DEVELOPMENT AND APPLICATION OF STATISTICAL METHODS FOR PROGNOSIS RESEARCH

By

KYM IRIS ERIKA SNELL

A thesis submitted to the University of Birmingham

for the degree of

DOCTOR OF PHILOSOPHY

School of Health and Population Sciences

University of Birmingham

May 2015

ABSTRACT

A pivotal component of prognosis research is the prediction of future outcome risk. This thesis applies, develops and evaluates novel statistical methods for development and validation of risk prediction (prognostic) models. In the first part, a literature review of published prediction models shows that the Cox model remains the most common approach for developing a model using survival data; however, this avoids modelling the baseline hazard and therefore restricts individualised predictions. Flexible parametric survival models are shown to address this by flexibly modelling the baseline hazard, thereby enabling individualised risk predictions over time. Clinical application reveals discrepant mortality rates for different hip replacement procedures, and identifies common issues when developing models using clinical trial data.

In the second part, univariate and multivariate random-effects meta-analyses are proposed to summarise a model’s performance across multiple validation studies. The multivariate approach accounts for correlation in multiple statistics (e.g. C-statistic and calibration slope), and allows joint predictions about expected model performance in applied settings. This allows competing implementation strategies (e.g. regarding baseline hazard choice) to be compared and ranked. A simulation study also provides recommendations for the scales on which to combine performance statistics to best satisfy the between-study normality assumption in random-effects meta-analysis.

ACKNOWLEDGEMENTS

This PhD was funded by the MRC Midland Hub for Trials Methodology Research, without which I would not have been able to complete this research.

I would like to express my gratitude to my supervisors, Prof. Richard Riley and Prof. Lucinda Billingham. Richard, I consider myself so lucky to have had the opportunity to work with you. You have always believed in me, built up my confidence, sent opportunities my way and helped me find my passion and shape my career. I cannot truly thank you enough for the endless support and guidance you have offered.

I would also like to thank the following people I have been lucky enough to work with: Thomas Debray for input and feedback on several chapters, Joie Ensor for all the discussions and feedback on chapters, Jon Deeks for his support in the last few months leading up to submission, and not forgetting Karen Biddle and Anne Walker for helping with anything and everything that they could. Thanks also go to my colleagues in Health and Population Sciences for all their encouragement.

To my family, thank you for all your love and support over the years. Mum, Dad, Matt, Jay and Dawn, thanks for always believing in me and supporting me in everything I do. I am so lucky to have the family I do and words cannot express how much I love you all.

Last but certainly not least, thanks go to my friends, old and new: Lozz, Hannah, Ruby, Elena, and the biggest thanks of all to a friend that has been there every single day of this journey, Dani. Since we started our PhDs together, you have been the best friend and my biggest support. We’ve been through it all together, you’ve been there to celebrate the highs and help me through the lows. You have made the last three and a half years an amazing experience. I am so thankful to have you in my life and proud of all we have achieved.

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

TABLE OF BOXES

ABBREVIATIONS

AIC	Akaike information criterion
AUC	Area under the (ROC) curve
BHR	Birmingham hip resurfacing
BIC	Bayesian information criterion
CI	Confidence interval
d.f.	Degrees of freedom
DVT	Deep vein thrombosis
E/O	Expected/observed number or proportion of events (performance statistic)
EPV	Events per variable
FP(1/2)	fractional polynomials (of first/second degree)
HR	Hazard ratio
IECV	Internal-external cross-validation
IPD	Individual participant data
IQR	Interquartile range
LP	Linear predictor
MFP	multivariable fractional polynomials
MI	Multiple imputation
OR	Odds ratio
RCT	Randomised controlled trial
REML	Restricted maximum likelihood
ROC	Receiver operator curve
SD	Standard deviation
SE	Standard error
THR	Total hip replacement

CHAPTER 1: INTRODUCTION

1.1 Introduction to research area

A pivotal component of clinical research is understanding and predicting future outcome risk in those with existing disease, to tailor treatment strategies and inform patient counselling. Prognosis research is crucial for this purpose, and a recent series outlined a framework of four types of prognosis research studies.[1-4] The most well-known is **clinical prediction modelling**, which has become more popular over time.[5] A clinical prediction model is a statistical model that is used to predict the risk of a defined outcome or event. There are different types of clinical prediction models depending on the aim of the model and the group of patients for which the model is intended. Two types of clinical prediction models are diagnostic and prognostic models. A **diagnostic model** is used to predict the probability of having a particular disease or condition in patients that are suspected of having the disease or condition. For example, Wells et al. published a model to help predict the probability of having a pulmonary embolism (PE) in patients suspected of having a PE.[6] The tool was developed to help identify patients that are unlikely to have a PE and therefore do not require further testing. A **prognostic model**, on the other hand, aims to predict the probability of a particular future outcome or event in patients that have a certain condition or disease of interest.[1] An example of a prognostic model developed for patients with advanced epithelial ovarian cancer is PIEPOC (Prognostic Index of EPithelial Ovarian Cancer) which is used to predict the 5-year probability of overall survival and classifies patients into low, intermediate or high risk of mortality.[7]

This thesis is focused on the application and development of statistical methods for prognosis research, with particular emphasis on the use of prognostic models for clinical prediction. The importance of prognosis research is being increasingly recognised,[1] especially with regard to the development and validation of prognostic models that are both clinically useful and applicable in clinical practice. However, current evidence suggests that prognosis research is limited by poor research standards and inadequate reporting, and methodology work is needed to improve the development and external validation of prognostic models to ensure they are robust for use in the clinical settings of intended use.[1] This thesis therefore aims to apply and develop novel methodology for prognosis and prognostic model research, in order to overcome some of the many challenges facing this field.

1.2 What is prognosis research?

Prognosis research aims to understand, explain and predict future outcomes in patients with an existing disease or health condition. It is an important area of clinical research for many reasons.[1] In particular it aims to identify measurable variables (such as biomarkers) whose values are associated with the risk of a future outcome of interest. In prognosis research, these variables may be referred to as prognostic factors, prognostic markers or predictors, and these names are often used interchangeably. Using multiple prognostic factors in conjunction, a multivariable prognostic model (a prognostic model with more than one variable in it) can be developed to make individualised predictions about a patient's outcome risk based on their own set of prognostic factor values. This helps patients understand the course of disease and their likely outcome, and helps clinical decision-making in terms of treatment selection for individual patients or for groups of patients with different prognoses.[8-10]

Prognosis research needs a clearly defined start point and patient population, and a clearly defined outcome (or outcomes) of interest. The start point for follow-up should be similar for all patients, for example recruiting patients soon after diagnosis of a particular disease and preferably at the same stage of disease.[11] The patient population that will be used to collect data for model development (and in which the model is intended to be used) needs to be well-defined, for example, a specific age group, or only including patients with advanced disease. The outcome of interest also needs to be clearly defined and is usually an event such as death (does the patient die, yes or no) or disease progression (does the patient's disease advance, yes or no). Outcomes in prognosis research are often binary and patients recruited into studies are usually followed up over time to observe if and when the event of interest occurs.

Prognostic studies often develop models using cohort data, in which patients are followed up over time to observe the event of interest. Cohorts can be retrospectively collected from one or more centres, for example by identifying patients with a particular condition or disease from hospital records. This type of study design is simple and low cost but may be subject to selection bias if information is incomplete.[12] Single centre studies may be limited to small sample sizes. Alternatively, prospective cohort studies could be run across multiple centres recruiting patients from a clearly defined group for follow-up. Prospective studies give the opportunity to collect information on any potential prognostic factors thought to be of interest. However, due to strict inclusion and exclusion criteria, any model developed may not be generalizable to wider patient populations.[12] For example, a model developed in patients with advanced disease may not be generalizable to patients with less advanced disease. Data may be collected prospectively for purposes other than developing a prognostic model, for example a randomised clinical trial with different treatment arms. Although prognosis research may not be the primary aim for collection of the data, it provides a rich opportunity to develop prognostic models for patients with the disease of interest. Chapter 4 aims to highlight some of the challenges using randomised control trial (RCT) data to develop a prognostic model. Other sources of data include registries (for example, cancer registries) and e-health record databases such as QResearch,[13] which offer large sample sizes and broader coverage of patient populations across many centres such as GP practices.

Below is an example of a prognostic model including details of the data such as the start and end points as well as a description of the model itself to illustrate the concepts above.

Example of a prognostic model:

The Nottingham Prognostic Index is a well-known and widely used prognostic model that was first published in 1982 by Haybittle et al.[14] The model was developed using retrospective data from 387 patients that had primary operable breast cancer (start point) and were followed up over five years for the event of death (outcome). Patients all had a simple mastectomy and triple-node biopsy and had all been treated by the same surgeon at Nottingham City Hospital. There were eight candidate predictors considered for inclusion in a multivariable Cox model. The authors fitted a multivariable model with all candidate predictors in it and then retained only those that were significant at the 5% level for the final model. This was simplified further (using a scaling factor to reduce the number of *β*-coefficients that were not equal to 1) to arrive at the final index,

$$\text{NPI} = 0.2 \times \text{size} + \text{lymph-node stage} + \text{tumour grade}$$

where size is measured in cm, lymph-node stage is coded as 1=A, 2=B, 3=C and tumour grade is coded as 1=I, 2=II, 3=III. As the value of the index increases, the prognosis for a patient gets worse. The index was classified as high (>4.4), medium (2.8 to 4.4) or low (<2.8) and survival curves were plotted for the different prognostic groups.

The NPI was later validated by applying it prospectively and following patients for up to six and a half years.[15] The model has also been tested in slightly different settings such as in women with small invasive breast cancers or in younger women with primary breast cancer (aged under 40 years at presentation).[16,17]

1.2.1 Framework for prognosis research

The development and validation of prognostic models is just one aspect of prognosis research. Although there is a lot of activity in prognosis research, the research is often of poor quality and has little impact on clinical practice.[18] Over the last few years, there has been a collaborative effort to provide more structure to the field and offer recommendations for improvement and areas that require further research.[1,8]

In 2013 a series of four articles were published by the PROGRESS (PROGnosis RESearch Strategy) partnership to improve prognosis research.[1-4] These articles provide a foundation for researchers interested in prognosis research and provide recommendations for overcoming some of the challenges in this field. PROGRESS started by proposing a framework which is given below,[1] and each of the four papers in the series focused on one of the research themes.

1 **Fundamental prognosis research**: Studying the natural course of a disease or condition with current clinical practice.[1]

2 **Prognostic factor research**: Studying individual factors that are associated with the outcome.[2]

3 **Prognostic model research**: The use of multiple prognostic factors to develop a model from which risk of the outcome can be predicted for individuals. This

research theme also includes validation and assessing the clinical impact of the prognostic model.[3]

4 **Stratified medicine research**: The use of prognostic information to make treatment decisions that are personalised to individuals.[4]

The content of this thesis mainly fits into research theme three (prognostic model research) as it will explore and apply methods for developing and validating multivariable prognostic models. An earlier series of four papers on 'Prognosis and prognostic research' from 2009 focused more on prognostic model research (theme 3). The first paper provided an introduction to prognostic research and its importance;[8] the other papers covered topics on developing a prognostic model,[19] validating a prognostic model,[20] and lastly the application and clinical impact of prognostic models.[21]

The next two sections introduce the fundamental statistical concepts of prognostic model research. In particular, the statistical models used to develop clinical prediction models are described for outcomes that are generally observed over a short follow-up and where follow-up is complete for all patients (logistic regression) and also for outcomes that require longer follow-up to observe events and in which follow-up duration differs for different patients (survival analysis).

1.3 Logistic regression

In clinical prediction studies, the outcome of interest is often binary. Logistic regression models are most often used to model binary outcomes when the duration of follow-up is relatively short.[3]

A logistic regression model can be written as

$$\text{logit}(p_i)=\alpha + \boldsymbol{\beta x_i} \qquad \textbf{(1.1)}$$

where the event probability for patient *i* (p_i=P(Y=1) for the binary outcome Y = 0 or 1) is modelled using a logit transformation, α is the estimated intercept, $\boldsymbol{x_i}=(x_1, x_2, x_3, \ldots)^T$ is the vector of predictor values for individual *i* and $\boldsymbol{\beta}=(\beta_1, \beta_2, \beta_3, \ldots)$ is the vector of coefficients (log odds ratios) estimated for the predictors in the model ($\boldsymbol{\beta x_i}=\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\ldots$). The logit transformation is

$$\text{logit}(p_i)=\log\left(\frac{p_i}{1-p_i}\right) \qquad \textbf{(1.2)}$$

Probabilities can only take values between 0 and 1 which is why the logit transformation is used. If the untransformed probabilities were modelled, fitted values and predictions could easily fall above one or below zero [22] However, using the logit function (Figure 1.1) means that modelled probabilities and future predictions remain between the bounds of 0 and 1.

The model (1.1) can be estimated using maximum likelihood to give the fitted model,

$$\text{logit}(\hat{p}_i)=\hat{\alpha} + \boldsymbol{\hat{\beta} x_i}$$

Odds ratios (OR) are typically reported for each variable in a logistic regression model instead of the $\hat{\beta}$-coefficients and can be calculated as $\exp(\hat{\beta}_j)$ for any variable *j* in the model.

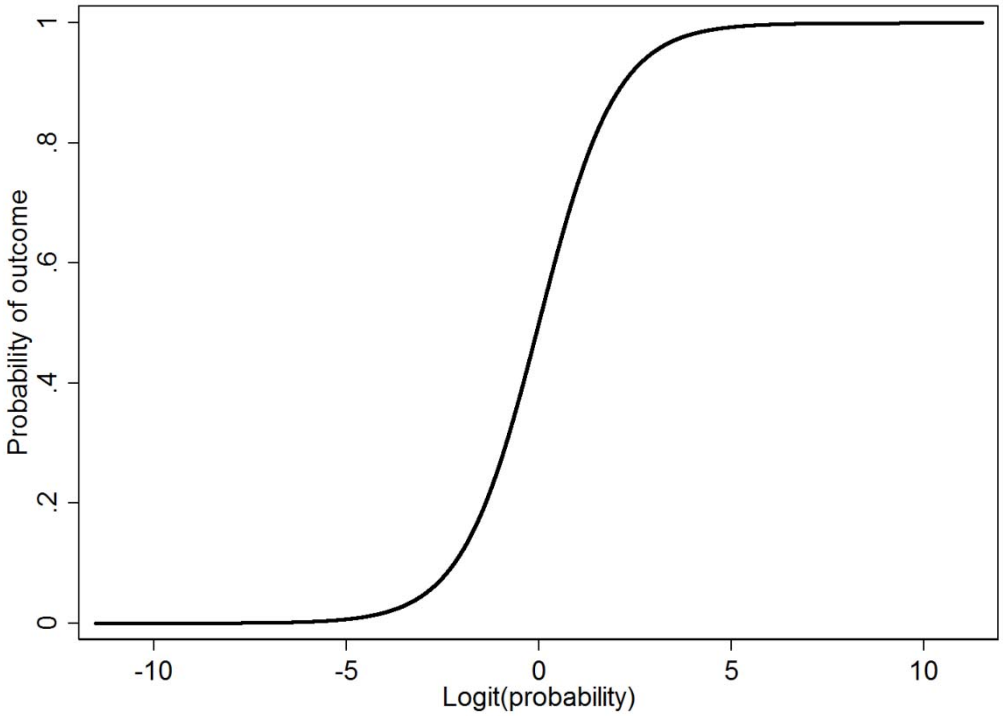


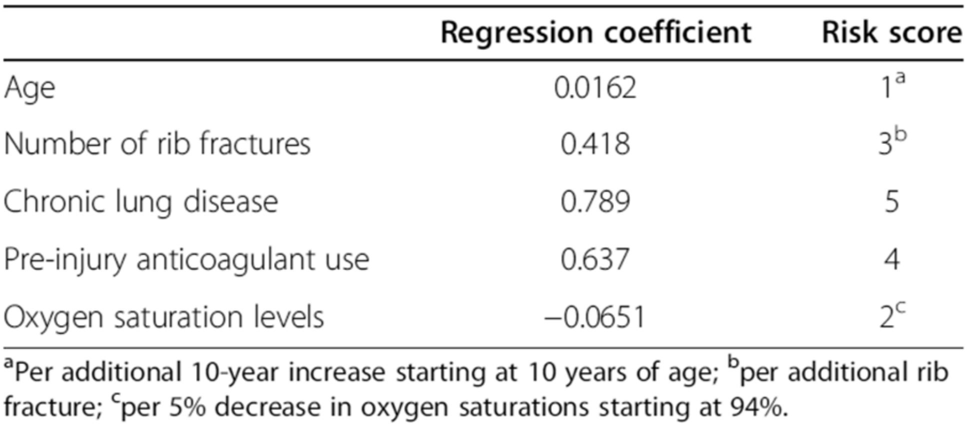
Figure 1.1: Logit function used to model the event probability in a logistic regression model.

For the purpose of predicting outcome probabilities in patients, the inverse transformation is applied to the estimated linear predictor (LP_i=$\hat{\alpha}$ + $\boldsymbol{\hat{\beta} x_i}$) to get the predicted probability for patient *i*. The inverse transformation to obtain the predicted probability is given by

$$\hat{p}_i = \frac{e^{LP_i}}{1+e^{LP_i}} \qquad \textbf{(1.3)}$$

1.3.1 Example prognostic model developed using logistic regression

A prognostic model was developed to predict the risk of developing complications after blunt chest wall trauma using retrospective data from 274 patients who presented to a single trauma centre in South Wales.[23] The aim of the model was to provide a tool to assist clinicians in the management of patients with blunt chest wall trauma. ‘Complications’ was a composite outcome including in-hospital mortality, morbidity including all pulmonary complications, ICU admission or a hospital stay of seven or more days. The authors developed a multivariable logistic model including five prognostic factors: age, number of rib fractures, chronic lung disease, use of pre-injury anticoagulants and oxygen saturations. They then simplified the model by scaling the regression coefficients to produce an easy to calculate risk score for patients (Figure 1.2). The mean probability of developing complications was reported for risk groups based on the total risk score (calculated by summing the risk scores for each prognostic factor). For example, a risk score of less than 10 has a (mean) predicted probability of complications of 13%, whereas a risk score of 31 or more has a (mean) predicted probability of complications of 88%.



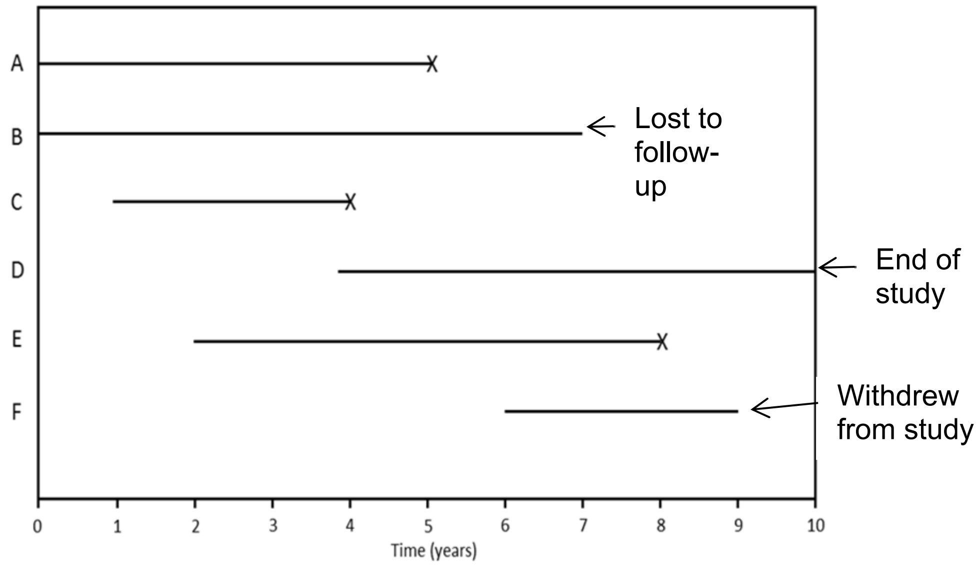
	Regression coefficient	Risk score
Age	0.0162	1[a]
Number of rib fractures	0.418	3[b]
Chronic lung disease	0.789	5
Pre-injury anticoagulant use	0.637	4
Oxygen saturation levels	−0.0651	2[c]

[a]Per additional 10-year increase starting at 10 years of age; [b]per additional rib fracture; [c]per 5% decrease in oxygen saturations starting at 94%.

Figure 1.2: Regression coefficients and risk scores for patients developing complications after blunt wall chest trauma. Original article by Battle et al.[23] Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0).

1.4 Survival analysis

In many diseases or conditions in prognosis research, many events may not be expected within a short time frame or the event may not occur at all for some patients. For example, consider the event of death following diagnosis of breast cancer. Patients would need to be followed up for many years to gain a sufficient number of events and many would not die within the study duration. Therefore, the time until the event occurs is generally not normally distributed and may be subject to censoring. There are different types of censoring, namely left, right and interval censoring. Left or interval censoring are less common especially within prognosis research so will not be considered in this thesis. Right censoring is very common in survival studies and a patient is said to be right censored if they choose to withdraw from the study before having the event of interest, they are lost to follow-up, or if the study ends prior to observing the event for that patient (Figure 1.3). In all these cases, patients are censored at the last time of follow-up where they were known to not have had the event of interest, and censoring is independent to the event occurrence. For right censored data, the data can only be analysed at a particular time point *t* (snapshot analysis) using logistic regression if the outcome of interest is binary (such as in the case of mortality – alive or dead at time point *t*) and if censoring does not occur before time *t*. Therefore survival analysis methods are usually preferred to logistic regression in prognostic model research, as censoring usually occurs in practice unless the follow-up period is short.



X = Event

Figure 1.3: Example of survival data with right censoring.

Censored patients still provide valuable information as it is known that they had not had the event of interest up until the time they were censored. Ideally, this information should not be wasted by excluding these patients from analyses. Survival analysis methods allow for censored patients to be included in the analysis up until the time at which they were censored. Some core concepts for survival methods are now described.

1.4.1 Functions in survival data

There are a few key functions in survival analysis which are described below for random variable T which is the survival time.

Survival function

The survival function $S(t)$ is defined as the probability of an individual surviving to time t or longer.[24] This can be written as

$$S(t) = P(T \geq t) \qquad \textbf{(1.4)}$$

where $0 < t < \infty$. The survival function can be estimated using a non-parametric technique known as the Kaplan-Meier method and then plotted as Kaplan-Meier curves.[25] This is done by estimating the survival probability at each unique event time (further details on calculating survival probabilities can be found in textbooks such as Collett (1994)[24]). The estimated survival probability remains the same until the next event time, creating a step function. An example of a Kaplan-Meier curve is shown below in Figure 1.4, using data from a trial in pancreatic cancer that will be used in Chapter 4. The probability of survival can only decrease over time as more patients experience the event. Kaplan-Meier curves can be used to compare groups (for example, by sex) as a preliminary step in survival analysis before modelling the data.

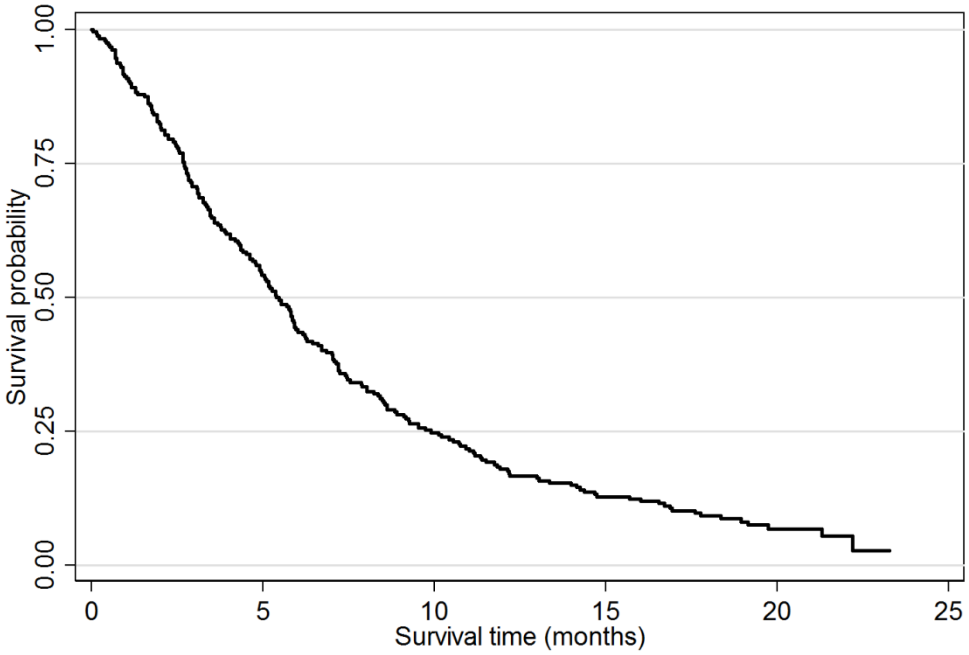


Figure 1.4: Kaplan-Meier curve showing survival probability for a trial in pancreatic cancer.

