldhaber SZ, Moll S, Bounameaux H, et al. D-dimer, factor VIII coagulant activity, low-intensity warfarin and the risk of recurrent venous thromboembolism. Journal of Thrombosis & Haemostasis. 2006;4(6):1208-14.
217. Palareti G, Legnani C, Cosmi B, Valdre L, Lunghi B, Bernardi F, et al. Predictive value of D-dimer test for recurrent venous thromboembolism after anticoagulation withdrawal in subjects with a previous idiopathic event and in carriers of congenital thrombophilia. Circulation. 2003;108(3):313-8.
218. Tait R, Lowe GDO, McColl MD, McMahon AD, Robertson L, King L. Predicting risk of recurrent venous thrombosis using a 5-point scoring system including fibrin D-dimer. Journal of Thrombosis and Haemostasis. 2007;5(Supplement 2 (O-M)):060.
219. Fattorini A, Crippa L, Vigano DAS, Pattarini E, D'Angelo A. Risk of deep vein thrombosis recurrence: high negative predictive value of D-dimer performed during oral anticoagulation. Thrombosis & Haemostasis. 2002;88(1):162-3.
220. Eichinger S, Hron G, Bialonczyk C, Hirschl M, Minar E, Wagner O, et al. Overweight, obesity, and the risk of recurrent venous thromboembolism. Archives of Internal Medicine. 2008;168(15):1678-83.
221. Heit JA, Mohr DN, Silverstein MD, Petterson TM, O'Fallon WM, Melton LJ, III. Predictors of recurrence after deep vein thrombosis and pulmonary embolism: a population-based cohort study. Archives of Internal Medicine. 2000;160(6):761-8.
222. Look MP, van Putten WL, Duffy MJ, Harbeck N, Christensen IJ, Thomssen C, et al. Pooled analysis of prognostic impact of urokinase-type plasminogen activator and its inhibitor PAI-1 in 8377 breast cancer patients. Journal of the National Cancer Institute. 2002;94(2):116-28.
223. Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. American journal of epidemiology. 2007;165(6):710-8.
224. 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(9):993-1000.
225. StataCorp. Stata Statistical Software: Release 14. College Station, TX: StataCorp LP 2015.
226. 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. 2016.
227. Chen H-l, Zhou M-q, Tian W, Meng K-x, He H-f. Effect of Age on Breast Cancer Patient Prognoses: A Population-Based Study Using the SEER 18 Database. PLOS ONE. 2016;11(10):e0165409.
228. Macaskill P, Gatsonis C, Deeks JJ, Harbord RM, Takwoingi Y. Chapter 10: Analysing and Presenting Results. In: Deeks JJ BP, Gatsonis C, editor. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 10: The Cochrane Collaboration.; 2010.
229. Korevaar DA, Ochodo EA, Bossuyt PM, Hooft L. Publication and reporting of test accuracy studies registered in ClinicalTrials.gov. Clinical chemistry. 2014;60(4):651-9.
230. Dukic V, Gatsonis C. Meta-analysis of diagnostic test accuracy assessment studies with varying number of thresholds. Biometrics. 2003;59(4):936-46.
231. Hamza TH, Arends LR, van Houwelingen HC, Stijnen T. Multivariate random effects meta-analysis of diagnostic tests with multiple thresholds. 2009;2009/11/12:73.
232. Putter H, Fiocco M, Stijnen T. Meta-analysis of diagnostic test accuracy studies with multiple thresholds using survival methods. Biometrical journal Biometrische Zeitschrift. 2010;52(1):95-110.
233. Martinez-Camblor P. Fully non-parametric receiver operating characteristic curve estimation for random-effects meta-analysis. Statistical methods in medical research. 2014.
234. Riley RD, Takwoingi Y, Trikalinos T, Guha A, Biswas A, Ensor J, et al. Meta-analysis of test accuracy studies with multiple and missing thresholds: a multivariate-normal model. J Biomed Biostat. 2014;5:100196.

235. Steinhauser S, Schumacher M, Rucker G. Modelling multiple thresholds in meta-analysis of diagnostic test accuracy studies. BMC medical research methodology. 2016;16(1):97.
236. Riley RD, Ahmed I, Ensor J, Takwoingi Y, Kirkham A, Morris RK, et al. Meta-analysis of test accuracy studies: an exploratory method for investigating the impact of missing thresholds. Systematic Reviews. 2015;4:12.
237. Gopalakrishna G, Langendam M, Scholten R, Bossuyt P, Leeflang M, Noel-Storr A, et al. Methods for Evaluating Medical Tests and Biomarkers. Diagnostic and Prognostic Research. 2017;1(1):7.
238. Chu H, Cole SR. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. 2006;2006/11/14:1331-2.
239. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. 2005;2005/09/20:982-90.
240. Riley RD, Abrams KR, Sutton AJ, Lambert PC, Thompson JR. Bivariate random-effects meta-analysis and the estimation of between-study correlation. 2007;2007/01/16:3.
241. Takwoingi Y, Guo B, Riley RD, Deeks JJ. Performance of methods for meta-analysis of diagnostic test accuracy with few studies or sparse data. Statistical methods in medical research. 2015.