Probability density function

The probability density function *f*(*t*) is given by the probability of the event occurring at time *t*. This is an unconditional probability and the function will normally be positively skewed. As with all probability density functions, the total area under the curve will equal one.

$$f(t) = \lim_{\delta \to 0} \frac{P(t \leq T < t+\delta)}{\delta} \qquad \textbf{(1.5)}$$

Hazard function

The hazard function $h(t)$ is frequently used in survival analysis and can be written as

$$h(t) = \lim_{\delta \to 0} \frac{P(t \leq T < t+\delta \mid T \geq t)}{\delta} \qquad \textbf{(1.6)}$$

The hazard function at time t is the instantaneous rate of experiencing the event at time t conditional on the event having not already occurred prior to time t.

Cumulative hazard function

The cumulative hazard function $H(t)$ is the total amount of hazard accrued up until time t.[24] It can be written as

$$H(t) = \int_0^t h(u)\, du \qquad \textbf{(1.7)}$$

Relationships between functions

The functions described above are related to each other and can be calculated from one another. For example, survival probability at time t can be calculated by transforming the cumulative hazard function at time t.[24]

$$S(t) = \mathrm{e}^{-H(t)}$$

Inversely the cumulative hazard function can be calculated from the survival function.

$$H(t) = -\ln S(t)$$

And the hazard function can be calculated from the probability density function and survival function,

$$h(t) = \frac{f(t)}{S(t)}.$$

The (cumulative) hazard function is usually the scale used for prognostic modelling and will be described further in the following sections. However, it is often the survival function that is easier to interpret and of interest for prognosis of patients, as it gives the probability of the patient not having the event within a certain time period; whereas the hazard function gives the rate of hazard at time *t* which is harder to interpret and can change over time.

1.4.2 Cox proportional hazard model

The Cox proportional hazards model is a semi-parametric model and can be written as

$$h_i\,(t) = h_0(t)\ e^{\beta x_i} \qquad \textbf{(1.8)}$$

where $h_0(t)$ is the baseline hazard function and this is multiplied by the exponential of the linear predictor made up of the linear combination of estimated *β*-coefficients and variables in the model. The Cox model became popular because it does not make any distributional assumptions for the shape of the baseline hazard function, yet the *β*-coefficients can still be estimated using maximum likelihood of the partial likelihood.[26] A hazard ratio (HR) is usually reported for each variable in the model where HR=exp($\hat{\beta}$).

For prediction, the survival function is more intuitive than the hazard function and can be written by transforming the model in (1.8).

$$S_i(t)=S_0(t)^{e^{\beta x_i}} \qquad \textbf{(1.9)}$$

Proportional hazards assumption

The Cox model, and the other survival models that will be described below, all assume that the hazard functions for all variables in the model are proportional, in other words that the effect of a variable (the HR) remains constant over time. For illustration, consider a simple model with sex as the only predictor (0=female, 1=male) and a HR=2. The assumption is that the hazard rate for males is twice the hazard rate for females at any time point. It is important to check that this assumption holds for all candidate variables. This can be done graphically using 'log-log' plots in which -ln(-ln(S(t)) is plotted against ln(t). Categorical variables can be plotted by category and continuous variables need to be categorised into groups before plotting. Lines should appear approximately parallel if the proportional hazards assumption is valid.[27] Figure 1.5 shows an example log-log plot for white blood cell count in patients with advanced stage pancreatic cancer where the lines are approximately parallel, in an application to be considered within Chapter 4.

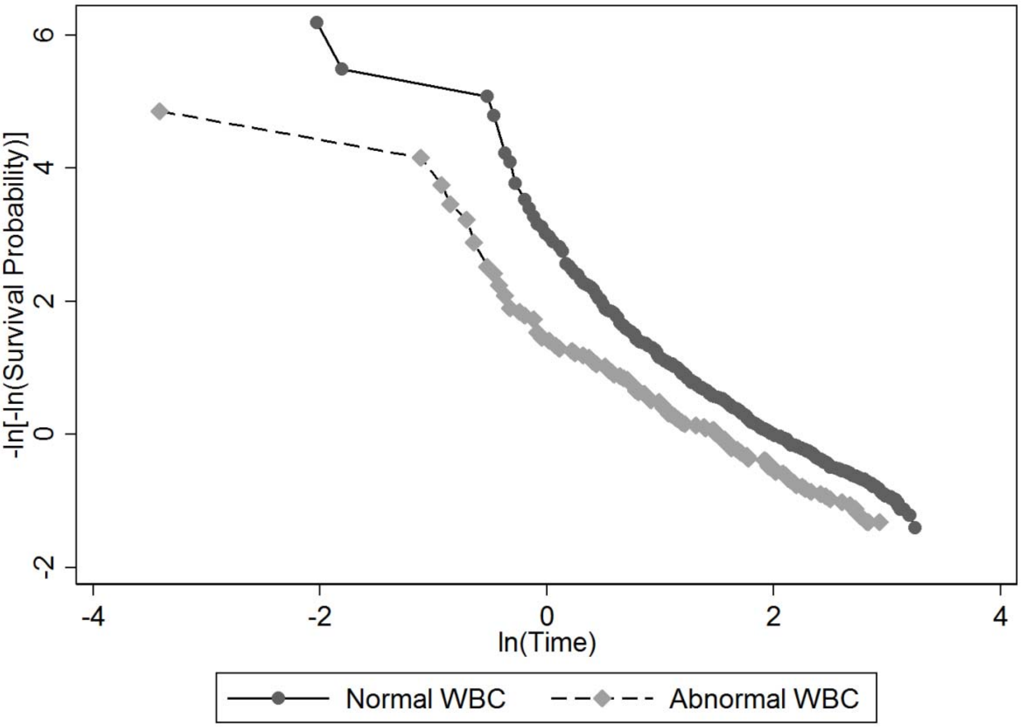


Figure 1.5: Log-log plot for white blood cell count (WBC) in advanced stage pancreatic cancer.

The proportional hazards assumption can also be checked by calculating Schoenfeld residuals.[28] Schoenfeld residuals are partial residuals that do not depend on time, therefore they can be plotted against time to check for any possible relationships. A test for the null hypothesis of zero slope (constant over time) can be performed for individual variables or as a global test for multiple variables in a model.[27,29]

1.4.3 Parametric models

The Cox model makes no distributional assumption for the shape of the baseline hazard, however different distributional assumptions can be used for the baseline hazard. These models are still proportional hazards models but are referred to as parametric models.

There are several distributional shapes that can be used for the hazard function but the exponential and Weibull are perhaps the most common. The simplest parametric model is the exponential model (1.10) which assumes that the hazard function is constant over time. The Weibull model (1.11) is more flexible in shape due to the additional parameter γ and reverts back to the exponential model when γ=1 (see Figure 1.6 for example shapes of the Weibull model). Other parametric models that can be fitted include the Gompertz and log-normal models.

Exponential $$h(t) = \lambda \, e^{\boldsymbol{\beta x_i}} \qquad \textbf{(1.10)}$$

Weibull $$h(t) = \lambda \, \gamma \, t^{\gamma-1} \, e^{\boldsymbol{\beta x_i}} \qquad \textbf{(1.11)}$$

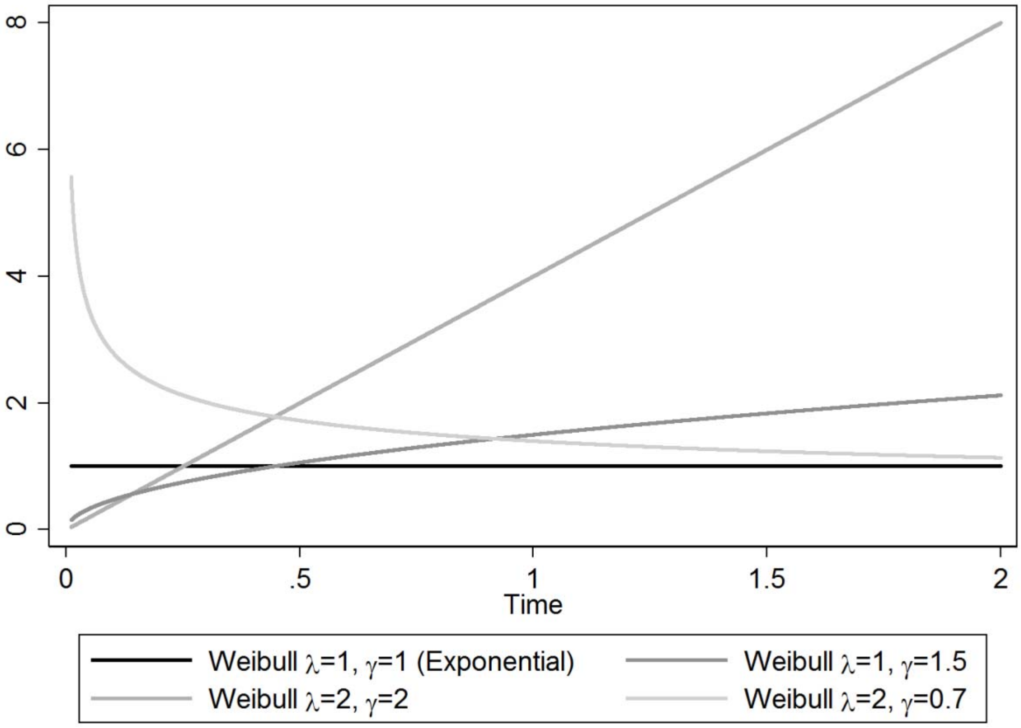


Figure 1.6: Examples of hazard functions using the Weibull distribution.

Parametric models are especially useful for making predictions because the baseline hazard function is explicitly modelled, however they are restricted in the shapes they can take. For example, using a Weibull model only allows monotonic increasing or decreasing functions for the baseline hazard. This often does not fit real data very well as it is possible that the hazard could both rise and fall over time. It can therefore be difficult to choose an appropriate parametric model (if one exists) and requires trying several models which could be quite time consuming. If the baseline hazard is not modelled appropriately, predictions from the model will not be valid and so more flexible parametric models have been proposed.

1.4.4 Flexible parametric models

Flexible parametric models seek to be more flexible in the shape of the baseline hazard function, in order to appropriately capture the shape observed in the data available. Though the idea of flexibly modelling the baseline hazard function using splines had been explored much earlier,[30,31] the use of flexible parametric models have become popularised through so-called Royston-Parmar models, as proposed by Royston and Parmar in 2001,[32] and then extended considerably by Royston and Lambert.[33,34] These authors focused on methodology to explicitly model the baseline hazard function as in parametric models but allow flexibility in the shape it can take by using restricted cubic splines, therefore overcoming the restricted shapes of hazard function in parametric models such as the Weibull model. Examples of baseline hazard functions modelled using flexible parametric models are given in Figure 1.7. These are two examples from data that will be used later in this thesis, but both have turning points in the hazard function that would be difficult to model using standard parametric survival models. Flexible parametric models and modelling the baseline hazard function is the focus of chapters 2 to 4.

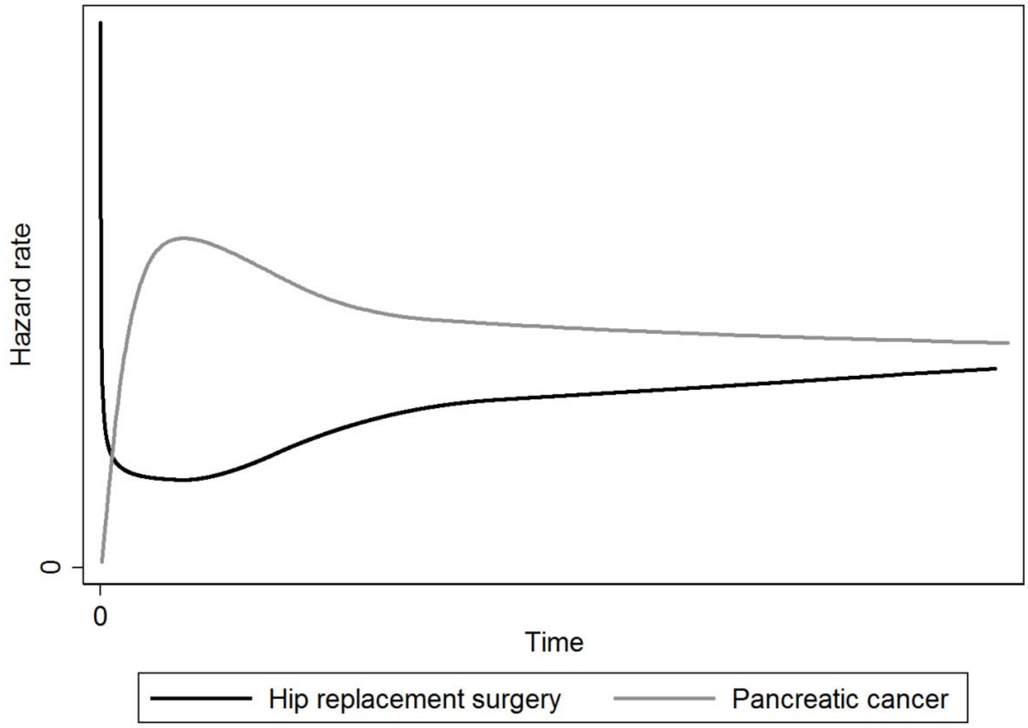


Figure 1.7: Examples of baseline hazard functions modelled using flexible parametric survival models.

Restricted cubic splines

Cubic splines can be used to create flexible and smooth functions that are able to fit tightly curved shapes. This is done by fitting a series of cubic functions and joining them at certain points (called knots). The cubic functions are connected smoothly by ensuring that their first and second derivatives are equal at the knot positions. Cubic splines can be unstable in the tail regions, especially when there is little data in these regions. Therefore restricted cubic splines are used to constrain the function to be linear in the tails (i.e. before the first knot and after the last knot).[22]

Consider having m+2 knots ($k_{\min} < k_1 < k_1 <\ldots< k_m < k_{\max}$) including the boundary knots at various positions of variable x that is to be modelled using restricted cubic splines. In survival analysis, the boundary knots are defined as the minimum and maximum survival times of uncensored observations (first and last event times). The spline function for variable x is given below in (1.12) where new variables are created ($z_1,\ldots,z_{m+1}$) and parameters ($\gamma_0,\ldots,\gamma_{m+1}$) are estimated.[34,35]

$$\text{spline}(x) = \gamma_0+\gamma_1 z_1+\gamma_2 z_2+\ldots+\gamma_{m+1} z_{m+1} \qquad \textbf{(1.12)}$$

and the z variables are calculated as follows:

$$\begin{aligned} z_1 &= x \\ z_j &= (x - k_j)_+^3 - \lambda_j (x - k_1)_+^3 - (1 - \lambda_j)(x - k_m)_+^3 \\ \lambda_j &= \frac{k_{\max} - k_j}{k_{\max} - k_{\min}} \end{aligned}$$

for j=2, …, m+1 and $u_+ = \begin{cases} u \text{ if } u > 0 \\ 0 \text{ if } u \leq 0 \end{cases}$.

Model

Royston-Parmar models are fitted on the log cumulative hazard scale and use ln(time) rather than the original time scale as it is more stable.[33] Therefore the log cumulative hazard function can be written as

$$\ln(H_i(t))=\ln H_0(t)+\boldsymbol{\beta x_i} \qquad \textbf{(1.13)}$$

where $\ln H_0(t)$ is the baseline cumulative hazard function that can be modelled using restricted cubic splines, giving

$$\ln(H_i(t))=\text{spline}(\ln t)+\boldsymbol{\beta x_i} \qquad \textbf{(1.14)}$$

If there are zero knots, spline(ln$t$)= $\gamma_0+\gamma_1\ln t$ which is the baseline hazard for a Weibull model. Royston-Parmar proportional hazards models are a generalisation of the Weibull model.[34]

The degrees of freedom for the baseline hazard function are calculated as d.f. = m +1. Royston and Lambert suggest that 2-3 d.f. are usually adequate to model the baseline hazard function in small datasets and 4-5 d.f. in larger datasets.[34] If the purpose of the model is to predict absolute risk for patients, an adequate number of knots should be used (usually more than 2) to capture the shape of the baseline hazard.[36] However, hazard functions using different d.f. can be fitted and plotted for comparison with each other and with the function obtained through non-parametric estimates. This will give an idea of whether the shape is reasonably captured and also check that it not over-fitted. The authors also suggest comparing AIC and BIC from models using different d.f. for the baseline hazard to help decide how many d.f. are adequate to capture the shape of the baseline hazard.

In Stata, the AIC and BIC are calculated as shown in (1.15) and (1.16) below:

$$\text{AIC}=(-2\times\ln(\textit{likelihood}))+(2\times k) \qquad \textbf{(1.15)}$$

$$\text{BIC}=(-2\times\ln(\text{l}\textit{ikelihood}))+(\ln(N)\times k) \qquad \textbf{(1.16)}$$

where k is the number of parameters estimated in the model and for survival analyses, N is the number of events.[37-39] Both the AIC and BIC are measures of model fit with a penalty for increasing the number of parameters in the model. The difference between them is that the BIC includes a stronger penalty due to the inclusion of the sample size in the calculation. A model with a smaller AIC or BIC is preferred.

The positions of the knots are usually based on centiles of the uncensored (log) event times. For example, if one internal knot is used it will be placed at the 50th percentile, if two internal knots are used they would be placed at the 33rd and 67th centiles. Royston and Lambert say that good fit can be achieved without optimal positioning of knots.[34]

In most cases when the baseline hazard function is modelled adequately, the hazard ratios for variables obtained from a Royston-Parmar model are almost exactly the same as for a Cox model. Therefore, apart from using a few additional d.f. to model the baseline hazard function, at a minimum the model estimates are the same as from a Cox model. A problem could arise if there was insufficient data to model the baseline hazard. For example, if there are too few events for the number of knots used. Royston-Parmar models are more flexible than standard parametric models and the hazard function is explicitly modelled,[34] thereby allowing the model to be used for predicting absolute outcome risk over time in patients of interest. The advantages of Royston-Parmar models will be demonstrated in Chapters 2 and 4.

In survival analysis, proportional hazards modelled (fitted on the hazard scale) are most common. Royston-Parmar models can also be fitted on the proportional-odds scale which is a generalisation of the log-normal model or fitted on the probit scale.[40] However, interpretation of estimates from models fitted using the odds or probit scales are more difficult and will not be focused on in this thesis.

1.4.5 Non-proportional hazards

The models described above all assume that hazards are proportional for all variables in the model. However, the proportional hazards assumption does not always hold. When necessary, time-dependent effects can be fitted for variables that are not proportional. Time-dependent effects are modelled by allowing an interaction between the β-coefficient and time. This can be modelled more complexly by including a restricted cubic splines function as will be explored in Chapter 4. In general, the development of a prognostic model should keep the fitted model as simple as possible, without unnecessary complexity as this complicates interpretation and implementation of the model and often is a consequence of over-fitting.

1.4.6 Example prognostic model developed using a flexible parametric survival model

The recently published Melanoma Severity Index was developed to predict the risk of death over time in patients diagnosed with a single invasive cutaneous melanoma.[41] The data used to develop the prognostic model came from a population-based cancer registry in Queensland, Australia (n=28654). The multivariable prognostic model was fitted using Royston-Parmar flexible parametric modelling on the probit scale and included eight prognostic factors: (1) gender, (2) age at diagnosis, (3) thickness, (4) smooth rank transformed thickness (an additional transformed variable for thickness), (5) body site, (6) ulceration, (7) positive lymph nodes, and (8) metastasis. Interpretation of the beta-coefficients for a probit survival model is more complex and the authors give the following interpretation:

> *A one-unit change in a covariate results in a one-beta change in risk on the probit (inverse normal probability) scale, where beta is the regression coefficient for the variable in question. However, in a more general sense, a positive beta coefficient means that an increase in the covariate raises the predicted probability of death from melanoma. Conversely, a negative beta coefficient means that an increase in the covariate reduces the predicted probability of death*.[41]

As flexible parametric modelling was used, the model predictions are not restricted to certain time points, for example 5 or 10-year survival probabilities. To demonstrate how the Melanoma Severity Index performs, the authors reported predicted probabilities of survival after 10 years for 12 hypothetical melanoma patients.[41]

1.5 Model development considerations

Chapters 2 to 4 focus primarily on the development of prognostic models (including a review of published prediction models in Chapter 3). Although complex statistical methods are being developed, there is little agreement on the right approach for developing reliable prognostic models.[9] There are many statistical considerations when developing a prediction model and it is good practice to write a protocol in advance of the analysis beginning.[42] The aim is to develop a model that discriminates well between individuals that will have the event and those that will not have the event (good discrimination) and also predicts accurately (good calibration between observed and predicted risk).[19] Ideally a model should be developed with clinical input rather than just relying on statistical procedures alone. For example, if a variable is known clinically to be prognostic, it should be included in the multivariable model regardless of statistical significance.[11] However, below are a few statistical considerations. This list is by no means exhaustive and these considerations are relevant whether modelling a binary or time-to-event outcome.

Selection of candidate prognostic factors

Many hypotheses are tested for inclusion/exclusion of each variable when developing a multivariable model and it is not always sensible to test every variable that is recorded in a dataset, especially when the sample size is small. The power to detect effects in a multivariable model is fairly complex to calculate and is based on the number of events rather than number of observations when the outcome is binary. Simulation studies have led to a rule of thumb that there should be a minimum of 10 events per variable (EPV).[43] Variables (also known as candidate prognostic factors or candidate predictors) should be selected based on subject knowledge (systematic review and/or expert judgement), distributions (ideally predictor distributions should be wide), missingness (if a predictor is not recorded in practice, it may not be useful in a model) and similarity to other variables (variables could be clustered or some excluded if highly correlated with others, using data reduction methods).[10,12,19,22] Candidate predictors should not be selected based on association with the outcome and therefore should not be based on univariable analyses which could lead to missing important predictors.[22,44]

Data quality and missing values

Data should be fit for purpose and measurement error should be minimal.[19] If values are missing for predictors, these can be handled in different ways. Complete case analysis (removing observations with any missing values) can be used if the proportion of observations with missing values is small, typically less than 5%.[22] Otherwise multiple imputation can be used to handle missing data; however imputation techniques assume that values are missing at random which may not always be a reasonable assumption.[45,46]

Data handling and modelling continuous predictors

Data handling can include creating new variables from old ones (potentially combining multiple variables into one new variable) or collapsing a categorical variable into fewer categories if required.[19] Continuous variables may not be linear in a multivariable prognostic model, therefore it is important to consider how to model them more appropriately. Creating a categorical variable with two or more categories from a continuous variable is not advised as there is a loss of information and power, nor is using optimal cut-points to categorise a continuous variable advisable.[44,47,48] Instead it is better to consider transformations of the continuous variable if it is not linear on its original scale. The variable may be linear after applying a simple transformation such as the natural logarithm. Alternatively, more complex methods can be used such as multivariable fractional polynomials which use combinations of transformations to achieve a better model fit.[49-51] However, with increased complexity in the functions, there is added difficulty in interpreting predictor effects and increased potential for over-fitting (i.e. developing a model that is over-fitted to the data observed by chance).

Variable selection strategy for inclusion in the multivariable model

There are multiple strategies that can be used to select variables to be included in the multivariable model, however there is little agreement over the 'best' approach as yet.[19,52] Commonly used approaches include fitting the full model (containing all candidate variables) or using automatic variable selection methods such as backward elimination.[19,52] The full model approach means that candidate variables are selected possibly using data reduction techniques and then all candidate predictors are included in the multivariable model. This method has been said to avoid selection bias, over-fitting and results in meaningful confidence intervals.[19,22] Backward elimination is an automatic selection procedure that starts with the full model and sequentially removes variables based on a series of hypothesis tests.[19] Automatic selection procedures are data-driven variable selection techniques that make decisions regarding inclusion or exclusion of variables based on hypothesis tests and a pre-specified significance level for inclusion/exclusion. In backward elimination, variables are removed sequentially if the p-value for a variable (usually using a Wald test) exceeds the specified significance level. Backward elimination is preferred to forward selection which begins with the null model (no variables in it) and sequentially adds 'significant' variables that meet the inclusion p-value.[53]

Other considerations

There are many other topics to consider when developing a multivariable prognostic model. These include deciding if there are any interactions between variables that should be tested or included in the model, trying to avoid over-fitting by having adequate data but also checking and adjusting the final model for this (called shrinkage),[22] checking for possible outliers and deciding if any sensitivity analyses should be planned.

1.6 Validating a prognostic model

Model development involves estimating model parameters to minimise errors in the dataset at hand; therefore parameter estimates and selected predictors are based on the information available within one particular dataset. However, as mentioned in Section 1.5, when tailoring model functions, the selection of predictors, and the specification of continuous predictors, over-fitting may be an issue such that the model does not perform as well when taken outside of the development data.

Therefore, once a prognostic model has been developed, validation of that model is required to quantify how well the model performs both internally (in the same data or population used to develop the model) and externally (in new data from an external but relevant population).[54,55] Statistically, a model should ideally have unbiased predictions and predict accurately across a wide range of different individuals (with different case-mix variations), and explain as much variation as possible. Clinically a model may be useful in two ways: if it reliably can classify patients into prognostic groups (i.e. groups that have different prognoses) or a model that can be used to estimate the prognosis for individual patients.[56] A model may be clinically useful, even when not statistically considered ideal. For example, if predictions in the high risk group are upwardly biased, then it may not be clinically important if the true risk is still high enough to trigger the same clinical action.

The Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) statement gives guidelines for the reporting of studies that develop and/or validate clinical prediction models.[42] Using TRIPOD, studies can be classified into different types depending on whether they perform any validation of the model developed and what type of validation. Type 1 studies have a single dataset available and the authors either develop a model and look at apparent performance (Type 1a), or develop and validate using resampling techniques for internal validation (Type 1b). Type 2 studies split a single dataset for development and validation, either using a random split (Type 2a) or non-random split (Type 2b). Type 3 studies have separate datasets for model development and validation, and Type 4 studies only validate an existing model.[42] The different types of validation and their purpose are now discussed.

1.6.1 Internal validation

The aim of internal validation is to evaluate the performance of the developed model using the same data (or same population) as that used for model development.

Apparent validation

When model performance and model fit are evaluated in exactly the same data used for model development, this is known as apparent validation. Usually this gives an indication of the best possible performance of that model, as over-fitting usually causes performance to deteriorate in other datasets. If apparent validation suggests poor model performance, it is unlikely that the model will perform adequately in other datasets.

Data splitting

When the dataset is initially split for model development and validation (also referred to as training and testing sets), this is known as data splitting. For example, 70% of the data may be used for model development and 30% reserved for validation. If data come from multiple centres, data can be split by centre to achieve the required proportions in each set. However, if the dataset is split randomly, it is likely to only differ due to chance variation and therefore this is not usually considered a good validation technique.[56]

To develop the best possible prediction model, as much data as possible should be used in model development to avoid missing genuine prognostic factors due to low power.[22] If only a proportion of the data is used for model development, parameter estimates may be less stable than if all the available data were used.[12] Therefore performing a single split is not very efficient,[57] nor does it provide an external dataset for validation as the population is the same as for model development. This will not give a true indication of how the model is likely to perform in other populations for which the model may be intended for use. The model performance could also vary depending on the actual split that is made. Steyerberg states that 'split-sample validation is a method that works when we do not need it. It should be replaced in medical research by more efficient internal validation techniques, and by attempts of external validation'.[12]

Cross validation

An improvement upon simple data splitting is to split the data multiple times, each time reserving a different portion of the data for validation and developing the model in the rest. Leave-one-out cross-validation means leaving one patient out of the analysis at a time and predicting the outcome for that individual using the model with *n*-1 patients. This is repeated *n* times and model performance is summarised across the *n* patients when excluded. However, reserving groups of patients for cross-validation, (for example, 10-fold cross-validation would divide the data into 10 parts) is more accurate.[22]

Bootstrapping

Bootstrapping is another technique that is often used in internal validation and does not require excluding any data for validation, therefore using the maximum amount of data for model development. This method validates the modelling process, for example, the variable selection procedure if performed in each bootstrap sample.[58] It also provides an estimate of the expected optimism, which can be used to down-weight the model parameter estimates.

Nonparametric bootstrapping is done by sampling with replacement from the original data to obtain a new sample of the same size as the original data, and this is called a bootstrap sample. The model is then developed afresh in this dataset using the same procedure (e.g. selection process) as in the development of the original model, as far as possible. The apparent performance of the new model is then estimated in the bootstrap sample. Then, its performance is also calculated in the original dataset, and its difference with the apparent performance is estimated ('optimism' estimate). This process is repeated many times (for example, taking 1000 bootstrap samples), and the average optimism estimate obtained. This indicates the potential optimism in the original developed model and thus 'optimism-adjusted' performance statistics can be derived for the original model, simply by subtracting the optimism from the apparent performance.[58,59]

An example of bootstrapping used for internal validation of a prognostic model can be given for PIEPOC (mentioned in Section 1.1), where the authors used bootstrapping with 200 samples to obtain bias-corrected estimates of calibration performance of observed and predicted probabilities.[7]

1.6.2 External validation

A model will nearly always perform better within the data used for development as it is optimised for that data, hence external validation is needed to check performance of the model and assess generalizability or transportability of the model in new but similar patients.[60] External validity is evaluated by assessing model performance in data that is completely external to that in which the model was developed.

Ideally, the model performance will be checked in multiple datasets to get a better idea of how well it performs in multiple different settings or different patient groups (with different case-mix variation). If multiple validation datasets are available, model performance can be evaluated in each of them and summarised using meta-analysis methods.[61,62] This will produce summary estimates of the average performance and the heterogeneity in performance across settings. Ideally, a model will have excellent average performance and no heterogeneity. Heterogeneity would suggest that the model performs better in some settings or populations than in others and may not be suitable for use in all of them. Novel meta-analysis methods are considered in Chapter 5 for pooling performance statistics obtained through 'internal-external cross validation' which is an approach that was proposed for both developing and ‘externally’ validating a model multiple times when several studies are available.

1.6.3 Validation statistics

Validation of a model requires calculating performance statistics. These can include measures of model fit, which test statistical validity (for example R^2 measures the proportion of explained variation), but also examining how well the model discriminates (separates) between individuals who had the event and those that did not (discrimination) and if the model’s predicted (expected) outcome risk agrees with the observed outcome risk on average (calibration).

Overall model performance statistics

Overall performance statistics such as explained variation (R^2 or Nagelkerke’s R^2 for survival data) and the Brier score are measures of overall model fit.[22] The Brier score can be calculated for binary outcomes and is the average squared difference between the observed outcome (0 or 1) and predicted probability (ranges from 0 to 1). The Brier score is not directly applicable to survival data due to censoring, but can be calculated at specific time points.[12] Measures of overall model performance can be broken down into components of calibration and discrimination.[12] In this thesis, calibration and discrimination will be considered separately as these are more clinically meaningful.

Calibration

The calibration of a model is a measure of how well the predicted probabilities from the model agree with the observed outcome. Assessing calibration of the model shows how much the model over or under-predicts absolute outcome risks over time. Recalibration of the model can be considered if the model does not calibrate well in external populations.[63] In particular recalibration may be needed if the baseline risk is higher or lower than in the development data or if the predictor effects differ. Unfortunately a recent systematic review of articles which reported some kind of external validation of a multivariable prediction model, found that 67% of the articles did not report calibration performance of the prediction model.[54] This shows that calibration, an important element of model performance, is often wrongly ignored and Chapters 5 and 6 will look at methods to address this.

The calibration statistics included in this thesis are:

- **Calibration-in-the-large:** For a logistic regression model, this is the difference between the mean number of predicted events (mean($\widehat{Y}$)) and the mean number of observed events (mean(Y)).[12,64] Mean($\widehat{Y}$) is calculated by averaging the predicted probabilities of the event (using the prognostic model) and the mean(Y) is calculated by averaging the binary event indicator for whether an event was observed or not (0 or 1). This can also be estimated by fitting a logistic model for the probability of the outcome (p) with the linear predictor (LP) as a covariate (offset term),

$$\text{logit}(p_i) = \alpha+\beta(LP_i) \qquad \textbf{(1.17)}$$

where the estimate of α given β=1 is the estimate of calibration-in-the-large.[12] Calibration-in-the-large should be close to zero for a well calibrated model.

Calibration-in-the-large can also be calculated for a survival model by evaluating the difference between mean observed and expected probabilities of events at specified time points.

- **Expected/observed ratio:** The ratio of expected and observed number of events should be close to one if the model calibrates well in the validation dataset. This is easily calculated for a logistic model using the expected and observed number of events or can be calculated as the ratio of expected and observed probabilities (of the event) at specified time points for a survival model.

- **Calibration slope:** Relates to the slope of the calibration plot where patients are often categorised into risk groups, for example using deciles of predicted probability of having the event. The predicted probability for each risk group is plotted (on the x-axis) against the observed outcome proportion in that risk group (on the y-axis). If a line is fitted through the points on the graph, the slope of the line is the calibration slope. A model that calibrates well in the validation dataset would result in a calibration slope=1. A calibration slope<1 indicates that a model over-predicts, and conversely a calibration slope>1 indicates that a model under-predicts.

  Rather than categorising patients into risk groups, a calibration model can be fitted in the validation dataset. Using a logistic model for example, the calibration model is given by (1.17) where p_i is the probability of the binary outcome Y_i and LP_i is the linear predictor from the developed model, then $\widehat{\beta}$ is the estimated calibration slope.[65,66]

Discrimination

Discrimination is a measure of how well a model can differentiate between individuals who have the event of interest and those that do not. A model is more likely to discriminate well if the range of predicted probabilities is wide. The C-statistic and the D-statistic are the two discrimination statistics of particular focus in this thesis:

- **C-statistic:** The probability of concordance between predicted and observed outcomes is calculated by looking at all possible pairs in which one individual had the event and the other did not have the event. The C-statistic is calculated as the proportion of pairs in which the individual that had the event had a higher predicted probability than the individual that did not have the event.[22] This is equivalent to the area under the ROC curve for a logistic model.[10] For a survival model, Harrell's C-statistic is the proportion of pairs in which the individual with the higher predicted survival probability survived longer than the other individual.[10] Pairs, in which both individuals are censored, both have the event at the same survival time or where one individual is censored at an earlier time than the other individual's survival time, cannot be ordered and therefore cannot be included in the calculation. The closer the C-statistic is to the value one, the better the model discriminates between individuals of high and low probability of the event. A value of 0.5 suggests that the model does no better than chance alone. Other C-statistics have been proposed for survival data,[67,68] but Harrell's C-statistic will be used in this thesis.

The following figure is taken from Collins et al.[54] and shows the C-statistic in internal and external validation across a range of reviews, highlighting again that discrimination in external validation data is generally worse than in the development data, due to over-fitting and thus over-optimism in the developed data.

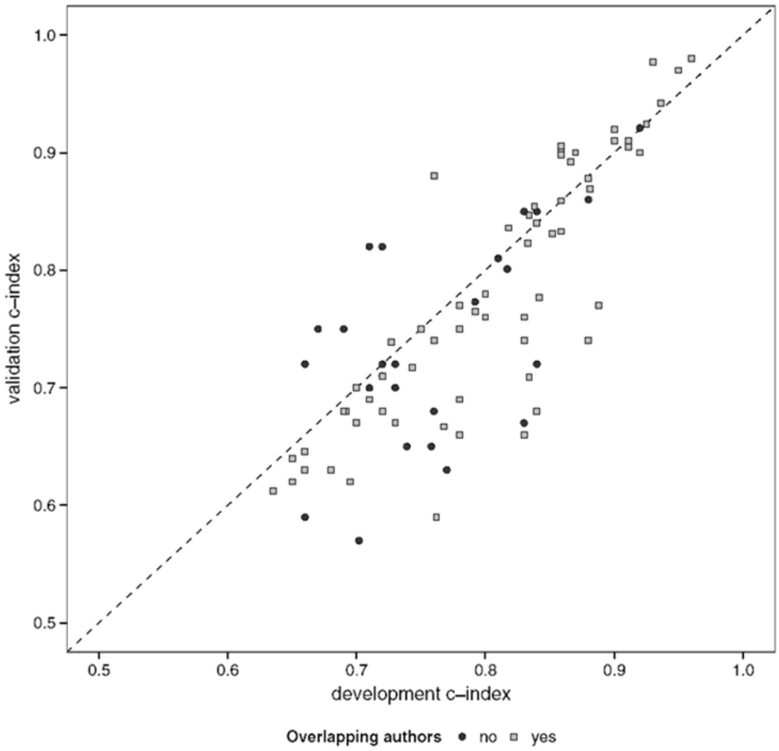


Figure 1.8: Paired C-statistics from development and external validation, also showing if authors overlap or not (originally published by Collins et al.[54]). Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0).

- **D-statistic:** Proposed by Royston and Sauerbrei, the D-statistic is used for survival models as a measure of separation. The D-statistic can be defined as

$$D = \kappa\sigma^*$$

where σ^* is an estimate of the standard deviation (σ) of the 'true' prognostic index ($\boldsymbol{x\beta}$) which is assumed to be Normal(μ, σ^2).[69] The σ^* is estimated by fitting a Cox model to the standard normal order statistics for $\boldsymbol{x_i\widehat{\beta}}$ of individuals in the study. This is then multiplied by the scaling factor $\kappa = \sqrt{8/\pi}$ to give the D-statistic. The κ-scaling is used so that the D-statistic is an estimate of the log hazard ratio for two prognostic groups, if the data were split into two equal groups using the median prognostic index.[70] So values of the D-statistic further from zero indicate better separation.

Reclassification statistics

Other metrics that could be calculated include reclassification statistics such as Net Reclassification Improvement (NRI) which is used to compare two nested models, for example, models with and without a new biomarker, to evaluate the improvement due to the additional marker.[71] For a binary outcome, reclassification tables can be constructed based on cut-points for probability of the event occurring and tabulating the number of patients that fall into each category based on each model. The NRI summarises the movement (reclassification) of patients either upwards (low to high risk) or downwards (high to low risk) in patients that had the event and patients that did not have the event. Therefore, the NRI is estimated as

$$\widehat{NRI} = \left(\hat{p}_{\text{up, events}} - \hat{p}_{\text{down, events}}\right) - \left(\hat{p}_{\text{up, nonevents}} - \hat{p}_{\text{down, nonevents}}\right)$$

where $\hat{p}_{\text{up, events}}$ is a probability calculated as the proportion of patients that had the event that moved up a category (comparing the original model with the new model) out of all patients that had the event. The other probabilities can be calculated in a similar way, for patients reclassified downwards and/or patients that did not have the event.[71] Further developments later meant that the NRI could also be calculated avoiding the use of cut-points, called the continuous NRI (NRI>0) and calculated for survival models.[72,73] However, more recent publications have shown that there are issues with reporting of reclassification in the literature,[74] and also with the methodology.[75-78] In particular, the NRI is affected by miscalibration of the model and can also result in optimistic p-values for biomarkers that do not have predictive value (shown through simulation).[77]

1.7 Presentation of prognostic models for clinical decision making

Prognostic models can be presented in different ways for use in clinical practice; the assistive approach provides predicted probabilities of the outcome of interest and allows the clinician to use their own judgement in how they manage the patient, while a directive approach recommends different management strategies for patients in different risk categories (such models are also called clinical prediction rules or clinical decision rules).[21] Most prediction models are presented as assistive tools, making the assumption that having accurate predictions helps clinical decisions, which may not always be true.[79] Therefore, following development and validation of a prognostic model, impact studies should be conducted to assess the influence of the model on clinical decisions and patient outcomes.[3]

Models can be presented in different formats depending on how the model is intended for use. The model formula for predicting outcome probability from a logistic model or survival model (using baseline survival probability at a given time point) could be published in the format given in equation (1.3) or (1.9). Alternatively, the model could be simplified to a scoring system, in which predictor coefficients are scaled and rounded to produce simple point scores which can be added up and used to look up the corresponding predicted probability of outcome. A graphical presentation for assigning risk scores could be used instead, such as a nomogram, for which risk scores are assigned using bars for each variable and summed up to give the overall risk score and corresponding probability of the outcome (example in Figure 1.9).[80]

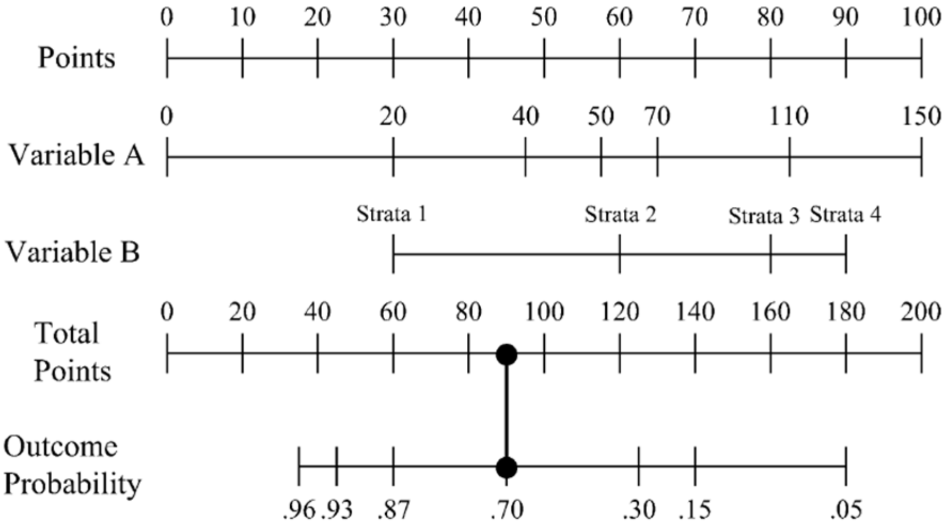


Figure 1.9: Example of a nomogram (Reprinted from the American Journal of Obstetrics and Gynecology, 194, Grobman and Stamilio,[80] Methods of clinical prediction, 888-894, Copyright 2006, with permission from Elsevier.

Presentation formats will be summarised for articles included in the literature review (Chapter 3), however, this thesis focuses on development and validation of the statistical model rather than presentation of the final model and how the model can be used to aid clinical decision making.

1.8 Importance of improving methodology in prognosis research

Many other areas of clinical research are well established with good methodology and reporting standards leading to advances in clinical practice. For example, there is a considerable amount of methodological literature about how to appropriately design, analyse and report a clinical trial to test new treatments, to ultimately identify more effective treatments for patients with particular conditions and lead to a change in practice. Despite a lot of activity in prognosis research, the improvement in the understanding of prognosis for many diseases or conditions is very slow and only a small proportion of the research impacts on clinical practice.[1,81] Methodological improvements are needed, as many of the problems arise from poor use of statistical methods and reporting. For example, publication bias of prognostic factor studies, inappropriate dichotomising of continuous prognostic factors is endemic, and therefore replication of initial evidence can be rare.[47,82]

The application and development of more sophisticated statistical methods therefore warrants urgent attention in prognosis research. For example, a systematic review of prognostic models in cancer found that almost all studies of time-to-event data used the Cox proportional hazards model.[83] The Cox model has potential limitations in prognostic modelling due to the baseline hazard not being explicitly modelled. Flexible parametric models such as Royston-Parmar models should be considered as an alternative, and they have potential advantages as predictions can be made *over time*.[34] It is therefore important to evaluate the use of Royston-Parmar models and their advantages in prognostic modelling, and this will be a major focus in this thesis.

There are also many prognostic models being developed and published, but very few of them are being externally validated. Several systematic reviews of prognostic or prediction models have looked at reporting of model validation. Mallett et al. published a review on the reporting of performance of prognostic models in cancer and found that only 34% of articles included some form of validation in the original article and only 21% of models were externally validated (either in original article or subsequent publication).[84] Another systematic review by Bouwmeester et al. stated that `the majority of model development studies reported predictive performance in the development data only' and 'only a very few model development studies reported an external validation of the model in the same paper'.[85] Collins et al. published a systematic review on the methodological conduct and reporting of external validation of multivariable prediction models and found that in general, reporting of external validation was poor and that calibration was often not reported. They concluded:

> *It may therefore not be surprising that an overwhelming majority of developed prediction models are not used in practice, when there is a dearth of well-conducted and clearly reported (external validation) studies describing their performance on independent participant data.*[54]

A suggestion made by PROGRESS is that there should be a shift of focus from developing new models to validating and updating existing models.[3] Research should be building upon what is already known and new biomarkers or genetic markers should be evaluated by considering if they improve prediction after accounting for known prognostic factors. It is also important to validate a model in multiple settings, and PROGRESS state:

> *The collation and synthesis of individual participant data from multiple studies offers a natural opportunity to increase sample size. Models can then be developed using data from a subset of studies and assessed on data from the remaining studies. Variation in model accuracy across studies and its causes can be explored.*[3]

Reporting of any external validation tends to be lacking and poor quality, therefore validating in multiple settings is even less common. Methods have been published for simultaneously developing and externally validating a model multiple times when individual participant data are available from multiple studies (internal-external cross-validation).[69] Few studies have considered how to combine performance statistics across multiple studies or clusters.[61,62,69] Meta-analysis methods for combining performance statistics across multiple studies will therefore be developed and evaluated in Chapters 5 and 6 of this thesis.

1.9 Aims and overview of thesis

The overarching aim of the thesis is to apply, develop and evaluate novel statistical methods for prognosis research, with particular focus on prognostic models. In particular, this thesis aims to:

- Apply the flexible parametric survival model of Royston and Parmar to address clinically relevant questions in hip replacement, pancreatic cancer and breast cancer research.
- Review current practice for modelling the baseline hazard in prognostic model research, and illustrate the benefits of the Royston-Parmar approach over the standard Cox model.
- Illustrate methodological issues for prognostic model development using randomised clinical trial data with multiple treatment groups, with recommendations.
- Propose and evaluate a novel multivariate meta-analysis approach for validating the performance and implementation of prognostic models across multiple settings.

The thesis has seven chapters. Chapters 2 to 4 focus primarily on model development issues and comparisons between the flexible parametric approach and the Cox model in novel settings. Therefore these chapters only consider time-to-event data and survival analysis. Chapters 5 and 6 focus on model validation in multiple settings. In these chapters logistic models are also considered. Several articles have been published on prognosis research since starting this PhD in 2011. For example, the PROGRESS Group published a framework and highlighted areas that require methodological research[1-4] To reflect these advances and keep up with the field, the original thesis plan was adapted to include validation as this was one of the areas identified as requiring methodological research. An outline of the chapters is given below.

Chapter 2 explores the use of flexible parametric survival models for the purpose of outcome risk prediction and comparison across relevant patient groups. The models are developed in registry data collected from osteoarthritis patients who had hip replacement surgery. This is a large dataset which is ideal for exploring the use flexible parametric methods for individual prediction over time. Here comparisons are made between different types of hip replacement. The aim of this chapter is to highlight the advantages of using flexible parametric models for absolute risk prediction and comparison of mortality rates, compared to the standard Cox model.

Chapter 3 reviews how recent prediction model articles developed and possibly validated a clinical prediction model using time-to-event data. The main aim is to establish if/how researchers are modelling the baseline hazard and if/how absolute risk predictions are being made for patients from the developed model, to ascertain areas for improvement.

Chapter 4 examines methodological issues when developing a new prognostic model using randomised trial data. Data from two clinical trials of patients with advanced stage pancreatic cancer are used, but challenges are present due to different treatment groups, missing data, and non-proportional hazards for treatment leading to a time-dependent effect being required. The model is internally validated but no external data were available at the time for external validation of the model.

Chapter 5 develops a novel multivariate random-effects meta-analysis method for validating a prediction model when multiple studies are available. The approach summarises the joint discrimination and calibration performance of a model, whilst accounting for their correlation. It produces a summary of calibration and discrimination performance, and quantifies the amount of heterogeneity in model performance across studies. The results of the meta-analysis are used to predict how well the model would be expected to perform in a new but similar study that was not included in the meta-analysis. It is also shown to help identify the best implementation strategy, in particular regarding recalibration of the model intercept (baseline hazard).

Chapter 6 uses simulation to evaluate the random-effects meta-analysis approach proposed in Chapter 5; in particular, to investigate whether the assumption that performance statistics across studies come from a normal distribution is plausible. A range of performance statistics are evaluated under a range of different scenarios, including differing levels of heterogeneity for either the intercept or predictor effects.

Chapter 7 includes discussion of the overall findings from this thesis and makes recommendations for further work.

CHAPTER 2: HIP REPLACEMENT SURGERY IN OSTEOARTHRITIS PATIENTS

2.1 Introduction

This chapter aims to demonstrate the use and advantages of flexible parametric survival methods in a real clinical dataset. Royston-Parmar models are fitted to compare mortality and revision rates over time between osteoarthritis patients receiving cemented and uncemented procedures in hip replacement surgery. This chapter shows how the shape of the baseline hazard function can be explored and demonstrates how it can be modelled (on the log-cumulative hazard scale) using restricted cubic splines which were introduced in Chapter 1. Clinical conclusions and recommendations for hip replacement are made based on the results identified. Further, the advantages of using flexible parametric models are shown in terms of predicting adjusted survival probabilities for groups of patients and also for individuals based on their own predictor values. The clinical findings of this chapter were published in the BMJ.[86]

2.2 Background to hip replacement procedures

A hip replacement may be necessary when the hip joint is damaged, causing pain and difficulty in daily activities such as walking. Surgery is considered when the problem cannot be treated non-surgically to relieve pain, such as by taking painkillers, the use of steroid injections or other pain relieving creams or gels.[87] Common reasons for damage to the hip joint include osteoarthritis, rheumatoid arthritis and hip fractures.[88] The investigations in this chapter look at primary total hip replacement (THR) in patients with osteoarthritis which is the most common type of arthritis.

THR has become a very common procedure with almost 69000 primary THRs recorded by the National Joint Registry in 2010.[89] A primary THR is the initial surgery in which the hip is first replaced, as opposed to a revision which refers to later surgery on the replaced hip. The surgical procedure involves removing the femoral head and any damaged cartilage from the acetabulum (hip socket). A metal stem is inserted into the hollow centre of the femur with a ball attached to the upper part of the stem. A prosthetic cup is then fixed into the acetabulum.[90]

As well as the different designs and materials used for prostheses, procedures can be categorised by the method for fixing the prostheses in place and as such, procedures can be classified as either cemented, uncemented or hybrid. An alternative to THR is hip resurfacing which preserves more of the patients' bone. A short description on each of these procedures is now given below.