242. Malin GL. The diagnostic/ prognostic value of neonatal findings for predicting childhood and adult morbidity: systematic reviews, meta-analysis and decision analytic modelling: University of Birmingham; 2013.
243. Apgar V. A proposal for a new method of evaluation of the newborn infant. . Current Researches in Anesthesia & Analgesia. 1953(32):260–7.
244. American Academy of P. The APGAR score. Advances in Neonatal Care. 2006;6(4):220-3.
245. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. Statistics in medicine. 2006;25(24):4279-92.
246. Little RJA, Rubin DB. Statistical analysis with missing data: John Wiley & Sons; 2014.
247. Youden WJ. Index for rating diagnostic tests. Cancer. 1950;3(1):32-5.
248. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. Journal of clinical epidemiology. 2005;58(9):882-93.
249. Begg CB, Berlin JA. Publication bias: a problem in interpreting medical data. Journal of the Royal Statistical Society Series A (Statistics in Society). 1988:419-63.
250. Begg CB, Berlin JA. Publication bias and dissemination of clinical research. Journal of the National Cancer Institute. 1989;81(2):107-15.
251. Abramowitz M, Stegun IA. Handbook of mathematical functions: with formulas, graphs, and mathematical tables: Courier Corporation; 1964.
252. Cook JA, Collins GS. The rise of big clinical databases. Br J Surg. 2015;102(2):e93-e101.
253. Dretzke J, Ensor J, Bayliss S, Hodgkinson J, Lordkipanidze M, Riley RD, et al. Methodological issues and recommendations for systematic reviews of prognostic studies: an example from cardiovascular disease. Syst Rev. 2014;3:140.
254. 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):772-89.
255. Riley RD, Ensor J, Jackson D, Burke DL. Deriving percentage study weights in multi-parameter meta-analysis models: with application to meta-regression, network meta-analysis and one-stage individual participant data models. Statistical methods in medical research. 2017:0962280216688033.

256. Steyerberg EW, Harrell FE, Jr. Prediction models need appropriate internal, internal-external, and external validation. Journal of clinical epidemiology. 2016;69:245-7.
257. Collins GS, Moons KG. Comparing risk prediction models. BMJ (Clinical research ed). 2012;344:e3186.
258. Rodger MA, Le Gal G, Wells P, Baglin T, Aujesky D, Righini M, et al. Clinical decision rules and D-Dimer in venous thromboembolism: current controversies and future research priorities. Thrombosis research. 2014;134(4):763-8.
259. Riley RD. Commentary: like it and lump it? Meta-analysis using individual participant data. 2010;2010/07/28:1359-61.
260. Riley RD, Lambert PC, bo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. 2010;2010/02/09:c221.
261. Stewart LA, Tierney JF. To IPD or not to IPD? Advantages and disadvantages of systematic reviews using individual patient data. 2002;2002/03/01:76-97.
262. Askie LM, Baur LA, Campbell K, Daniels LA, Hesketh K, Magarey A, et al. The Early Prevention of Obesity in CHildren (EPOCH) Collaboration--an individual patient data prospective meta-analysis. BMC Public Health. 2010;10:728.
263. Pitcher A, Emberson J, Lacro RV, Sleeper LA, Stylianou M, Mahony L, et al. Design and rationale of a prospective, collaborative meta-analysis of all randomized controlled trials of angiotensin receptor antagonists in Marfan syndrome, based on individual patient data: A report from the Marfan Treatment Trialists' Collaboration. Am Heart J. 2015;169(5):605-12.
264. Baigent C, Keech A, Kearney PM, Blackwell L, Buck G, Pollicino C, et al. Efficacy and safety of cholesterol-lowering treatment: prospective meta-analysis of data from 90,056 participants in 14 randomised trials of statins. Lancet (London, England). 2005;366(9493):1267-78.
265. Reade MC, Delaney A, Bailey MJ, Harrison DA, Yealy DM, Jones PG, et al. Prospective meta-analysis using individual patient data in intensive care medicine. Intensive Care Med. 2010;36(1):11-21.
266. Ioannidis J. Next-generation systematic reviews: prospective meta-analysis, individual-level data, networks and umbrella reviews. Br J Sports Med. 2017.
267. 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. 2008;27:1870-93.
268. Riley RD, Simmonds MC, Look MP. Evidence synthesis combining individual patient data and aggregate data: a systematic review identified current practice and possible methods. 2007;2007/04/11:431-9.
269. 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;12:56.
270. Debray TP, Riley RD, Rovers MM, Reitsma JB, Moons KG, Cochrane IPDM-aMg. Individual participant data (IPD) meta-analyses of diagnostic and prognostic modeling studies: guidance on their use. PLoS Med. 2015;12(10):e1001886.
271. Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, van Staa T, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). Int J Epidemiol. 2015;44(3):827-36.
272. Douketis J, Iorio A, Marcucci M, Baglin T, Cushman M, Eichinger S, et al. Does the clinical presentation of venous thromboembolism predict the risk for and type of thrombosis recurrence? Journal of Thrombosis and Haemostasis. 2009;Conference: 22nd Congress of the International Society of Thrombosis and Haemostasis Boston, MA United States. Conference Start: 20090711 Conference End: 20090716. Conference Publication:(var.pagings):723-4.