2.2.1 Cemented procedures

In cemented THRs, the prostheses are fixed in place using bone cement to fill the space between the stem and the bone. The cementing technique has moved away from finger-packing high viscosity cement into the bone, towards using a cement gun with low viscosity cement to force the cement into the space between the stem and the bone. This change in method of cement fixation has resulted in a lower implant failure rate as there is less loosening of the implant over time.[91,92] An example of a stem used in a cemented THR is shown in Figure 2.1. A cemented THR is often considered more suitable for older patients and those in poor health as the recovery time is shorter than other procedures and allows the patient to bear weight on their hip fairly soon after surgery.[93]

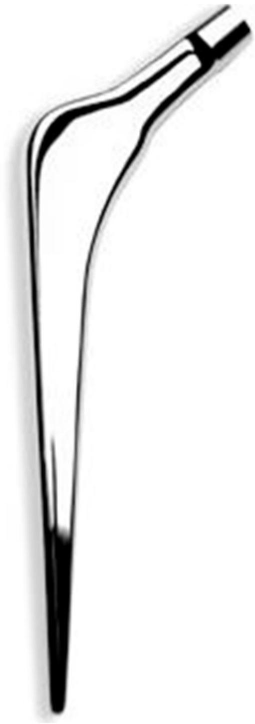


Figure 2.1: A cemented stem (image courtesy of The McMinn Centre and Smith & Nephew).

2.2.2 Uncemented procedures

The prostheses used in an uncemented THR have a textured surface or coating (seen in Figure 2.2). The prostheses are fitted very close to the surface of the bone and the porous surface encourages bone growth to fix the bone to the prosthesis. For the bone to bind to the prosthesis, the prosthesis must be fitted very tightly to the bone by being no more than 1-2mm apart.[93] The recovery time for this procedure is longer than for a cemented procedure and the patient is unable to bear weight on the hip as the bone needs time to bind to the prostheses. However, after recovery, uncemented THRs have been shown to be successful in younger and more active patients as there is less loosening over time than with cement,[94] and thus time to revision is thought to be prolonged.

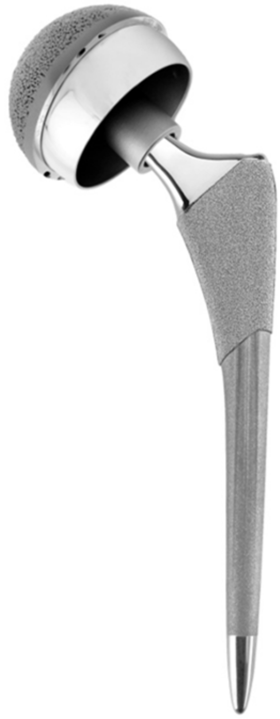


Figure 2.2: An uncemented stem and cup (image courtesy of The McMinn Centre and Smith & Nephew).

2.2.3 Hybrid procedures

THRs are classified as a hybrid hip replacement if an uncemented acetabular cup and a cemented stem are used. A reverse-hybrid hip replacement is where a cemented acetabular cup and an uncemented stem are used.

2.2.4 Birmingham Hip Resurfacing

Hip resurfacing is a conservative procedure in which very little bone is removed and capped with a large diameter metal bearing (Figure 2.3). Birmingham hip resurfacing (BHR) first became available in 1997.[95] It uses a metal head and a metal acetabular cup, and this is referred to as a 'metal-on-metal' joint. BHRs are not suitable for all patients and studies have raised some concerns, for example by showing an increased failure rate in women.[96] However, the BHR has been shown to be successful in younger (aged under 55) and more active males.[97]

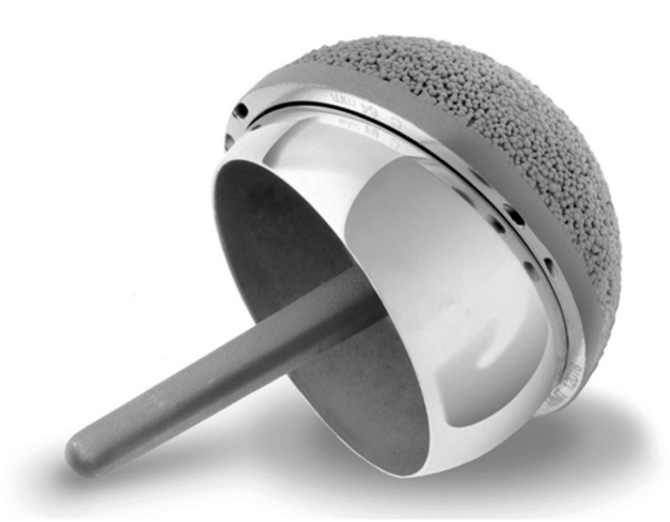


Figure 2.3: A Birmingham Hip Resurfacing implant (image courtesy of The McMinn Centre and Smith & Nephew).

2.3 Data

A database of 335841 primary THRs or hip resurfacings performed on osteoarthritis patients, recorded between 2003 and 2011 was obtained from the National Joint Registry (NJR) by Smith and Nephew in August 2011. This dataset included cemented, uncemented and hybrid THRs, as well as BHRs in men only. The dataset was formally obtained by Smith and Nephew to be used by their clinical teams to investigate survivorship. The dataset was passed to Mr Derek McMinn (pioneer of the BHR and an unpaid consultant for Smith and Nephew and part of their clinical team) for such research purposes. Subsequently, Mr McMinn asked Prof. Richard Riley to provide professional statistical analysis of the data. Kym Snell then wrote the analysis plan, and performed all the analyses under the supervision of Prof. Riley. The variables recorded in the dataset, along with a brief description, are displayed in Table 2.1. American Society of Anesthesiologists (ASA) grade definitions are reported as presented in the NJR 8th Annual Report 2011.[89]

Table 2.1: Variables in hip replacement dataset with brief description.

Variable	Description
Index Number	Patient identifier
Age	Age of patient at primary surgery
Gender	Male or female
ASA Grade	Pre-operative physical status classification system P1: Fit and healthy P2: Mild disease not incapacitating P3: Incapacitating systemic disease P4: Life threatening disease P5: Expected to die within 24 hours with or without an operation
Side	Left or right side of hip
Primary operation date	Date of primary hip replacement
Procedure type	Cemented, uncemented, hybrid or BHR
Surgery approach	Approach to hip used by surgeon
Complexity	Complex or non-complex procedure
Endpoint type	Unrevised, revised or death
Revision	Binary indicator for revision as endpoint
Death	Binary indicator for death as endpoint
Time to event (years)	Time from primary surgery to either revision, death or end of follow-up

Patients were followed up from the date of primary surgery until either death (from any cause) or first revision. Thus, for each patient, his or her time to death, time to revision or time until censored (whichever came first) was recorded in years. The aim was to investigate survivorship (of hip or patient) up to their first revision, and so for any patient that received a revision, no mortality information was utilised after the revision surgery. Patients were censored due to loss to follow-up or if they had not experienced an event (revision or death) before the end of the study.

2.4 Objectives

2.4.1 Clinical objectives

The pre-specified primary clinical objective was to assess if there was any difference in the mortality and revision rates over time between osteoarthritis patients receiving cemented and uncemented THRs for a first hip replacement.

Secondary objectives included:

- Comparing mortality and revision rates for cemented and uncemented THRs in specific subgroups of patients such as different ASA grades.
- Comparing mortality and revision rates for cemented THRs, uncemented THRs and BHRs in men under 55 years of age (i.e. the subset of patients BHRs are aimed at).

2.4.2 Statistical objectives

Alongside the clinical objectives, for this thesis there is also a statistical objective, which is to implement flexible parametric modelling techniques and demonstrate the advantages of modelling the baseline hazard function by fitting Royston-Parmar models rather than Cox regression models. In particular, by modelling the baseline hazard function, it will be shown how it is possible to obtain absolute survival estimates in addition to relative risk estimates such as hazard ratios, normally reported after fitting a Cox regression model. Comparisons of absolute survival probabilities will also be made at specific time points, and population-averaged survival curves will be predicted and plotted for particular covariates, such as procedure type, whilst adjusting for confounding factors included in the model. These advantages of modelling the baseline hazard function will be highlighted throughout the results, shown in Section 2.6 and discussed further at the end of the chapter.

2.5 Methods

The pre-specified statistical analysis plan is now summarised.

2.5.1 Data cleaning, inclusion and exclusion criteria

Any patients missing data for age, gender, procedure type or ASA grade were excluded from all analyses. This resulted in 11 patients being removed for missing age or gender. This represented a very small proportion of the total number of patients, and so more sophisticated methods for handling missing data (such as multiple imputation) were not considered necessary. Surgery approach was missing for 5.5% of patients and therefore this variable was excluded from the analyses. Later sensitivity analyses including this variable suggested that it did not impact upon the investigations of the survival difference between procedure types (Appendix A1). In the dataset, duplicate entries existed for 911 patients that underwent hip replacement surgery on both hip sides on the same date. Any such patient that had a different procedure type on each side was excluded from all analyses (24 patients); those patients receiving the same procedure type on both sides remained in all analyses (887 patients), but their duplicate entry was removed and a variable named ‘both sides’ was created to identify them from patients having a single hip replacement. Patients that received a ‘hybrid’ hip replacement were excluded from all analyses as this labelling makes no distinction between hybrid and reverse hybrid, and such patients were not relevant to the clinical objectives. This resulted in a further 51530 patients being excluded and left a total of 283365 patients for use in our analyses (including BHR men).

2.5.2 Summary of data

Baseline characteristics of patients and follow-up information (e.g. number censored and number of deaths) were summarised for the dataset as a whole as well as by procedure type. Mean, standard deviation (SD), median, interquartile range (IQR) and the minimum and maximum values were reported for continuous variables, and total numbers and proportions were reported for categorical variables. Patient follow-up was summarised by the median duration (reverse Kaplan-Meier method)[98] and total follow-up in person-years for each procedure type and overall.

2.5.3 Analysis of primary outcomes

Survival analysis methods were used to investigate the two primary outcomes, time to death and time to revision, and the association with each of five variables available in the dataset: procedure type, gender, age at primary surgery, ASA grade before the operation, complexity and if both sides were replaced.

Initially, Kaplan-Meier plots were produced for each variable to show unadjusted differences in the probability of survival over time. To quantify these associations and investigate the primary outcomes listed above, Royston-Parmar survival models were then fitted to obtain both unadjusted and adjusted results for each variable.[40] Royston-Parmar models were introduced in Chapter 1. The hazard ratios estimated using this approach were practically identical to those obtained through fitting a Cox regression model.

The multivariable (adjusted) models were fitted using a backward elimination procedure which forced procedure type to remain in the model and retained any other variables that remained statistically significant (as defined by $p<0.1$). Age was included as a continuous variable and a linear association with outcome was assumed. The linearity assumption was checked by plotting martingale residuals from the multivariable model against age and using a smoother.[99] The smoothed line did not deviate from the line $y=0$, supporting the linear assumption for age.

Using the estimates from the multivariable model, adjusted survival curves were predicted and plotted by calculating the population-averaged survival curve for each procedure type. This was done by predicting the survival curve for each patient in the dataset, using their own predictor values but assuming they received a particular procedure (e.g. cemented). The individual survival curves were then averaged to give the population-averaged survival curve for that procedure type.[34] This was repeated for each procedure type, so that the entire dataset was used for each survival curve. This also allowed the difference in the mean absolute survival probabilities between procedure types to be calculated at different time points, after adjusting for the other variables in the model.

For analyses relating to mortality as the outcome, any patient that had a revision was censored on the date of revision and for analyses of revision, patients that died before having a revision were censored on the date of death. It is not possible to know that censoring due to revision is not informative for the outcome of death; therefore the results from the mortality analyses relate only to patients that had not had a revision prior to time *t*. For example, interpreting a hazard ratio at time 2 years relates to the mortality rate of patients that had not had a revision before 2 years.

2.5.4 Assessing the proportional hazards assumption

In survival analysis, hazard ratios are assumed to remain constant over time. If the proportional hazards assumption is violated, this implies that the hazard ratio changes over time. Non-proportional hazards can be incorporated into the model by including time-dependent effects. One of the many advantages of Royston-Parmar models is the ease in which time-dependent effects can be incorporated into the model, fitted using restricted cubic splines.[33] `Log-log’ plots were used to check the proportional hazards assumption for each variable. Non-parallel lines suggest that the proportional hazards assumption does not hold,[100] and time-dependent effects should be considered. Continuous variables such as age were categorised into quartiles to produce the ‘log-log’ plot.

2.5.5 Number of knots for the baseline hazard function

The baseline hazard function is modelled in Royston-Parmar models using restricted cubic splines with a specified number of knots. A suitable number of knots for the data was decided by plotting the baseline hazard function for the null model using different numbers of knots and using this to inform the decision. The AIC and BIC were also considered to aid the choice of knots (defined in Chapter 1, equations (1.15) and (1.16)). A model with a smaller AIC or BIC is preferred.

2.5.6 Analysis of secondary outcomes

A secondary analysis aimed to check if any differences observed between procedure types remained consistent across the different ASA grades. Therefore, the multivariable model described in Section 2.5.3 (but excluding ASA grade) was re-fitted for patients in each of the ASA grades separately, and the hazard ratio for cemented versus uncemented procedures was estimated for each ASA grade.

Data on BHR procedures were only available for males. Thus, another secondary analysis was conducted to compare BHR to cemented and uncemented THRs in males under 55 years of age as this is the subset of patients that BHR is often aimed at.[97] A similar modelling strategy was followed to that described in Section 2.5.3, although no adjustment was required for gender. Adjusted survival curves were estimated and plotted for the three procedure types.

2.6 Results

2.6.1 Summary of data for cemented and uncemented THRs

In the dataset, a total of 154996 patients received cemented THRs and 120017 patients received uncemented THRs between April 2003 and July 2011. The cemented group were followed up for a median length of 3.55 years (range: 0.001 to 9.65 years) and the uncemented group for a median length of 2.45 years (range: 0.001 to 8.58 years).

Table 2.2 shows that baseline characteristics were not balanced between the cemented and uncemented groups. The mean age in the cemented group was almost 7 years older than the uncemented group. In the uncemented group, there were more males (42.1% compared to 34.5% in the cemented group) and a higher proportion of patients with ASA grade 1 (20.2% compared to 14.4% in the cemented group). There were also more patients categorised as complex procedures in the cemented group (8.5%) than in the uncemented group (0.7%). Approach appeared relatively balanced between cemented and uncemented groups, although the lateral approach was used more frequently in the cemented group (21.1% compared to 12.3% in the cemented group) and the posterior approach which was the most common approach in both groups, was used more in the uncemented group than the cemented group (52.3% compared to 41.9% in the cemented group).

During follow-up, in the cemented group 11745 (7.6%) patients died and 1589 (1.0%) had a revision, whilst in the uncemented group 3728 (3.1%) patients died and 1917 (1.6%) had a revision.

Table 2.2: Summary of baseline characteristics, outcome and follow-up by procedure type.

		Cemented (n=154996)	Uncemented (n=120017)	Overall (n=275013)
Baseline characteristics				
Age, years	Mean (SD)	73.20 (8.69)	66.69 (10.09)	70.36 (9.87)
	Median	73.79	67.00	71.15
	IQR	67.96 – 79.24	60.46 – 73.67	64.16 – 77.31
	Range	15.93 – 103.41	15.40 – 106.15	15.40 – 106.15
Gender, n (%)	Male	53409 (34.46)	50529 (42.10)	103938 (37.79)
	Female	101587 (65.54)	69488 (57.90)	171075 (62.21)
ASA grade, n (%)	1	22336 (14.41)	24276 (20.23)	46612 (16.95)
	2	107395 (69.29)	82104 (68.41)	189499 (68.91)
	3	24369 (15.72)	13151 (10.95)	37520 (13.64)
	4	852 (0.55)	456 (0.38)	1308 (0.48)
	5	44 (0.03)	30 (0.02)	74 (0.03)
Surgery				
Approach, n (%)	Anterior	485 (0.34)	429 (0.37)	914 (0.35)
	Antero-lateral	12223 (8.56)	7161 (6.12)	19384 (7.46)
	Hardinge	34366 (24.06)	28834 (24.66)	63200 (24.33)
	Lateral (inc. harding)	30137 (21.10)	14321 (12.25)	44458 (17.11)
	Posterior	59855 (41.90)	61086 (52.25)	120941 (46.56)
	Trochanteric osteotomy	1164 (0.81)	96 (0.08)	1260 (0.49)
	Other	4620 (3.23)	4989 (4.27)	9609 (3.70)
	Missing	12146 (7.84)	3101 (2.58)	15247 (5.54)
Complexity, n (%)	Non-complex	141825 (91.50)	119172 (99.30)	260997 (94.90)
	Complex	13171 (8.50)	845 (0.70)	14016 (5.10)
Both sides, n (%)	No	154798 (99.87)	119497 (99.57)	274295 (99.74)
	Yes	198 (0.13)	520 (0.43)	718 (0.26)
Follow-up				
Endpoint, n (%)	Death	11745 (7.58)	3728 (3.11)	15473 (5.63)
	Revision	1589 (1.03)	1917 (1.60)	3506 (1.27)
	Unrevised	141662 (91.40)	114372 (95.30)	256034 (93.10)
Length of follow-up, person-years	Total	535035	323477	858512

2.6.2 Proportional hazards assumption

By visual inspection of ‘log-log’ plots, the proportional hazards assumption was assessed for all of the variables considered in modelling both outcomes. Figure 2.4 shows approximately parallel curves for procedure type with mortality as the outcome. Parallel curves suggest that there is no problem with the proportional hazards assumption, and this was the case for the majority of plots. An exception was gender for the outcome of revision, Figure 2.5 shows a section where both curves are overlaid. However, there is little spacing between the curves in this figure and only a small amount of crossover, so this was not considered of great concern. For these reasons, all the Royston-Parmar models were fitted assuming proportional hazards. The complete set of log-log plots for all variables can be found in Appendix A1.

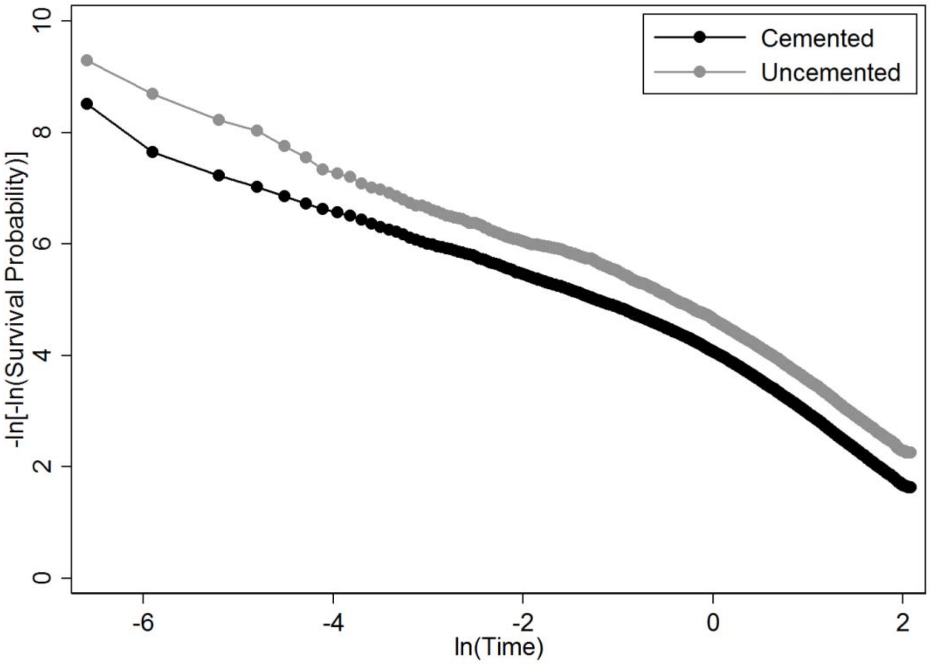


Figure 2.4: ‘Log-log’ plot for procedure type and the outcome of mortality.

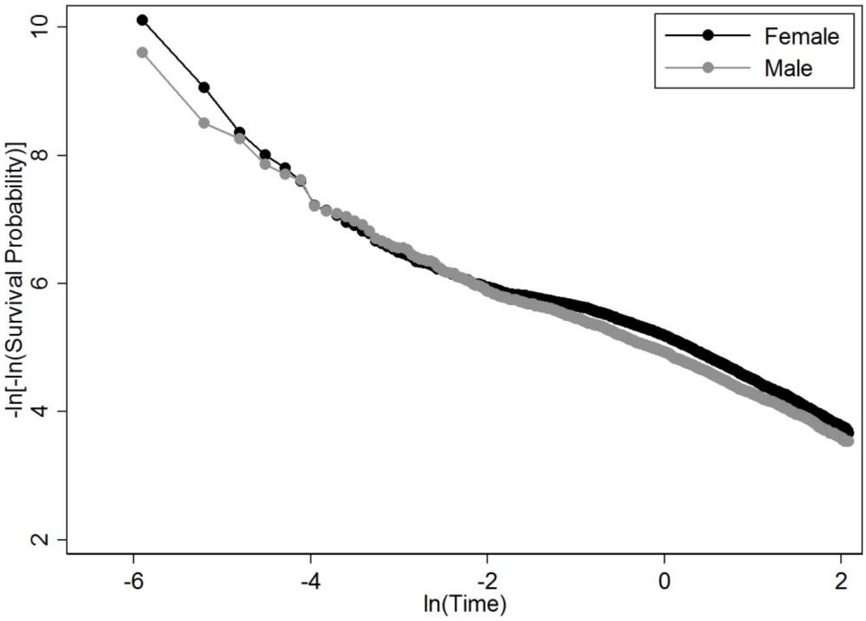


Figure 2.5: 'Log-log' plot for gender and the outcome of revision.

2.6.3 Number of knots for the baseline hazard function

Royston-Parmar models with between 2 and 10 d.f. for the baseline hazard function were fitted for each of the outcomes and plotted against time. Figure 2.6 shows little difference in the shape of the estimated baseline hazard function using 1 knot (2 d.f.) compared to as many as 9 knots (10 d.f.) for the outcomes of mortality and revision.

The AIC and BIC can aid the choice of d.f. to use in modelling the baseline hazard. In this case, the sample size was large and so increasing the d.f. resulted in a large reduction in AIC and BIC, even though the function did not appear to change much visually. Based on AIC alone, 10 d.f. was best for both outcomes. Based on BIC alone, 10 d.f. was best for mortality and 6 d.f. for revision. However, in cases of large datasets, Royston and Lambert suggest basing the decision on 'feel' rather than formal statistics.[34] Therefore it was decided that 5 d.f. would be adequate and is recommended as the default number for large datasets. Figure 2.6 suggests that visually any more knots would result in what appears to be over-fitting.

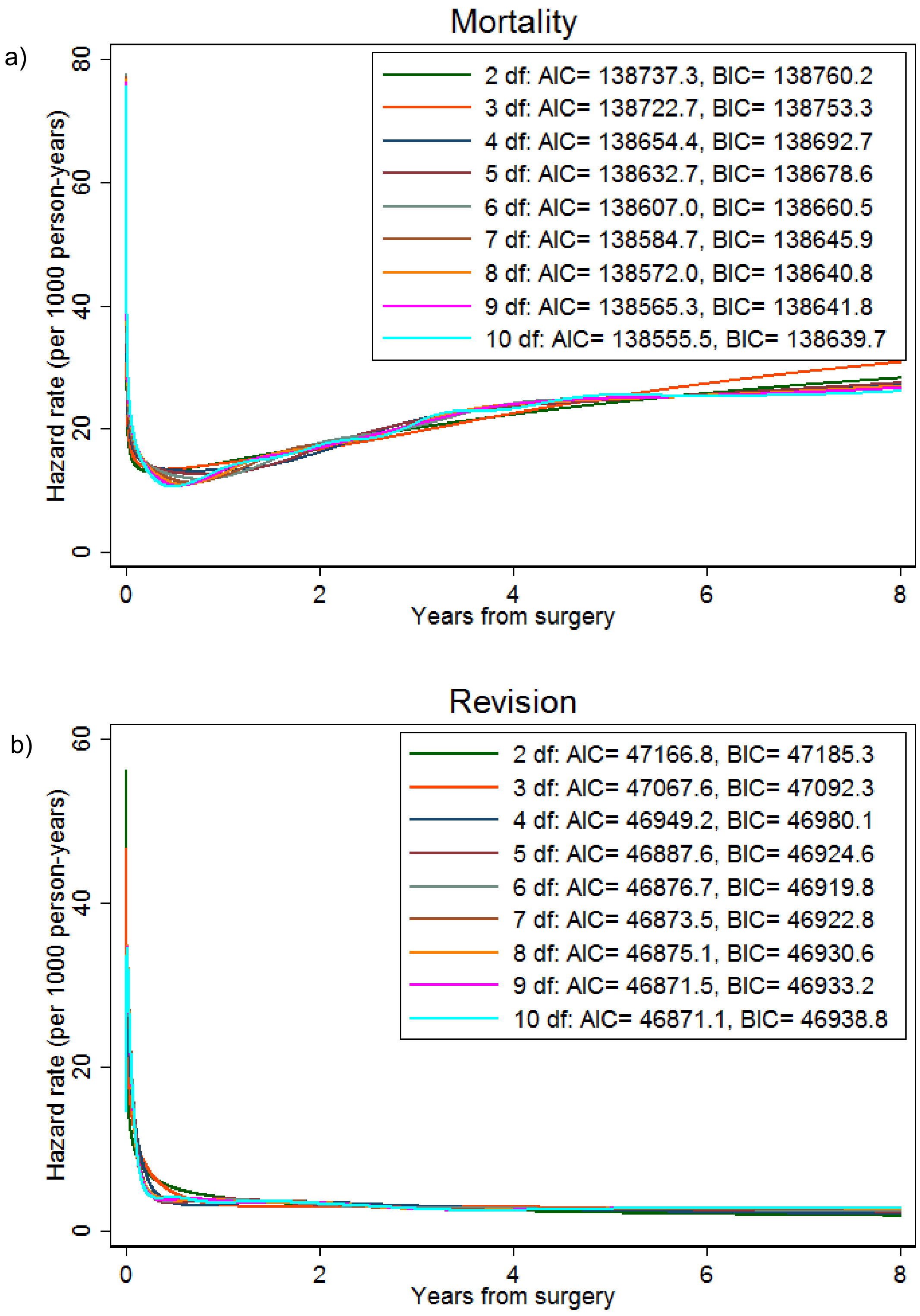


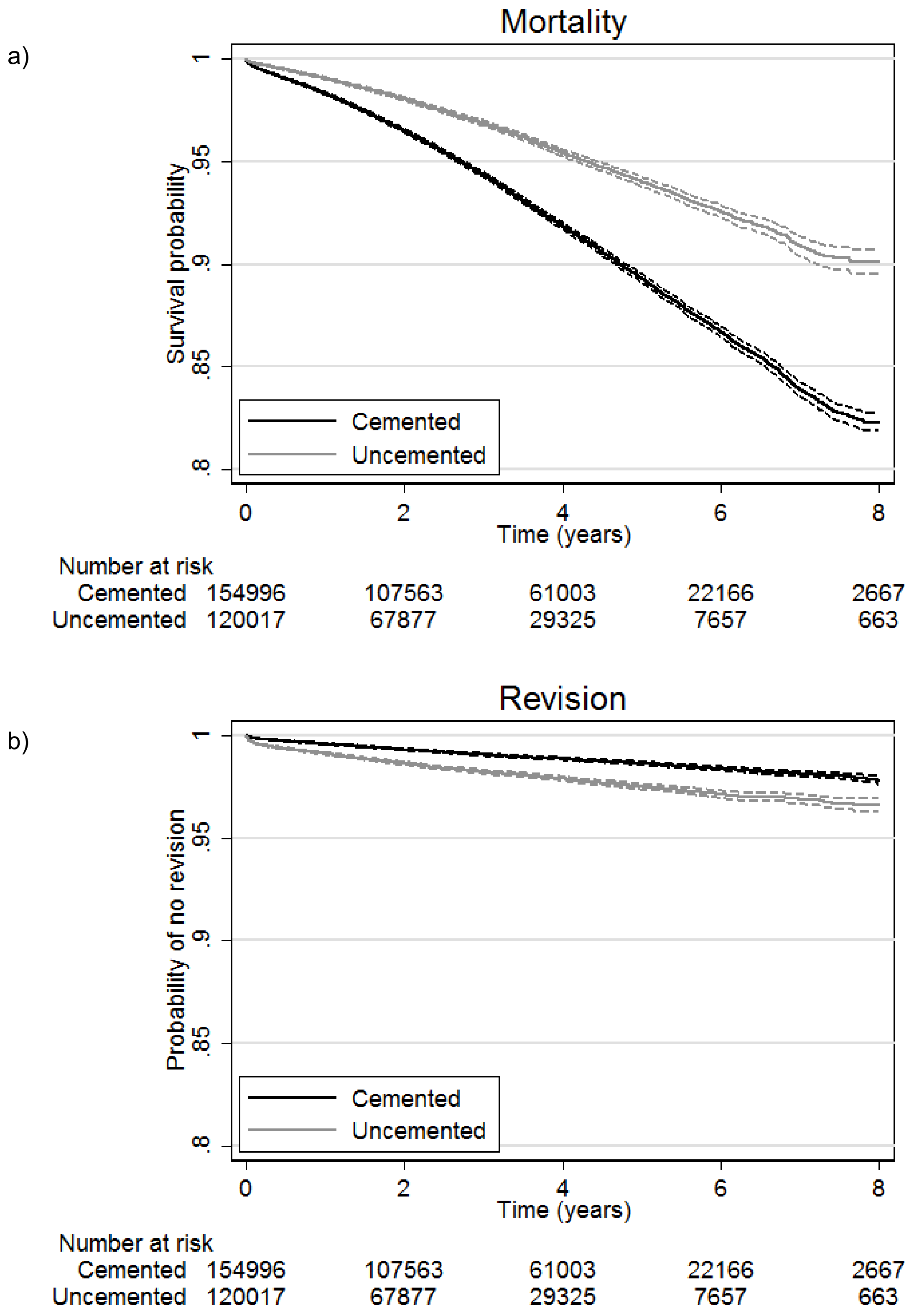
Figure 2.6: The baseline hazard function estimated using different degrees of freedom for the outcomes of (a) mortality and (b) revision. AIC=Akaike information criterion, BIC=Bayesian information criterion.

The baseline hazard function relates to the hazard function when all the covariates in the model are equal to zero.[34] Assuming a model with all the variables included, the baseline hazard would relate to a female that had a non-complex, uncemented procedure on one side only, was aged 0 years and categorised as ASA grade 1. This does not relate to any patient in the dataset due to age 0, but alternatively age could be mean-centred to make the baseline more meaningful and it would then relate to someone of average age. The shape of the baseline hazard function (Figure 2.6) for both mortality and revision reflect the high initial hazard immediately following surgery, after which the hazard rate rapidly declines and remains low for revision but increases slowly over time for mortality. Cox proportional hazards models do not explicitly model the baseline hazard function, and are thus inferior to the Royston-Parmar approach in this regard **(STATISTICAL ADVANTAGE OF ROYSTON-PARMAR MODELS 1: Estimating and displaying the baseline hazard function)**.

2.6.4 Primary outcome analyses

Unadjusted survival analysis

The Kaplan-Meier unadjusted survival curves for procedure type are shown in Figure 2.7. This shows that patients receiving a cemented THR had a lower probability of survival and a higher probability of no revision over time when compared to patients receiving an uncemented THR. The unadjusted probability of patient survival at 8 years is 90.9% (95% CI: 89.5% to 90.7%) in the uncemented group and 82.3% (95% CI: 81.9% to 82.7%) in the cemented group. The difference between cemented and uncemented procedures in terms of revision is much smaller than for survival as the outcome, with an unadjusted probability of no revision at 8 years of 97.8% (95% CI: 97.6% to 98.0%) and 96.6% (95% CI: 96.3% to 96.9%) in cemented and uncemented procedures respectively.



Number at risk					
Cemented	154996	107563	61003	22166	2667
Uncemented	120017	67877	29325	7657	663

Number at risk					
Cemented	154996	107563	61003	22166	2667
Uncemented	120017	67877	29325	7657	663

Figure 2.7: Kaplan-Meier unadjusted survival curves for procedure type with (a) mortality and (b) revision as the outcomes (95% CIs given by dashed lines).

Kaplan-Meier plots for the other variables were also produced (see Appendix A2). The Kaplan-Meier plots suggested that women had a better outcome than men in terms of both mortality and revision. The probability of survival was highest for the youngest quartile of patients and lowest for the oldest quartile of patients as would be expected. Conversely, the youngest quartile of patients had the lowest probability of no revision over time. Patients categorised as complex procedures had a lower probability of survival over time compared to patients that had non-complex procedures.

Fitting unadjusted Royston-Parmar models for mortality as the outcome resulted in an unadjusted hazard ratio of 1.83 (95% CI: 1.76 to 1.90), suggesting a higher mortality rate in the cemented group compared to the uncemented group. A hazard ratio of 0.53 (95% CI: 0.50 to 0.57) was estimated for revision as the outcome, indicating a hazard of revision 47% lower in cemented versus uncemented procedures. The univariable hazard ratio estimates obtained by fitting Royston-Parmar models (Table 2.3) are practically identical to those from Cox regression for most variables **(STATISTICAL ADVANTAGE OF ROYSTON PARMAR MODELS 2: For proportional hazards, the hazard ratio estimates are practically identical to Cox regression but one additionally obtains the baseline hazard)**. However, as there are large imbalances in baseline characteristics between the two procedure groups (seen in Table 2.2), these univariable hazard ratio estimates are likely to be confounded and therefore adjusted analyses are more appropriate.

Table 2.3: Univariable (unadjusted) estimates from Royston-Parmar and Cox models for mortality and revision.

Variable	Mortality				Revision			
	Royston-Parmar		Cox		Royston-Parmar		Cox	
	HR (95% CI)	P-value	HR (95% CI)	P-value	HR (95% CI)	P-value	HR (95% CI)	P-value
Uncemented procedure	1	-	1	-	1	-	1	-
Cemented procedure	1.826 (1.760 to 1.895)	<0.001	1.826 (1.760 to 1.895)	<0.001	0.534 (0.500 to 0.571)	<0.001	0.534 (0.499 to 0.571)	<0.001
Age (years)	1.095 (1.093 to 1.098)	<0.001	1.095 (1.093 to 1.098)	<0.001	0.980 (0.976 to 0.983)	<0.001	0.980 (0.977 to 0.983)	<0.001
Female	1	-	1	-	1	-	1	-
Male	1.302 (1.261 to 1.344)	<0.001	1.302 (1.261 to 1.344)	<0.001	1.227 (1.147 to 1.312)	<0.001	1.226 (1.147 to 1.312)	<0.001
ASA grade 1	1	-	1	-	1	-	1	-
ASA grade 2	1.577 (1.499 to 1.659)	<0.001	1.577 (1.499 to 1.659)	<0.001	0.965 (0.885 to 1.052)	0.422	0.966 (0.886 to 1.053)	0.433
ASA grade 3	3.654 (3.456 to 3.864)	<0.001	3.653 (3.455 to 3.863)	<0.001	1.129 (1.006 to 1.266)	0.039	1.130 (1.007 to 1.267)	0.037
ASA grade 4	6.702 (5.894 to 7.620)	<0.001	6.700 (5.893 to 7.618)	<0.001	0.492 (0.245 to 0.988)	0.046	0.492 (0.492 to 0.175)	0.046
ASA grade 5*	4.245 (2.460 to 2.878)	<0.001	4.244 (2.460 to 7.324)	<0.001	-	-	-	-
Non-complex	1	-	1	-	1	-	1	-
Complex	1.492 (1.426 to 1.561)	<0.001	1.505 (1.438 to 1.574)	<0.001	0.805 (0.707 to 0.916)	0.001	0.791 (0.693 to 0.903)	0.001
Single side	1	-	1	-	1	-	1	-
Both sides	0.533 (0.347 to 0.817)	0.004	0.533 (0.347 to 0.817)	0.004	1.672 (1.007 to 2.777)	0.047	1.670 (1.006 to 2.773)	0.047

*Not estimable for revision as the outcome

Adjusted results

Mortality

To adjust for baseline confounding in the mortality comparisons, procedure type was forced into the Royston-Parmar model and the variables age, gender, ASA grade and complexity were identified as statistically significant and thus also included. The model estimates (Table 2.4) are again practically identical to those from Cox regression. The adjusted hazard ratio for cemented compared to uncemented procedures is 1.11 (95% CI: 1.07 to 1.16) which remained statistically significant (p<0.001), but was substantially smaller than the unadjusted estimate (HR=1.83) due to the adjustment for confounding. The hazard of mortality is estimated to be 11% higher in patients that received a cemented procedure compared to patients that received an uncemented procedure, assuming that they had not had a revision before this time. Other estimates from this multivariable model suggest a 1-year increase in age is associated with an increase in the hazard of death of 9% (95% CI: 6.9% to 15.5%) and the hazard for ASA grade relative to ASA grade 1 increases as grade increases. The hazard of death is also 53.7% (95% CI: 48.9% to 58.7%) higher for males compared to females and 39.5% (95% CI: 33.2% to 46.2%) higher for complex procedures compared to non-complex procedures.

A sensitivity analysis was performed by refitting the model for mortality but including the variable ‘approach’ which was missing in over 15000 patients (see Appendix A3), hazard ratios were similar for the variables included in Table 2.4 and there was still a highly significant difference between cemented and uncemented procedures.

Table 2.4: Multivariable Royston-Parmar model estimates for mortality as the outcome.

Variable	Hazard ratio	95% Confidence interval	P-value
Uncemented procedure	1	-	-
Cemented procedure	1.111	1.069 to 1.155	<0.001
Age (years)	1.090	1.088 to 1.092	<0.001
Female	1	-	-
Male	1.537	1.488 to 1.587	<0.001
ASA grade 1	1	-	-
ASA grade 2	1.192	1.133 to 1.255	<0.001
ASA grade 3	2.152	2.033 to 2.278	<0.001
ASA grade 4	3.517	3.091 to 4.002	<0.001
ASA grade 5	2.937	1.702 to 5.070	<0.001
Non-complex	1	-	-
Complex	1.397	1.333 to 1.464	<0.001

Unlike Cox regression, Royston-Parmar models also estimate the baseline hazard. This allows the population-averaged survival curves for the cemented and uncemented procedure types to be estimated. These can be thought of as ‘adjusted’ for the included variables of age, gender, ASA grade and complexity **(STATISTICAL ADVANTAGE OF ROYSTON PARMAR MODELS 3: Population-averaged `adjusted’ survival curves)**. The mean probability of survival at 8 years in patients receiving either a cemented or uncemented THR is quite high, 0.850 (95% CI: 0.846 to 0.854) for cemented and 0.863 (95% CI: 0.858 to 0.868) for uncemented patients (Figure 2.8). There was a significant absolute difference in the mean probability of survival between the cemented and uncemented groups over time, although the difference in survival curves is much smaller than that seen in the unadjusted Kaplan-Meier plot (Figure 2.7a).

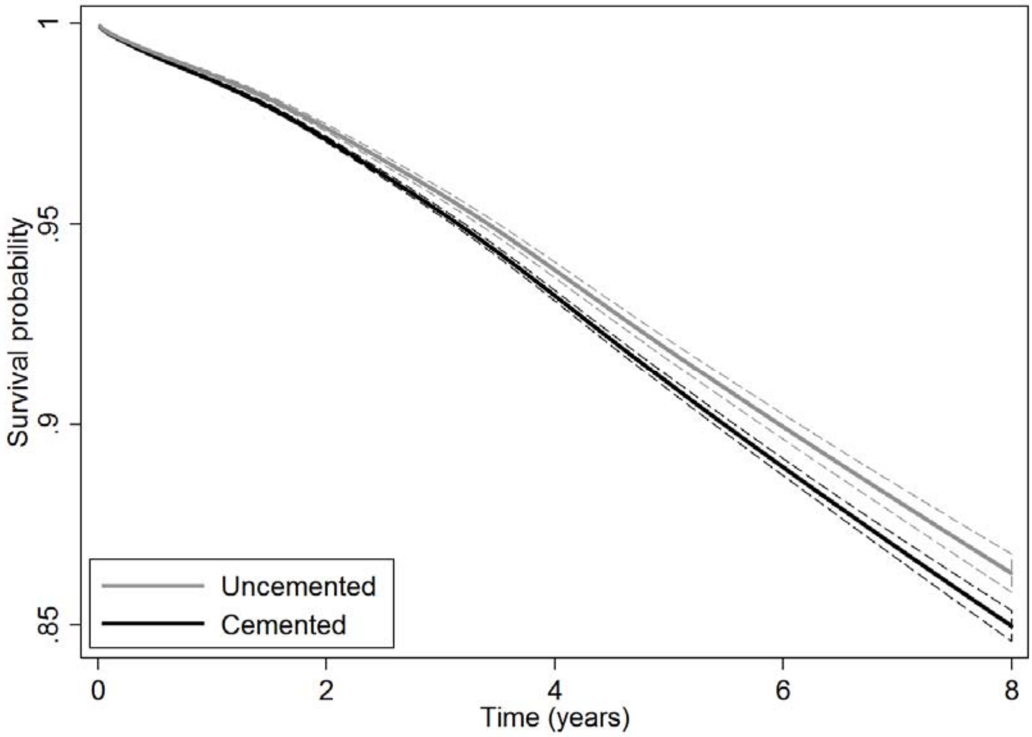


Figure 2.8: Population-averaged (adjusted) survival curves for procedure type and mortality as the outcome (95% CIs given by dashed lines).

Importantly however, although the absolute difference between cemented and uncemented procedures is statistically significant, the difference in predicted mean survival probabilities at any given time is very small **(STATISTICAL ADVANTAGE OF ROYSTON PARMAR MODELS 4: Estimating differences in absolute *S*(*t*) over time, after adjustment for covariates in the model)**. Table 2.5 shows the absolute mean survival probabilities for the cemented and uncemented groups and the difference in mean absolute survival probabilities at different times following surgery. The difference increases over time and the largest difference is seen at 8 years with a difference in mean survival probabilities of 0.013 (95% CI: 0.007 to 0.019).

Table 2.5: Mean survival probabilities and difference in mean survival probabilities between procedure types at selected time points.

Time point	Cemented *S*(*t*) (95% CI)	Uncemented *S*(*t*) (95% CI)	Difference (95% CI)
1 year	0.986 (0.985 to 0.986)	0.987 (0.987 to 0.988)	0.001 (0.001 to 0.002)
2 year	0.971 (0.970 to 0.972)	0.974 (0.973 to 0.975)	0.003 (0.002 to 0.004)
3 year	0.953 (0.952 to 0.954)	0.958 (0.956 to 0.959)	0.005 (0.003 to 0.006)
4 year	0.932 (0.931 to 0.933)	0.939 (0.937 to 0.941)	0.006 (0.004 to 0.009)
5 year	0.910 (0.908 to 0.912)	0.919 (0.916 to 0.921)	0.008 (0.005 to 0.011)
8 year	0.850 (0.846 to 0.854)	0.863 (0.858 to 0.868)	0.013 (0.007 to 0.019)

Revision

The multivariable model for revision included procedure type along with adjustment for age, gender and ASA grade. Complexity and both sides were not significant in the stepwise selection process when modelling revision and therefore not included in the final model. The adjusted hazard ratio for cemented procedures compared to uncemented procedures was 0.57 (95% CI: 0.53 to 0.62), shown in Table 2.6. This suggests that the hazard of revision was 43% lower at any time point in patients who received a cemented procedure compared to an uncemented procedure. As mortality is a competing risk to revision, this result relates only to patients that had not died before this time. The hazard of revision was also higher for males compared to females but decreased with age and with increasing ASA grade up to grade 3 (a lower hazard ratio was observed for grade 4 compared to grade 1 although not significant and a hazard ratio was not estimable for grade 5 due to small numbers in this category).

Table 2.6: Multivariable Royston-Parmar model estimates for revision as the outcome.

Variable	Hazard ratio	95% Confidence interval	P-value
Uncemented procedure	1	-	-
Cemented procedure	0.578	0.539 to 0.621	<0.001
Age (years)	0.987	0.984 to 0.991	<0.001
Female	1	-	-
Male	1.151	1.076 to 1.232	<0.001
ASA grade 1	1	-	-
ASA grade 2	1.087	0.995 to 1.187	0.065
ASA grade 3	1.377	1.223 to 1.550	<0.001
ASA grade 4	0.599	0.298 to 1.204	0.150
ASA grade 5*	-	-	-

*Not estimable

The population-averaged (adjusted) survival curves in Figure 2.9 show a greater probability of no revision for the cemented group compared to the uncemented group, with similar absolute probabilities to those seen in the unadjusted Kaplan-Meier plot (Figure 2.7b). However, as for mortality at 8 years of follow-up, the mean predicted probability of no revision was 0.979 (95% CI: 0.978 to 0.981) in the cemented group and 0.964 (95% CI: 0.962 to 0.967) in the uncemented group which are both higher than for mortality. There was a difference in mean predicted revision probabilities of 0.015 at 8 years (95% CI: 0.012 to 0.017) assuming patients were still alive at this time.

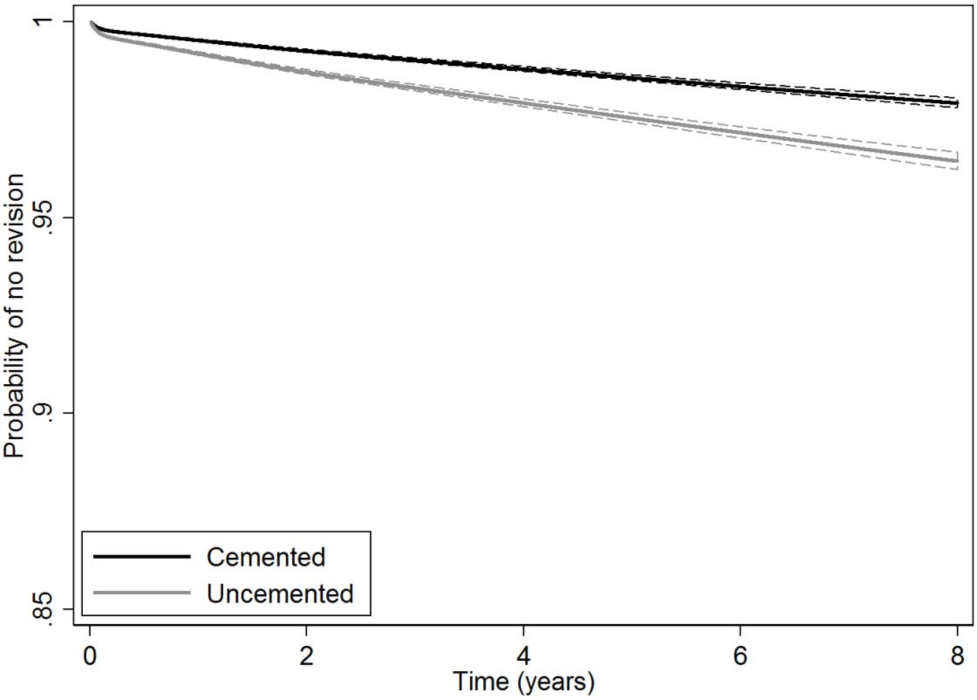


Figure 2.9: Population-averaged (adjusted) survival curves for procedure type and revision as the outcome (95% CIs given by dashed lines).

Predictions for individuals

One of the advantages of modelling the baseline hazard function using Royston-Parmar models is the ability to make predictions, not only for patient groups but for individual patients, based on the specific values that they take for covariates in the model **(STATISTICAL ADVANTAGE OF ROYSTON PARMAR MODELS 5: Making risk predictions for groups of patients and individuals)**. The predicted difference in mortality rates and survival probabilities between cemented and uncemented procedures at 1, 5 and 8 years following the THR are given in Table 2.7. The differences were predicted for individuals of different ages, gender, ASA grades and complexity. Table 2.7 shows that for a 50 year old male, categorised as ASA grade 2 that had a non-complex procedure, the predicted difference in mortality rate between cemented and uncemented procedures is 0.36 (95% CI: 0.22 to 0.50) per 1000 person-years at 8 years. Whereas for someone of the same gender, ASA grade and complexity but of age 80, the predicted difference in mortality rate is 4.79 per

1000 person-years (95% CI: 3.02 to 6.56). The largest predicted difference in 8 year survival between procedure types (cemented – uncemented) is 0.0377 (95% CI: 0.0240 to 0.0515), observed for an 80 year old male, categorised as ASA grade 3 with a complex primary THR.

Table 2.7: Predicted difference between procedure types in mortality rate (cemented – uncemented) and survival probability (uncemented – cemented) at 1, 5 and 8 years.

Age	Gender	Grade	Complex	Difference in mortality rate (per 1000 person-years)			Difference in survival probability		
				1 year	5 year	8 year	1 year	5 year	8 year
50	M	1	No	0.150	0.299	0.302	0.0002	0.0012	0.0020
50	M	1	Yes	0.209	0.417	0.421	0.0002	0.0016	0.0028
50	M	2	No	0.178	0.356	0.360	0.0002	0.0014	0.0024
50	M	2	Yes	0.249	0.497	0.502	0.0003	0.0019	0.0033
50	M	3	No	0.322	0.643	0.650	0.0004	0.0025	0.0043
50	M	3	Yes	0.449	0.896	0.906	0.0005	0.0034	0.0059
50	F	1	No	0.096	0.203	0.210	0.0001	0.0008	0.0013
50	F	1	Yes	0.134	0.283	0.293	0.0002	0.0011	0.0019
50	F	2	No	0.114	0.242	0.250	0.0001	0.0009	0.0016
50	F	2	Yes	0.160	0.337	0.349	0.0002	0.0012	0.0022
50	F	3	No	0.207	0.436	0.452	0.0002	0.0016	0.0028
50	F	3	Yes	0.288	0.609	0.631	0.0003	0.0022	0.0039
80	M	1	No	1.992	3.977	4.021	0.0022	0.0134	0.0213
80	M	1	Yes	2.778	5.547	5.607	0.0031	0.0176	0.0268
80	M	2	No	2.374	4.740	4.792	0.0026	0.0155	0.0241
80	M	2	Yes	3.311	6.611	6.683	0.0036	0.0202	0.0297
80	M	3	No	4.286	8.558	8.651	0.0047	0.0243	0.0338
80	M	3	Yes	5.977	11.935	12.066	0.0064	0.0299	0.0377
80	F	1	No	1.278	2.700	2.798	0.0015	0.0092	0.0152
80	F	1	Yes	1.782	3.766	3.902	0.0020	0.0123	0.0198
80	F	2	No	1.523	3.218	3.334	0.0017	0.0107	0.0175
80	F	2	Yes	2.124	4.488	4.650	0.0024	0.0143	0.0225
80	F	3	No	2.750	5.810	6.019	0.0031	0.0177	0.0268
80	F	3	Yes	3.835	8.103	8.395	0.0043	0.0227	0.0324

The hazard function is a rate that can change over time. Thus, although the hazard ratio is assumed to be constant over time in the Royston-Parmar models fitted, the difference in absolute survival probabilities or mortality rates between groups can vary as the baseline hazard rate changes over time. The absolute difference in mortality rates will be larger at time points when the baseline hazard rate is higher. Using a 60 year old male, categorised as ASA grade 2 having a non-complex primary THR as an example, the predicted difference in mortality rate between cemented and uncemented procedures is greatest immediately following surgery, after which it reduces to 0.42 per 1000 person-years (95% CI: 0.27 to 0.58) at 1 year after surgery and then remains fairly constant around 0.85 from 4 to 8 years following surgery (shown in Figure 2.10).

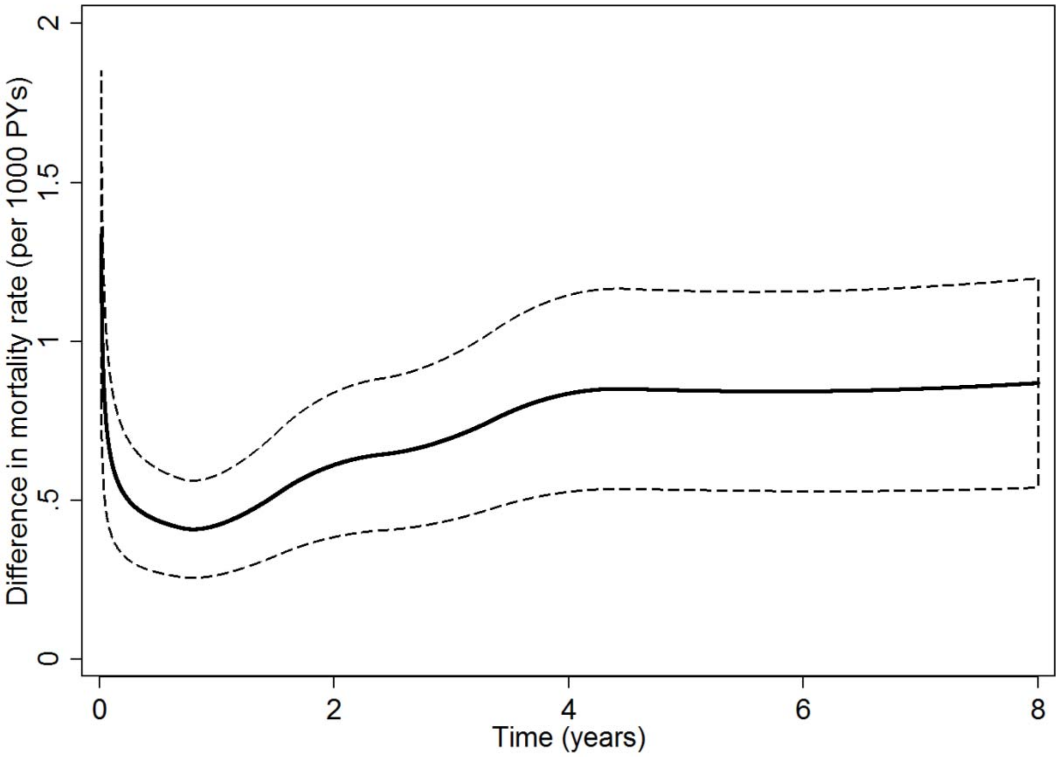


Figure 2.10: Difference in mortality rate (per 1000 person-years) between procedure types (cemented – uncemented) for a 60 year old male, ASA grade 2 having a non-complex primary THR (95% CI given by dashed lines).

2.6.5 Secondary analyses

Results by ASA grade

Mortality

Each ASA grade was examined separately, with adjustment made for age, gender and complexity. There were too few patients in ASA grade 5 for the model to be fitted in that subgroup of patients. A significant difference in mortality between cemented and uncemented procedures was seen for ASA grade 2 patients with a hazard ratio of 1.15 (95% CI: 1.10 to 1.21) and ASA grade 4 patients with a hazard ratio of 1.37 (95% CI: 1.03 to 1.83), yet there was no significant difference observed for ASA grade 1 or grade 3 patients with hazard ratios close to one (Table 2.8).

Table 2.8: Adjusted procedure type hazard ratios (cemented versus uncemented) for mortality, fitted for each ASA grade separately.

ASA Grade	N	Procedure type hazard ratio	95% Confidence interval	P-value
1	46612	0.992	0.884 to 1.112	0.885
2	189499	1.153	1.096 to 1.213	<0.001
3	37520	1.038	0.963 to 1.118	0.328
4	1308	1.371	1.027 to 1.832	0.032

Revision

Adjusting for age and gender, there was a significant difference in the hazard of revision between cemented and uncemented procedures in ASA grades 1, 2 and 3 separately (all p<0.001), hazard ratios were unestimable for ASA grade 4 and 5 subgroups. The hazard of revision was 39% lower in ASA grades 1 and 2, and 54% lower in ASA grade 3 patients receiving a cemented procedure compared to an uncemented procedure (Table 2.9).

Table 2.9: Adjusted procedure type hazard ratios (cemented versus uncemented) for revision, fitted for each ASA grade separately.

ASA Grade	N	Procedure type hazard ratio	95% Confidence interval	P-value
1	46612	0.605	0.512 to 0.715	<0.001
2	189499	0.604	0.554 to 0.659	<0.001
3	37520	0.460	0.384 to 0.550	<0.001

Inclusion of Birmingham Hip Resurfacing – under 55 male only analyses

BHRs are primarily aimed at younger patients (<55 years of age).[97] Therefore secondary analyses were conducted on a subset of patients which match the age range in which BHRs are intended for use. Although the mean age of patients was more similar in the three procedure groups (ranging from 49.16 in BHR to 51.01 in cemented) after removal of patients older than 55 years, baseline characteristics were still not balanced for ASA grade or complexity (Table 2.10). The number of patients included in analyses were also reduced (N=11483) with fewer events (deaths=98, revisions=195).

Table 2.10: Summary of baseline characteristics, outcome and follow-up by procedure type in males under 55 years of age.

		BHR (n=3560)	Uncemented (n=6193)	Cemented (n=1730)
Baseline characteristics				
Age, years	Mean (SD)	47.92 (5.50)	48.49 (6.09)	49.13 (5.87)
	Median	49.16	50.33	51.01
	IQR	44.66 – 52.28	45.76 – 53.02	46.71 – 53.37
	Range	19.03 – 54.99	17.13 – 54.99	18.10 – 54.99
ASA Grade, n (%)	1	2077 (58.34)	2595 (41.90)	692 (40.00)
	2	1421 (39.92)	3225 (52.07)	925 (53.47)
	3	59 (1.66)	360 (5.81)	108 (6.24)
	4	2 (0.06)	12 (0.19)	4 (0.23)
	5	1 (0.03)	1 (0.02)	1 (0.06)
Complexity	Non-complex	3511 (98.62)	6119 (98.81)	1525 (88.15)
	Complex	49 (1.38)	74 (1.19)	205 (11.85)
Both sides, n (%)	No	3531 (99.19)	6143 (99.19)	1716 (99.19)
	Yes	29 (0.81)	50 (0.81)	14 (0.18)
Follow-up				
Endpoint, n (%)	Death	10 (0.28)	56 (0.90)	32 (1.85)
	Revision	60 (1.69)	107 (1.73)	28 (1.62)
	Unrevised	3490 (98.03)	6030 (97.37)	1670 (96.53)
Length of follow-up, person-years	Total	11637	15886	5991

Mortality

Table 2.11 shows the multivariable model estimates for mortality in males under 55 years of age. This is used to demonstrate a situation where hazard ratios alone can be very misleading. Here, the hazard ratio for cemented compared to BHR was 3.86 (95% CI: 1.82 to 8.18) and was highly significant ($p<0.001$). Thus, the hazard of mortality was estimated to be almost four times higher in males under 55 years of age that received the cemented procedure relative to BHR. This sounds dramatic, but knowledge of the baseline hazard is required to make this result more clinically meaningful.

Table 2.11: Multivariable Royston-Parmar model estimates for mortality as the outcome in males under 55 years of age.

Variable	Hazard ratio	95% Confidence interval	P-value
BHR procedure	1	-	-
Cemented procedure	3.863	1.824 to 8.182	<0.001
Uncemented procedure	3.073	1.552 to 6.082	0.001
Age (years)	1.045	1.002 to 1.089	0.039
ASA grade 1	1	-	-
ASA grade 2	1.756	1.056 to 2.920	0.030
ASA grade 3	9.340	5.255 to 16.601	<0.001
ASA grade 4	11.527	1.550 to 85.70	0.017
ASA grade 5*	-	-	-
Non-complex	1	-	-
Complex	1.840	0.901 to 3.758	0.094

*Not estimable

To illustrate that the baseline hazard is necessary to interpret hazard ratios correctly, Table 2.12 shows the relevant results from the analysis of all males as well as the analysis of males under 55 years of age (the full results of the all male analysis can be found in Appendix A4). The hazard ratio for cemented versus BHR was much larger in the under 55 male analysis (HR=3.86) than in the analysis of males of all ages, where the hazard ratio for cemented compared to BHR was 1.67. However, even though the hazard ratio is larger, the absolute difference in $S(t)$ is actually smaller in the under 55 group, due to the lower baseline hazard rate.

This is more easily understood using a simple example; say there is a constant hazard rate for a group A and group B of 0.2 and 0.8 respectively; this would give a hazard ratio of 4 for group B relative to group A. However, if the hazard rates were 0.6 and 1.2 for groups A and B respectively, the hazard ratio would be 2 yet the absolute difference in hazard rates would be 0.6 in both cases. Thus the larger hazard ratio observed for the under 55 male only analysis is relative to the lower baseline hazard in this subgroup of patients than that seen in males of all ages, and therefore the hazard ratio has to be larger to show even a small absolute difference. At 6 years, the baseline hazard for all males is estimated to be 1.848 per 1000 person-years and only 0.458 per 1000 person-years for males under 55 years.

In absolute terms, the difference in survival probabilities between cemented and BHR procedures was only 0.018 (95% CI: 0.008 to 0.028) at 6 years (Table 2.12). This difference is smaller than the difference between the two procedures in males of all ages which was 0.044 (95% CI: 0.029 to 0.060) at 6 years, even though the hazard ratio was much larger. So, larger hazard ratios do not necessarily imply larger absolute differences as they depend on the baseline hazard to which the comparison is being made **(STATISTICAL ADVANTAGE OF ROYSTON PARMAR MODELS 6: Identify significant hazard ratios that are clinically important in relation to baseline risk)**.

Table 2.12: Comparison of hazard ratios and absolute survival probabilities in mortality analyses of males of all ages and males under 55 years of age.

Time point	Analysis	HR cemented vs. BHR (95% CI)	Predicted BHR *S*(*t*) (95% CI)	Predicted cemented *S*(*t*) (95% CI)	Absolute diff in *S*(*t*) BHR – cemented (95% CI)	Baseline hazard *h*(*t*) (per 1000 person-years)[**]
1 year	All males	1.667 (1.349 to 2.061)	0.990 (0.988 to 0.992)	0.983 (0.982 to 0.984)	0.006 (0.004 to 0.009)	0.923
1 year	Males <55 years	3.863 (1.824 to 8.182)	0.999 (0.998 to 1.000)	0.995 (0.993 to 0.998)	0.003 (0.001 to 0.006)	0.475
6 years	All males	1.667 (1.349 to 2.061)	0.921 (0.906 to 0.936)	0.877 (0.873 to 0.881)	0.044 (0.029 to 0.060)	1.848
6 years	Males <55 years	3.863 (1.824 to 8.182)	0.994 (0.990 to 0.998)	0.976 (0.966 to 0.985)	0.018 (0.008 to 0.028)	0.458

[**] Baseline hazard relates to a BHR patient, aged 50 years, ASA grade 1 having a non-complex procedure.

The population-averaged survival curves are given in Figure 2.11. The probability of survival over time was very high in males under 55 years of age receiving any of the three procedures. However, small differences between procedures were observed and patients receiving a BHR had the greatest probability of survival at 6 years with a probability of 0.994 (0.990 to 0.998).

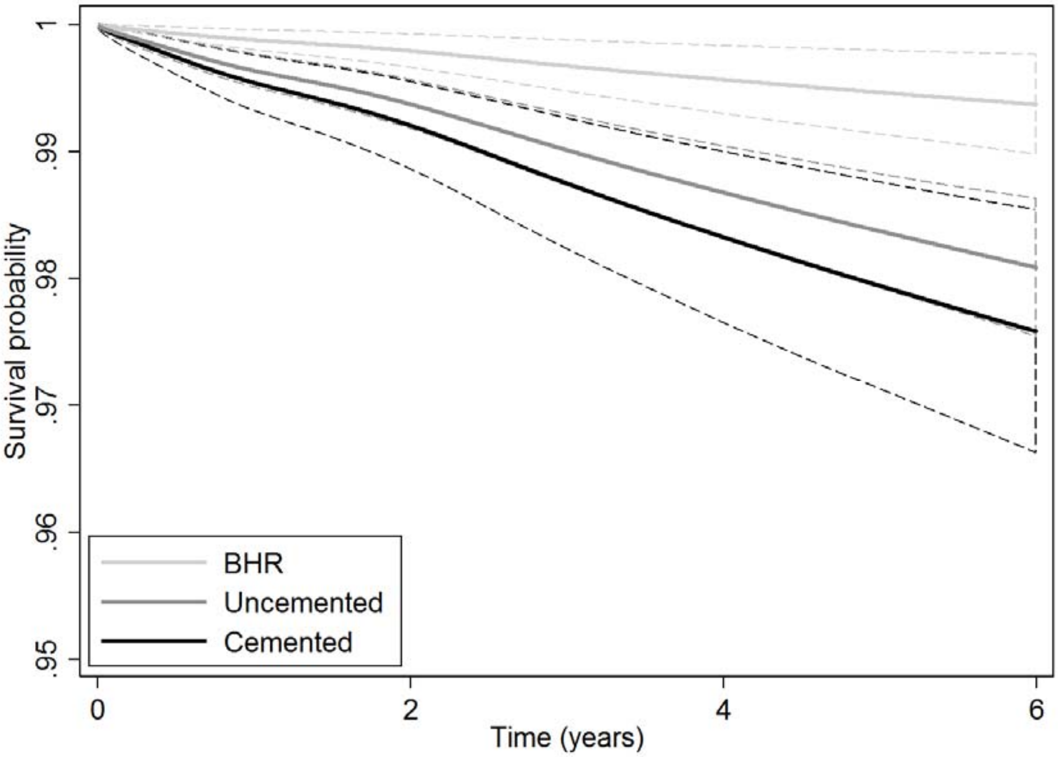


Figure 2.11: Adjusted survival curve for procedure type including BHR in men under the age of 55 where the outcome is mortality (95% CIs given by dashed lines).

Revision

There was no significant difference in revision rates between BHR and either cemented or uncemented procedures as seen in Table 2.13 (Wald test p=0.213 for procedure type).

Table 2.13: Multivariable Royston-Parmar model estimates for revision as the outcome in males under 55 years of age.

Variable	Hazard ratio	95% Confidence interval	P-value
BHR procedure	1	-	-
Cemented procedure	0.833	0.529 to 1.313	0.432
Uncemented procedure	1.185	0.860 to 1.634	0.300
ASA grade 1	1	-	-
ASA grade 2	1.404	1.045 to 1.886	0.024
ASA grade 3	1.654	0.897 to 3.051	0.107
ASA grade 4*	-	-	-
ASA grade 5*	-	-	-

*Not estimable

Figure 2.12 shows that the difference between survival curves of all three procedure types is very small.

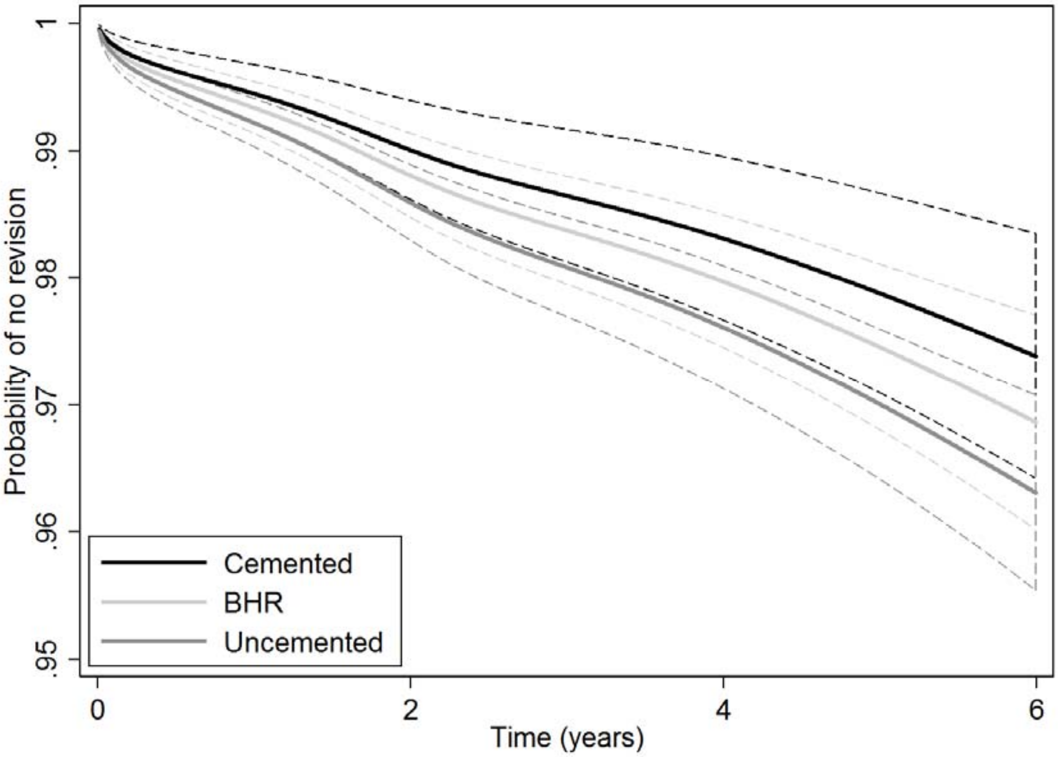


Figure 2.12: Adjusted survival curve for procedure type including BHR in men under the age of 55 where the outcome is revision (95% CIs given by dashed lines).

2.7 Discussion

In this chapter, an in-depth analysis has compared THR procedures in terms of mortality and revision, whilst identifying statistical advantages of Royston-Parmar models. Here, the key conclusions and limitations are summarised.

2.7.1 Summary of clinical findings

The clinical findings were published in the BMJ (Appendix A5).[86] Annual reports by the NJR have presented analyses of revision rates and reported differences between procedures, also using flexible parametric models. However, very limited analysis has been performed by the NJR for the outcome of mortality.[89,101,102] The NJR annual reports give unadjusted 30 and 90-day mortality rates but do not include mortality analyses of the same depth as for revision; in particular, adjusted analyses are not presented.

This study of the NJR data found small but significant differences between survival probabilities of patients receiving cemented and uncemented procedures. After adjusting for age, gender, ASA grade and complexity of procedure, the hazard of death was 11.1% (95% CI: 6.9% to 15.5%) higher in patients that received a cemented THR compared to an uncemented THR, assuming patients had not already had a revision. This hazard ratio corresponded to a difference in mean predicted survival probabilities of 0.013 (95% CI: 0.007 to 0.019) at 8 years. The hazard of revision was 42.2% (95% CI: 37.9% to 46.1%) lower in patients that received a cemented compared to an uncemented THR and this corresponded to a difference in mean predicted probabilities of 0.015 (95% CI: 0.012 to 0.017) at 8 years, assuming they had not died before this time. Although these differences are small, due to the number of hip replacements that are performed each year, if genuine it could still have important implications and is therefore of public health interest. There were around 71 500 primary hip replacements recorded by the NJR for 2011.[102] If hypothetically, these were all planned to be cemented procedures and changed to receive uncemented procedures, assuming the model results to be true, a predicted 929 fewer deaths (95% CI: 500 to 1359) would occur within 8 years, for patients who had not received a revision by this time. The analysis of males under the age of 55 suggested an increased hazard of mortality in cemented and uncemented THRs relative to BHRs but no significant difference between procedures for revision (p=0.213).

After examining each ASA grade separately, there was a significant increase in hazard associated with cemented rather than uncemented procedures found within ASA grade 2 (p<0.001) and ASA grade 4 patients (p=0.032). The procedure hazard ratio was not significant in ASA grade 1 (p=0.885) or ASA grade 3 (p=0.328) and was unestimable in ASA grade 5 patients. A significant procedure hazard ratio for revision was found in ASA grades 1, 2 and 3 (all p<0.001) but was unestimable in ASA grade 4 and 5 patients. If there was a genuine increased mortality risk from cemented, it would be expected to appear in each of the subgroups of patients even if the effect size differed, but it does not. Therefore, this raises some doubt about the causality of the cemented effect for mortality as it was not significant in all grades even with large numbers. However, the reduced revision risk for cemented procedures was observed in each ASA grade for revision.

Since this work was published in the BMJ, other studies have also been published supporting some of the findings of this chapter. One study compared metal-on-metal hip resurfacing to cemented and uncemented THRs using a propensity matched analysis of the hospital episode statistics database, and found that patients receiving a metal-on-metal resurfacing had a higher survival probability than patients receiving cemented (HR=0.51, 95% CI:0.45 to 0.59) and uncemented (HR=0.55, 95% CI:0.47 to 0.65) THRs.[103] Confounders such as a comorbidity index, rurality, area deprivation and surgical volume were included in their study but not available in the analyses included in this chapter. This suggests that even after adjusting for further confounders, the difference in mortality rates between hip resurfacings and THRs still remains. A systematic review and meta-analysis of randomised controlled trials comparing cemented and uncemented THRs concluded that there was no significant difference in revision rates between cemented and uncemented THRs.[104] They also found no difference in the mortality rates between cemented and uncemented THRs. These findings could be due to relatively few RCTs in the literature, and thus low power to detect small differences, compared to the large cohort studies used in this chapter and other published studies. Events for both mortality and revision are relatively rare so the sample size required to detect any 'significant' differences would need to be large. If an RCT was to be powered for the difference observed between cemented and uncemented procedures for mortality (hazard ratio=1.111), 2834 events would be required for 80% power and type I error of 0.05. Therefore, based on the 5-year estimated population-averaged survival probabilities, at least 33350 patients would need to be recruited to observe that number of deaths within 5 years (without even allowing for loss to follow-up). Another potential reason for no significant difference between procedures in the systematic review could be because patient characteristics should be more balanced in RCTs and differences seen between procedure types in cohort studies could be due to residual confounding not measured as patients are not randomised to procedure type.

One of the major limitations to the data used in this chapter is the limited number of variables recorded and the potential for residual confounding. The analyses have considered all information available in the dataset, however important factors such as smoking, activity level, deprivation levels and comorbidity scores were not recorded in the dataset and therefore could not be adjusted for. There may also be errors in the entry of data, for example patients recorded as ASA grade 5 are unlikely to receive hip replacement surgery as they are expected to die within 24 hours with or without an operation. Therefore it is more likely that these patients are ASA grade 1-4 but recorded as ASA grade 5 in error. Another limitation is that patients were only considered up to the time of revision if they had one, and so mortality after revision was not considered. Further, clinical studies with additional patient information should be undertaken to confirm any differences in mortality and revision found in these analyses.

2.7.2 Statistical advantages of flexible parametric models in this dataset

One of the key features of the Cox proportional hazards model is that it is semi-parametric.[105] This means that no distributional assumptions are made for the baseline hazard function. This makes Cox models very easy to fit and the most popular modelling technique for time-to-event data. Parametric models such as the exponential or Weibull models use a parametric distribution for the baseline hazard function, and as such often do not adequately fit the underlying shape of the hazard function in real data as they are limited in the shapes that can be fitted.[34]

Although Cox regression models are very easy to fit and do not make any distributional assumptions, they are limited in their use for prediction modelling. Royston-Parmar models however, are just as easy to implement in software packages such as Stata and have many advantages due to the explicit modelling of the baseline (cumulative) hazard function. Throughout this chapter, six key statistical advantages of using Royston-Parmar models were identified.

ADVANTAGE 1: Estimating and displaying the baseline hazard function

Using restricted cubic splines with varying degrees of freedom makes it possible to model the cumulative baseline hazard function flexibly. Whereas standard parametric models are not very flexible and most cannot model functions that have turning points, Royston-Parmar models are very flexible. Software packages such as those in Stata also make it easy to display the baseline hazard function.

ADVANTAGE 2: For proportional hazards, the hazard ratio estimates are practically identical to Cox regression but one additionally obtains the baseline hazard

Royston-Parmar models still produce hazard ratios which (when the baseline hazard is modelled correctly) are almost identical to those obtained from a Cox proportional hazards model. However, by modelling the baseline hazard function, it is possible to make predictions over time that would not be possible using a Cox model.

ADVANTAGE 3: Population-averaged (`adjusted') survival curves

Population-average survival curves can be plotted to graphically show survival functions for groups of patients (for example procedure type), adjusting for other covariates in the model This is a good alternative to Kaplan-Meier survival curves which are generally produced for one variable and therefore are unadjusted step functions. Using Royston-Parmar models, predicted survival curves are averaged over the population assuming each procedure type for example, resulting in survival curves that are 'adjusted' for the other variables in the model.

ADVANTAGE 4: Estimating differences in absolute *S*(*t*) over time, after adjustment for covariates in the model

From the population-averaged survival curves, the differences in absolute *S*(*t*) can be calculated, after adjusting for covariates in the model. This gives the average difference in survival functions at particular time points and allows the absolute survival difference to be reported in addition to the hazard ratios that are usually reported.

ADVANTAGE 5: Making risk predictions for groups of patients and individuals

By modelling the baseline hazard function using Royston-Parmar models, predictions are possible for individuals by using their specific covariate values in the model and therefore being able to predict their survival and hazard over time. This is important for prognosis research and if such a model were used in clinic, it would be possible to use the information to help make clinical decisions for an individual based on their characteristics rather than averages for groups of patients.

ADVANTAGE 6: Identify significant hazard ratios that are clinically important in relation to baseline risk

Advantage 6 is an important one as a hazard ratio is only useful when we know what it is relative to i.e. the baseline. Using the hazard ratio from the under 55 male only analysis as an example, it is possible to conclude that there is a significant four-fold increase in the hazard of mortality for patients that received the cemented procedure compared to patients that received the BHR (HR=3.96, p<0.001). This hazard ratio is very similar to that obtained by fitting a multivariable Cox model. The hazard ratio only gives the hazard of death for cemented patients *relative* to BHR patients. This does not give any information about their absolute risk of mortality which is low with a (predicted) mean probability of death of 0.024 in patients that received a cemented procedure compared to 0.006 in the BHR group at 6 years resulting in a small difference in mean probabilities of 0.018 at 6 years.

2.7.3 Potential pitfalls and situations when Royston-Parmar models are not required

Royston-Parmar models may not always be necessary to use instead of Cox proportional hazards models. In some cases, only relative measures such as hazard ratios may be required. For example, if testing whether a new prognostic factor is significantly associated with the outcome of interest. The benefits of modelling the baseline hazard are clear when absolute risk prediction is of interest as the baseline hazard function facilitates absolute risk prediction over time and individualised predictions. However, it is also important to ensure that there is adequate data (sufficient number of events) available as additional d.f. are used to fit the restricted cubic spline function for the baseline hazard. Caution should also be taken not to overfit the baseline hazard function to the data. Overfitting of the baseline hazard could occur if too many knots are used and/or there are an insufficient number of events.

2.7.4 Further work

As an extension, competing risks survival models might be used for the analysis of revision accounting for the competing risk of death, in order to derive absolute risk predictions in a ‘real world’ setting where death can prevent revision occurring. In this chapter, analyses for mortality assumed that a patient had not had a revision before that time, as individuals that had a revision were censored at their revision time. The results of the mortality analyses are therefore in a hypothetical setting where no revision would occur beforehand. Ideally, mortality information would be available even if a patient had subsequent revisions after the initial THR. Methods for handling competing risks involve fitting separate models for the competing events which has been done in this study or fitting a model stratified by competing events.[34,106] This would need some exploration as strictly speaking, death is a competing event for revision but revision is not a competing event for death, even though patients have not been followed up for death after having a revision.

As a further analysis, patients could be matched using propensity scores. This approach attempts to account for the differences in baseline characteristics between treatments (in this case, procedure type) by matching patients on their propensity score,

$$\Pr(T_i\!=\!1|\boldsymbol{X}_i)$$

which is the conditional probability of receiving treatment T based on covariates $\boldsymbol{X}$ for individual i, usually calculated using a logistic or probit model.[107,108] The results of such an analysis could be compared to those reported in this chapter. However, residual confounding could still be an issue and cannot be dealt with properly until additional variables (such as smoking and comorbidities) are recorded or made available. In addition to this, the propensity score approach may not be any better than multivariable modelling. In an editorial by Winkelmayer and Kurth, the authors say that in most situations ‘the use of propensity scores has no apparent advantage compared with traditional methods’.[109] Also saying, ‘Only if the outcome is rare relative to the number of confounders and the number of study subjects in the smaller exposure group is sufficiently large to warrant multivariable propensity score estimation, then this statistical technique has a legitimate role to potentially reduce bias and expand the possibilities in observational outcomes research’.[109,110]

2.8 Conclusion

Clinically, a small but significant difference has been found between cemented and uncemented procedures suggesting that patients receiving a cemented THR had a higher mortality rate but lower revision rate than patients receiving an uncemented THR. Whether this is genuine or due to residual confounding requires further research. Ideally a RCT in which the baseline characteristics of patients receiving the different procedure types were balanced could give more conclusive results as to whether the observed effect is real. However, due to the length of follow-up required and the number of patients that would be needed to detect such a small difference, this is unlikely to happen. In addition to this, sometimes a particular procedure may be best suited to an individual due to the requirements of that individual, for example, older patients receiving cemented THRs and young active males receiving BHRs. It is important to consider the other advantages and disadvantages of each procedure when the differences in mortality and revision are small.

Statistically, this dataset also highlighted the importance of modelling the baseline hazard function and the advantages of using Royston-Parmar models rather than Cox regression models. This is particularly important in prediction modelling, where ideally predictions should be made on an individual basis for patients and risk of an outcome presented in absolute terms rather than hazard ratios alone or predictions for groups of patients.

In the following chapter, a literature review is performed to assess how clinical prediction models have been developed and reported in some of the leading journals. This will give an indication of the types of models fitted to time-to-event data, if and how the baseline hazard function is modelled and how results are reported. In particular, to ascertain whether flexible parametric modelling is being used in making predictions for individuals and if not, summarise what methods are being used for absolute risk predictions.

CHAPTER 3: ESTIMATING THE BASELINE HAZARD AND ABSOLUTE RISK IN MULTIVARIABLE PREDICTION MODELS: A REVIEW OF CURRENT PRACTICE

3.1 Introduction and objectives

Chapter 2 illustrated how the baseline hazard can be modelled using Royston-Parmar models. It also highlighted some of the statistical advantages of modelling the baseline hazard, such as the ability to predict survival curves for individuals as well as for groups of patients. Royston-Parmar models were first published in 2001 and can easily be fitted in software packages such as Stata and R.[32,34,40,111] Therefore, it is of interest to evaluate if and how these (or other methods for baseline hazard estimation) are being implemented in the development of multivariable prediction models using time-to-event data. Such models consider multiple prognostic factors in combination, usually to examine the independent prognostic value of each factor and/or to use the overall model for making absolute risk predictions for new individuals. Therefore modelling of the baseline hazard could be advantageous but do researchers model the baseline hazard at all and, if not, how do they propose to use their model for making absolute risk predictions?

The primary objective of this chapter is therefore to review published journal articles that develop a multivariable prediction model (for any purpose), and to thereby identify:

(i) If the baseline hazard is being modelled in practice and, if so, what methods are being used.

(ii) How absolute risk predictions are being presented from the developed model and whether the baseline hazard is being used to do so.

(iii) How the developed model enables absolute risk predictions for new individuals and whether the baseline hazard is utilised toward such predictions. This will enable the different methods used for development and validation of the models to be identified and summarised, especially with regard to estimation of the baseline (cumulative) hazard function and absolute risks.

Secondary objectives of the review are to:

- Assess how the fitted models were reported; for example, how the baseline hazard was reported (if modelled) and whether beta estimates were reported or only hazard ratios.
- Assess if and how the proportional hazards assumption was checked.
- Assess how continuous variables were included in the multivariable model i.e. linearly, categorised or transformed.
- Summarise how authors validated their prediction model (where relevant).

The findings of the review should therefore help identify areas for improvement, current good practice, and recommendations for better use of the baseline hazard in prediction research. Note that a review published by Bouwmeester et al. aimed to investigate the reporting and methods used for prediction studies.[85] The authors conducted a systematic review of all prediction studies published in six high impact journals in 2008. Some of the studies included in the Bouwmeester et al. review are also included in the review done in this chapter; however, whereas they were interested in all prediction studies, this chapter is only interested in prediction models using time-to-event data as primary interest is in the baseline hazard function and how absolute risk predictions are derived.

3.2 Method

3.2.1 Identifying a set of articles for review

A database of multivariable prediction models was available to use for the literature review. This was kindly provided by collaborators at UMC Utrecht, Prof. Karel Moons and Thomas Debray. The database was created as part of a systematic review looking at the development, validation and impact of prediction models, published as part of the PROGRESS series.[3] The database consisted of 71 prediction models, published in six leading clinical journals between 2006 and 2009. These journals were:

- Annals of Internal Medicine
- British Medical Journal
- Journal of American Medical Association
- Lancet
- New England Journal of Medicine and
- PLOS Medicine

3.2.2 Inclusion/exclusion criteria

The original review and database included publications describing the development, external validation, or impact assessment (or combination) of a prognostic model and no other selection criteria. From this database, articles were screened and included in this current review if a multivariable prediction model was developed for time-to-event data and survival models were used. For this reason, prediction models developed using logistic regression or other methods not suitable for time-to-event data were excluded. The word ‘developed’ here is used to encompass any type of risk prediction model that was either completely newly created or was a modification of an existing model (e.g. removal or addition of one or more predictors). Exclusions were also made for articles modelling data that were not right censored, for example data that used interval censoring, as only right censored data are of interest in this thesis.

3.2.3 Review process

The selection of articles for inclusion was done independently by two reviewers (Snell and Debray) and agreed upon with discussion from a third reviewer (Riley) if necessary. Note that the review did not aim to identify all published prediction models, but rather to obtain a sample in leading medical journals for qualitative evaluation. The review team felt that qualitative saturation of the issues and methods could be obtained using this existing database. Also, as the review aimed to evaluate methodology used for model development for clinical prediction, no restriction was considered necessary with regard to the patient population at baseline. This was consistent with the previous work by Bouwmeester et al.[85] Therefore relevant articles could be for either healthy or diseased individuals at baseline or, in other words, prognostic models (diseased) and risk prediction models (healthy) were both included as long as time-to-event data was of interest using survival models.

3.2.4 Evaluation of relevant articles

A protocol specifying the aims of the literature review and the information to be recorded from each relevant journal article was written prior to starting the review. The protocol was then checked and amended by the review team before being finalised. A spreadsheet containing all relevant questions was used by each reviewer to record information from all relevant journal articles independently. The first reviewer (Snell) then compared and combined (where necessary) the information recorded by both reviewers. Where reviewers differed in their answers, the article was read a further time and an answer agreed upon by the first reviewer (Snell) and third reviewer (Riley). A summary of the information collected for each article is given below and the full protocol can be found in Appendix B1. Five sections were of interest:

Background

Background information was collected to give an overview of the types of studies included in the literature review and the clinical research areas they covered. This included the type of populations and disease areas of interest, the starting point at baseline (for example, healthy patients recruited, diagnosis of a disease etc.), the primary aims of the study (such as evaluation of candidate risk/prognostic factors, or derivation of a model for absolute risk prediction) and the outcome modelled (such as death in patients with a disease or diagnosis of a disease in healthy patients).

Description of development data

Information was collected from each article about the data used to develop the prediction models, including a summary of model sample size, number of events and candidate predictors, as well as summarising whether authors reported missing data, their approach to dealing with missing information and how the length of follow-up was reported.

Model development methods and baseline hazard

Information was collected from each article about the techniques used to develop the prediction models, whether the baseline hazard was explicitly estimated and, if so, how. Also, whether the authors checked the proportional hazards assumption (if a proportional hazards model was fitted), how continuous variables were modelled, whether univariable analyses were reported in addition to multivariable analyses, and how variables were selected for inclusion in the multivariable model.

Reporting of results and presentation of absolute risk

For each article, information was extracted about how results were reported from the analyses, for example whether authors reported hazard ratios or the betas (log HRs) from the developed model. Also, for articles that modelled the baseline hazard, details on how they reported it were recorded. It was also of interest to record how estimates of absolute risk were reported and presented, and to summarise the process needed to obtain absolute risk estimates from the developed model.

Validation

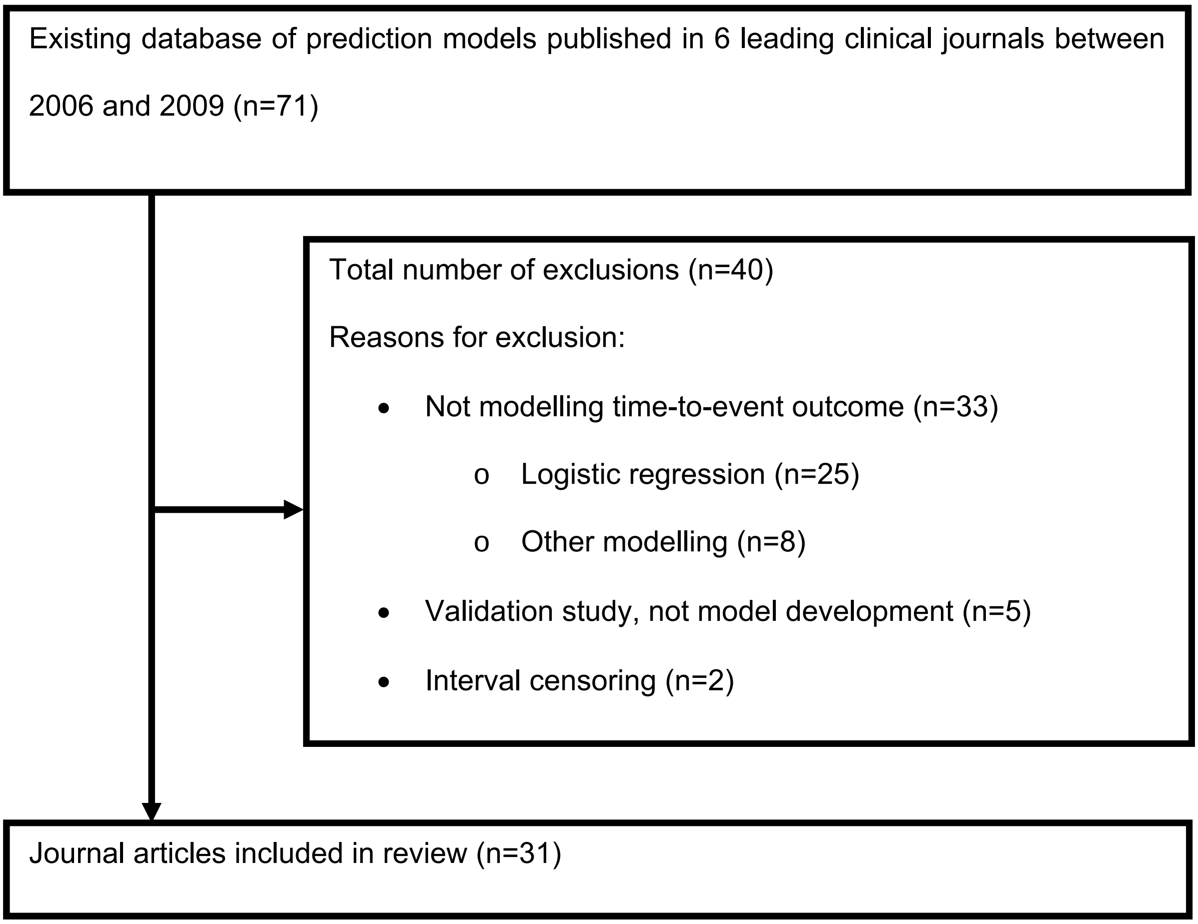
All articles included in the review developed a multivariable model. For the articles that included model validation in addition to model development, the methods used for validation were also summarised. This included whether internal and/or external validation was performed (and how), what validation statistics were reported, whether the baseline hazard was compared or discussed (if modelled), and whether absolute risk was compared between the development and validation datasets (or between groups within a dataset).

3.3 Results

3.3.1 Identification of relevant articles

The original database of multivariable prediction models contained 71 articles published between 2006 and 2009. From this, 40 were excluded leaving 31 articles eligible for inclusion in the review (Figure 3.1, see Appendix B2 for full list of articles excluded). Thirty-three of the exclusions were due to the methods used for development of the prediction models not modelling a time-to-event outcome: 25 used logistic regression and 8 articles used other methods including linear regression, Poisson mixed models for rate per 1000 live births (not rate over time) and Bayesian models such as a naive Bayesian classifier which aims to predict future risk of an event at *any* time point rather than modelling a rate over time (see

Appendix B2). Five articles were excluded because the primary aim was to *validate* an existing prediction model and not to develop a model, and two further articles were excluded because of interval censored data , rather than right censored.



Existing database of prediction models published in 6 leading clinical journals between 2006 and 2009 (n=71)

Total number of exclusions (n=40)

Reasons for exclusion:

- Not modelling time-to-event outcome (n=33)
    - Logistic regression (n=25)
    - Other modelling (n=8)
- Validation study, not model development (n=5)
- Interval censoring (n=2)

Journal articles included in review (n=31)

Figure 3.1: Flow diagram of journal articles for inclusion in literature review.

3.3.2 Summary of articles included in the review

The 31 articles included in the literature review published studies that included the development of a multivariable prediction model using time-to-event data. Table 3.1 gives a summary of the types of articles that were included in the review. The two most common disease areas were cardiovascular disease in 20 articles (64.5%) and oncology in seven articles (22.6%).[112-118] Other research areas included diabetes,[119,120] liver transplantation,[121] HIV/AIDS,[117] hormonal contraceptive,[122] and osteoporotic fracture.[123] Two articles covered more than one research area.[117,120] Three articles reported genetic studies in oncology.[112-114] These studies were genetic in that they focused on gene dosage and/or gene expression in relation to survival.

Study populations

The patient populations differed across the articles: 13 (41.9%) related to patients with a disease or condition of interest at baseline,[112-117,120,121,124-128] 17 articles (54.8%) used healthy individuals at baseline,[118-120,122,123,129-140] and three of the articles included patients with suspected disease or individuals at high risk of the disease of interest.[120,141,142] One article included all three patient populations as it included two studies, one which consisted of patients with cardiovascular disease or were considered to have a high risk of cardiovascular disease, and the second study included healthy patients free from cardiovascular disease.[120] One article differed from the others in that it did not try to predict onset of a disease in healthy patients nor an outcome in patients with a disease, but instead followed healthy males who had received a hormonal contraceptive and looked at the time to recovery of sperm to various thresholds after treatment had stopped.[122]

Study objectives

Though all studies developed a multivariable prediction model, the primary aim of the studies varied widely. Fourteen of the articles (45.2%) aimed to develop a completely new prediction model either to be used in a clinical setting or elsewhere for making absolute risk predictions,[115,117-119,123-125,128,134,136,138-140,142] and of these, three compared the developed model to an existing model.[134,138,139] Six articles (19.4%) aimed to develop a new model by adding new factors to an existing model and then comparing the models.[129,132,133,135,137,141]

Two articles (6.5%) aimed to compare model performance as their primary aim.[126,131] For example, Gaziano et al. compared a model that included laboratory information to a model that did not include laboratory information in the assessment of cardiovascular risk, but did not give details on how the variables were selected for each model.[131]

The remaining nine articles (29.0%) had a primary aim to investigate association between the variables in the model and the outcome, therefore using the developed model for absolute risk prediction was not the main intention.[112-114,116,120-122,127,130] However, four of these nine articles still reported or considered absolute risk in some way (see Table 3.1).[112-114,122]

Table 3.1: Summary of research areas, populations included and whether absolute risk estimation was of interest.

Article (First Author, ref.)	Clinical research area	Population at baseline	Interested in absolute risk estimates? If not, what was of interest?
Daly[124]	Angina	Patients newly diagnosed with stable angina	Yes
Fox[125]	Acute coronary syndrome (ACS)	Patients with acute coronary syndrome	Yes
Patel[116]	Prostate cancer	Patients with T1c to T3b, node negative and nonmetastatic adenocarcinoma of the prostate	No, compared prognostic significance of Gleason score 7 with tertiary grade 5 vs. other Gleason scores.
Schnabel[140]	Atrial fibrillation	Individuals without atrial fibrillation (retrospectively selected)	Yes
Montalvo[126]	Hypertension	Patients with hypertension	Yes, used in comparing predictive power of two models.
Sekhri[141]	Angina	Patients with suspected angina	Yes, to assess incremental prognostic value of exercise ECG.
Buckley[127]	Angina	Patients newly diagnosed with angina	No, only associations between risk factors and outcome by reporting HRs.
de Ruijter[129]	Cardiovascular disease (CVD)	Patients without history of CVD (birth cohort)	Yes, used in comparison of models with new biomarkers in terms of predictive performance.
Stebbing[117]	HIV/AIDS & associated cancer	Patients with AIDS-associated Kaposi's sarcoma	Yes
Parimon[115]	Allogeneic hematopoietic cell transplantation (HCT) and cancer	Patients having first allogeneic HCT	Yes
Lauer[142]	Coronary artery disease (CAD)	Patients with suspected coronary disease and normal electrocardiogram	Yes
Denes[130]	Coronary heart disease (CHD) and cardiovascular disease	Healthy postmenopausal asymptomatic women with an intact uterus	No, examined association of ECG abnormalities with outcome using HRs.
Moylan[121]	Liver transplantation	Patients registered on liver transplantation waiting list pre- and post-MELD	No, association between factors and outcome pre- and post-MELD.
Liu[122]	Spermatogenesis & hormonal contraceptive	Healthy eugonadal men	Yes, look at recovery probability over time.
Gaziano[131]	Cardiovascular disease	Patients without history of CVD	Yes, used to compare models.

Table 3.1 continued…

Article (First Author, ref.)	Clinical research area	Population at baseline	Interested in absolute risk estimates? If not, what was of interest?
Sattar[120]	Cardiovascular disease & diabetes	PROSPER: non-diabetics with pre-existing vascular disease or raised risk of disease. BRHS: non-diabetic males without (self-reported) CVD.	No, interested in associations between metabolic syndrome and risk of outcomes by reporting HRs
Melander[132]	Cardiovascular disease	Patients without CVD (population based study)	Yes, used to compare models to assess inclusion of contemporary biomarkers.
Rassi[128]	Chagas' heart disease	Patients with Chagas' disease	Yes
Chen[113]	Non-small-cell lung cancer (NSCLC)	Patients who underwent surgical resection of NSCLC.	Yes, used to compare risk groups for gene-signatures developed.
Zethelius[133]	Cardiovascular disease	Elderly men (around age 71)	Yes, used to assess improvement in risk stratification by adding biomarkers to model.
Cook[134]	Cardiovascular disease	Healthy non-diabetic women free of CVD and cancer	Yes
Ingelsson[135]	Coronary heart disease	Patients without CVD	No, reported HRs and compared performance of models using discrimination, calibration and risk reclassification.
Ridker[136]	Cardiovascular disease	Healthy women free of CVD and cancer	Yes
Paynter[137]	Cardiovascular disease	Healthy women free of major chronic disease including CVD and cancer	Yes, to assess whether a particular genetic variation improved risk prediction.
Hippisley-Cox[138]	Cardiovascular disease	Patients without diabetes and CVD	Yes
Crijns[114]	Ovarian cancer	Patients having surgery for advanced stage serous ovarian cancer	Yes, to show high and low risk profiles for the gene expression profile that was developed.
Tice[118]	Breast cancer	Women undergoing mammography with no previous diagnosis of breast cancer	Yes, developed tool for 5-year risk prediction
Hippisley-Cox[119]	Type II diabetes	Patients without prior diagnosis of type I or II diabetes (primary care patients)	Yes
Bredel[112]	Gliomas	Patients with gliomas	Yes, to compare survival for risk groups of the multigene risk scoring model.
Hippisley-Cox[139]	Cardiovascular disease	Patients without CVD or cerebrovascular disease (primary care patients)	Yes
Hippisley-Cox[123]	Osteoporotic fracture	Patients without previous hip, distal radius or vertebral fracture (primary care patients)	Yes

3.3.3 Development data description and size

Table 3.2 shows the sample sizes for each prediction model along with the number of events and candidate predictors. Sample sizes of the development cohort ranged from 63 in Chen et al. to 2540753 in Hippisley-Cox et al.[113,119] The problem of missing participant data was common with 26 of the 31 articles (83.9%) reporting this issue and the other five articles (16.1%) not reporting whether there was any missing data or not.[113,114,116,122,142] Complete case analysis was performed in 22 of the 26 articles (84.6%) with reported missing data, two of which (by the same first author) used a complete case analysis first and then used multiple imputation to account for the missing information,[123,138] while another article used complete case analysis but stated that 'imputation was tested but did not influence the identification of multivariable predictors or the discriminative power of the model'.[125] Another two articles just used multiple imputation (without mentioning the use of complete case analysis),[119,139] one other used median values for imputation,[117] and one was not clear on how they had dealt with missing data for model development.[118]

The number of candidate predictors ranged from 6 to 48 with a median of 13 (excluding two of the genetic studies investigating gene-signatures which considered hundreds or more genes). Using the rule of thumb for sample size required to develop a multivariable prediction model, there should be at least 10 events for each predictor considered in the model.[43] Using this rule, at least nine articles (29.0%) include at least one multivariable model in which the number of events was not large enough for the number of candidate predictors.[112,120,124,126-129,133,141] The ratio of events to predictors could not be calculated for six studies that did not report the number of events. Also worth noting is that the Hippisley-Cox et al. studies had extremely high power due to the use of a large database of health records (QResearch) consisting of information collected from around 530 practices, totalling over 2 million participants.[119,123,138,139]

Table 3.2: Sample size, number of events and candidate predictors by model.

First Author, ref.	Outcome	Description (if >1 model developed)	Sample size for development	No. Events	No. candidate predictors	No. of events per predictor
Daly[124]	Death/MI		2528	93	20	4.7
Fox[125]	Death		21688	1757	48	36.6
	Death/MI		21688	3110		64.8
Patel[116]	PSA failure		2370	613	6	102.2
Schnabel[140]	Atrial fibrillation		4764	457	21	21.8
Montalvo[126]	Cardiovascular events		504	76	15	5.1
	Death		504	74		4.9
Sekhri[141]	CHD death/ACS	Whole cohort	8176	576	21	27.4
		Summary ECG subset	4848	351		16.7
		Detailed ECG subset	1422	110		5.2
Buckley[127]	Acute MI		1785	116	10	11.6
	Coronary artery bypass grafting		1785	152		15.2
	Percutaneous transluminal coronary angioplasty		1785	108		10.8
	Death from ischaemic heart disease		1785	84		8.4
	Death from any cause		1785	175		17.5
de Ruijter[129]	Cardiovascular death		302	35	11	3.2
Stebbing[117]	Death		326	NS	13	
Parimon[115]	Death		1401	688	14	49.1
Lauer[142]	Death		33268	1619	12	134.9
Denes[130]	Incident CHD events		14749	246	11	22.4
	Incident CVD events		14749	595		54.1
Moylan[121]	Transplant	HCC	2365	1617	10	161.7
		Without HCC	43323	20353		2035.3
	Death	HCC	2365	197		19.7
		Without HCC	43323	6998		699.8
	Too sick for transplant	HCC	2365	187		18.7
		Without HCC	43323	2052		205.2

Table 3.2 continued…

First Author, ref.	Outcome	Description (if >1 model developed)	Sample size for development	No. Events	No. candidate predictors	No. of events per predictor
Liu[122]	Recovery to thresholds		1549	NS	19	
Gaziano[131]	Cardiovascular events		6186	3400	8	425
Sattar[120]	Incident CVD events	PROSPER	4812	772	13	59.4
		BRHS	2737	440	11	40
	Incident diabetes	PROSPER	4812	287	13	22.1
		BRHS	2737	105	11	9.5
Melander[132]	Cardiovascular events		4483	364	18	20.2
	Coronary events		4600	216		12
Rassi[128]	Death		331	130	24	5.4
Chen[113]	Death/recurrence of cancer	16-gene signature	63	NS	4 + 485 genes	
		5-gene signature	101	NS	4 + 16 genes	
Zethelius[133]	Death	Whole cohort	1135	315	19	16.6
		Without CVD	661	149		7.8
	Death from CVD	Whole cohort	1135	136		7.2
		Without CVD	661	54		2.8
Cook[134]	Cardiovascular events		15048	390	8	48.8
Ingelsson[135]	Cardiovascular events		3322	291	11	26.5
Ridker[136]	Cardiovascular events		16400	504	35	14.4
Paynter[137]	Cardiovascular events		22129	NS	12	
Hippisley-Cox[138]	Incident CVD events		1283174	65671	14	4690.8
Crijns[114]	Cancer death		157	NS	6 + 15909 genes	
Tice[118]	Diagnosis of cancer		377440	NS	7	
Hippisley-Cox[119]	Diagnosis of diabetes		2540753	78081	9	8675.7
Bredel[112]	Death		189	192	27	7.1
Hippisley-Cox[139]	Cardiovascular events		1535583	96709	13	7439.2
Hippisley-Cox[123]	Incident osteoporotic fracture		2357895	21184	18	1176.9
	Incident hip fracture		2357895	12369		687.2

NS: Not stated in the article.

Length of follow-up was reported using the median time in 14 articles (45.2%),[112-114,116,124,132,133,135-138,141,142] with interquartile ranges in six of the 14 articles. Mean follow-up was reported in six articles (19.4%),[120,126,128,130,134,139] but only one reported a standard deviation along with it.[126] The range of follow-up duration was reported in eight articles (25.8%), five of which had reported median follow-up and the other three had reported mean follow-up. The maximum follow-up was given in 12 articles (41.9%),[113,115,117,122,124,125,127,129-131,135,140] some of which were only observed from Kaplan-Meier plots rather than being reported in the text. Four of the 12 articles reporting maximum follow-up, reported this alongside other statistics such as the mean or median. Three articles (9.7%) reported the number of person-years for follow-up (one of which had also reported maximum follow-up duration),[119,123,131] and one article only reported the minimum follow-up duration, without other follow-up information.[121]

3.3.4 Model development methods

Modelling method

All 31 articles (100%) used the semi-parametric Cox proportional hazards model to develop the prediction models. Therefore none of the articles explicitly modelled the baseline hazard function (discussed in Section 3.3.6) and thus Royston-Parmar models or other such methods that model the baseline hazard were not utilised in this cohort, indicating the overwhelming dominance of Cox models in this field.

Proportional hazards assumption

The proportional hazards assumption was only checked in 17 (54.8%) of the studies. The other fourteen articles (45.2%) did not discuss the proportional hazards assumption or time-dependent effects of predictors; therefore it is unknown whether this was checked as part of the analysis.[113,114,117,119,120,124,126,128,129,131,134,136,139,141] None of their prediction models included time-dependent effects for any of the variables in the prediction model.

Continuous variables

Continuous variables may not have a linear association with outcome and therefore it is necessary to explore non-linear functions for such variables as part of the model development. In general, few authors specifically stated how continuous variables were modelled unless they used transformations or more complex non-linear functions. For this reason, information on how continuous variables were modelled was extracted primarily by observing how variables were reported in the results. Five articles (16.1%) used, or at least considered, fractional polynomials or restricted cubic splines for non-linear functions of continuous variables,[119,123,134,138,142] and nine (29.0%) used or considered a particular transformation (e.g. ln(x)) for continuous variables.[114,121,131-134,136,137,139] Ten articles (32.3%) categorised at least one continuous variable which may result in a loss of prognostic information from that variable.[112,113,115-118,120,126,128,130]

Variable selection

Variable selection using automatic methods such as stepwise, forward and backward selection methods was performed in 10 articles (32.3%).[113,115,117,121,122,124,125,128,132,136] In three of these 10 articles, automatic selection strategies were applied after selecting variables for inclusion based on univariable results.[115,121,125] Of the remaining 21 articles, two used the results of univariable analysis to determine variables to include in the multivariable model,[112,141] one of which included all significant variables (from the univariable analysis) and the other was not clear on the multivariable selection strategy.[141] Eleven articles (35.5%) pre-specified variables for the multivariable model and two articles (6.5%) included all variables in the multivariable model.[116,142] Other model selection strategies included principal component analysis in one article,[114] and selection based on BIC in three articles (9.7%).[119,123,139] The variable selection strategy in the other two articles was not clear.[118,140]

3.3.5 Reporting of results

Reporting of univariable analyses

Univariable analyses were reported in 11 articles (35.5%),[113,114,116,117,122,124,125,127,128,133,141] and age and sex adjusted analyses reported as a base analysis, rather than univariable analyses, in one article (3.2%).[140] Of the 11 articles that reported univariable results, all of them reported hazard ratios, 10 (90.9%) reported 95% confidence intervals, and eight (72.7%) gave p-values. Only one article reported beta estimates for univariable analysis.[117] The article that reported age and sex adjusted analyses as the base analysis instead of univariable analyses, reported the age and sex adjusted hazard ratios (with p-values).[140]

Reporting of multivariable analyses

Compared to univariable analysis, there was a wider variety in the statistics reported for multivariable/adjusted analyses with 26 of the 29 articles (89.7%) reporting hazard ratios, 22 (75.9%) reporting 95% confidence intervals, 21 (72.4%) reporting p-values, nine (31.0%) reporting beta estimates, six (20.7%) reporting standard errors and one article (3.4%) reporting the Chi-squared values.

Multivariable results (model beta estimates or HRs) were reported for all included variables in the model for 21 articles (67.7%), and one article (3.2%) reported the multivariable results but omitted two variables for which fractional polynomial functions had been fitted.[123] Seven articles (22.6%) reported adjusted results for only a partial set of included variables (rather than the full fitted model),[112,118,121,130,132,133,135] and two (6.5%) did not report any results from their multivariable model, only results on how well the model performed.[126,129]

3.3.6 Modelling the baseline hazard and reporting absolute risk predictions

Modelling the baseline hazard

One of the primary objectives of the literature review was to assess if and how the baseline hazard was modelled in the multivariable prediction models. All 31 articles used Cox proportional hazards models in which the baseline hazard is not explicitly modelled over time. However, eight articles (25.8%) estimated baseline survival/hazard at a particular time point.[118,119,123,134,136,138-140] The description of baseline hazard information was often very brief and without explicit details, and can be summarised across the eight articles as follows:

i. Four of the eight articles used a 10-year baseline survival estimate from the baseline survival function.[119,123,138,139] In these four articles (by the same first author), the baseline survival function was estimated as a step function that would be equivalent to the Kaplan-Meier estimate if there were no variables in the model (null model).[27] However, estimating the baseline survival function after fitting a Cox model results in a function that is adjusted for the variables in the model and in these four articles, the baseline survival function is centred on mean values of continuous variables (because variables were mean-centred in the model). The whole function was not of interest in these four articles and only the survival estimate at 10 years was extracted.

ii. One article used Kaplan-Meier estimates (product-limit estimator) for the average 10-year survival, calibrated to the Framingham data.[134] They stated 'to address the generalisability of the final WHS risk prediction model with hsCRP, we calibrated the predicted probabilities to observed risk in the Framingham Heart Study', and 'The projected 10-year risk from the WHS models was calibrated to the 10-year rate of cardiovascular outcomes among women in the Framingham data'. However, no further details are given of how their developed model was calibrated to the Framingham data.

iii. One article developed a risk prediction model for 10-year risk which includes a baseline survival term. However, the authors do not describe how this value was estimated.[136]

iv. One article reported 10-year baseline survival with the coefficients of their prediction model but did not state how this was estimated.[140]

v. One article used multiple stages to obtain age and race specific incidence rates and adjusted for those dying from other causes over time.[118] See section on 'Absolute risk for prediction in new individuals' for a description of the developed model for absolute risk prediction.

In summary, with the exception of (v) which used incidence rates as a baseline, seven of the eight articles all estimated baseline survival at a particular time point. Though the exact estimation methods for the baseline survival function were not detailed, it is likely that an approach available in statistics software such as the Kalbfleisch and Prentice estimator was used (see Appendix B3).[143]

Presenting absolute risks

The second primary objective was to assess how absolute risk was reported from the developed model. This section summarises how absolute risks from the developed models were reported, whilst the subsequent section considers how authors presented the model in a way that allow others to use it to derive absolute risks for new individuals. There is overlap between these two sections as sometimes the authors used the same approach for reporting their model as for telling others how to use their model for new predictions. Recall that 25 articles were in some way interested in absolute risk (Table 3.1) and six were not interested in absolute risk. For this reason, only 25 articles are considered in this section and the following section on 'Absolute risk for prediction in new individuals'.

Of the 25 articles that were interested in absolute risk, 14 articles provided absolute risks for reporting purposes only, five articles provided absolute risks for informing new prediction only and six articles provided absolute risks that could be considered for both reporting and new predictions (Figure 3.2). This section focuses on *reporting* of absolute risks, therefore only discusses the 20 articles that provided absolute risks for this purpose.

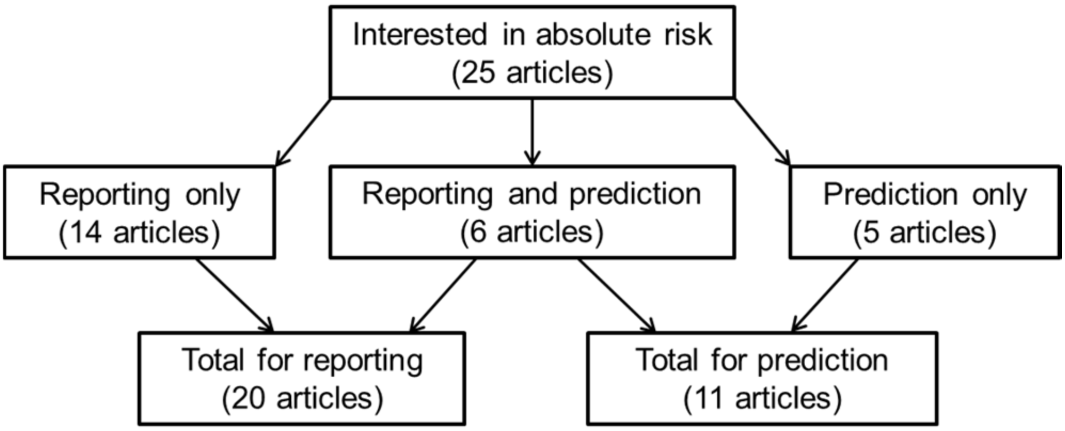

Interested in absolute risk (25 articles)
Reporting only (14 articles)
Reporting and prediction (6 articles)
Prediction only (5 articles)
Total for reporting (20 articles)
Total for prediction (11 articles)

Figure 3.2: Flow chart of articles interested in absolute risk.

Absolute risk estimates were reported in a variety of ways, both graphically and tabulated. Eleven of the 20 articles reported absolute risk over time and nine articles reported absolute risk at a particular time point. The 20 articles are now summarised:

Absolute risk over time

After a model was developed, the most common method to show absolute risks derived from the model was figures of cumulative risk or incidence such as Kaplan-Meier plots for risk groups, where the risk groups are created by categorising the risk score (linear predictor, $X\beta$) from a Cox model and a Kaplan-Meier curve is then generated separately for each group. This was done in nine articles, in which survival/incidence curves were plotted for between two and four risk groups.[112-115,117,126,128,129,132,137,142] An example of this is by Rassi et al. shown in Figure 3.3, in which the risk score was categorised into low, intermediate and high risk groups in the development and validation cohorts and survival curves produced using Kaplan-Meier estimates for each group.[128] Absolute risk over time can then be seen for each risk group from the curves. Another article produced survival curves for risk groups derived from a comparative model, but not the model developed in the article,[142] and one other article displayed 'adjusted' survival curves for three genotypes of a particular SNP. The authors did not give information on how this was done other than stating, 'Adjusted survival curves were generated by stratifying Cox proportional hazards models by genotype'.[137]

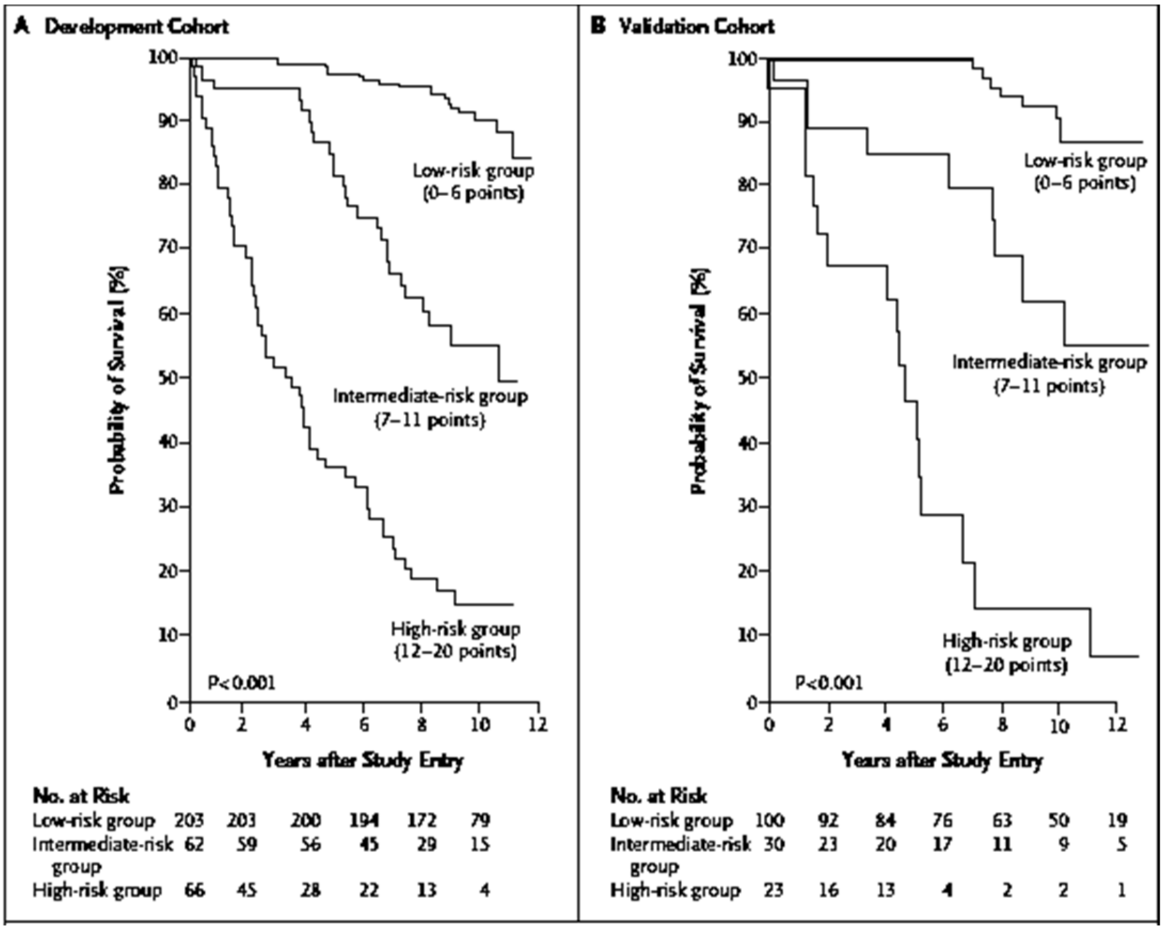
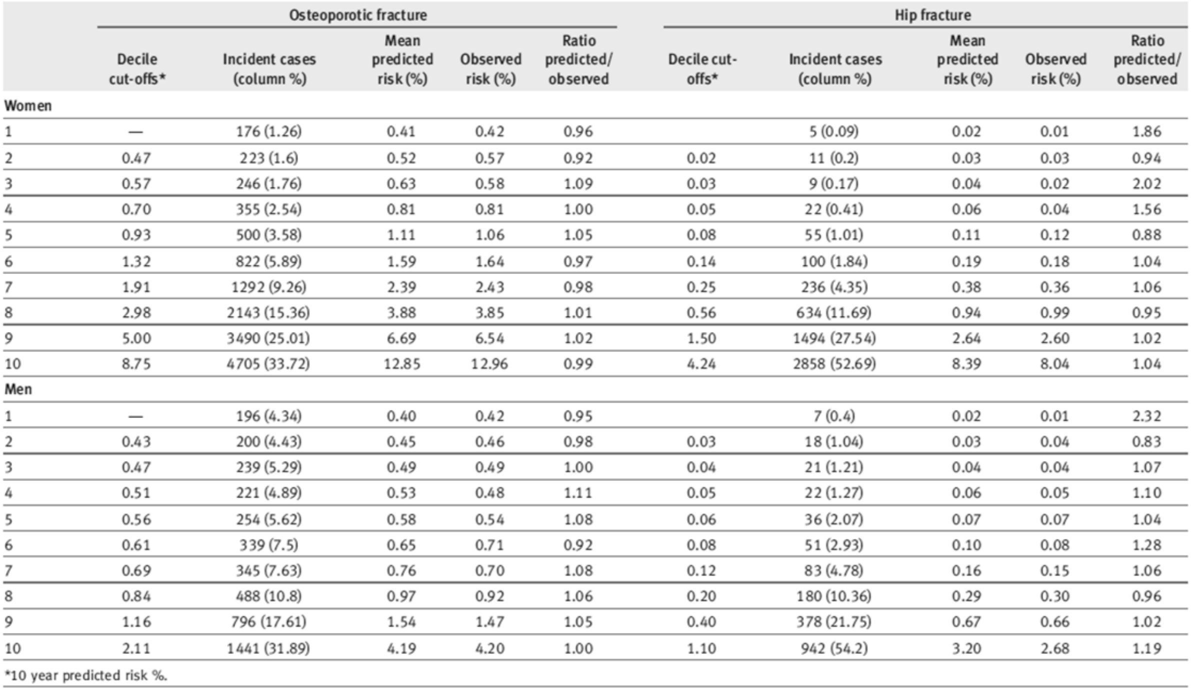


Figure 3.3: Example of how absolute risk estimates can be reported using Kaplan-Meier survival curves for risk groups derived from a developed multivariable prediction model. Reproduced with permission from Rassi et al.[128] Copyright Massachusetts Medical Society.

Absolute risk at particular times

Another way in which absolute risk was reported from the developed model was by displaying tables of risk score or risk groups and corresponding risk estimates at *one particular time point*. Two articles by the same first author, reported tables of deciles of risk score in men and women separately with the predicted and observed risk at 10 years (Figure 3.4).[123,138] Observed and expected risk or number of events were also plotted rather than tabulated in five articles.[119,123,131,139,140] Another article tabulated the cumulative probability of an event for each of three risk groups in three models of increasing complexity.[141] One article gave probability of recovery (of sperm after hormonal contraceptive) to different thresholds at four time-points. These probabilities were model-based and given for a hypothetical man.[122]

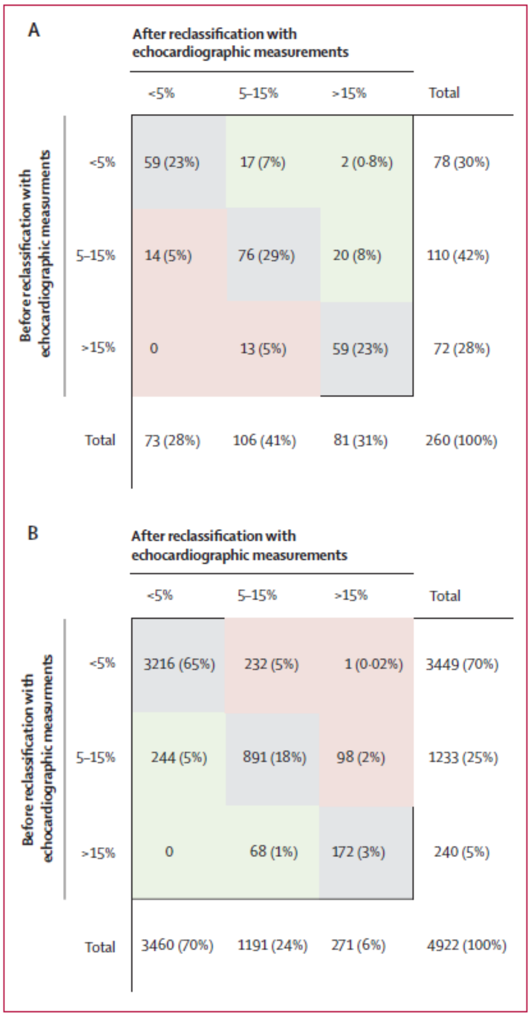


	Osteoporotic fracture					Hip fracture				
	Decile cut-offs*	Incident cases (column %)	Mean predicted risk (%)	Observed risk (%)	Ratio predicted/ observed	Decile cut-offs*	Incident cases (column %)	Mean predicted risk (%)	Observed risk (%)	Ratio predicted/ observed
Women										
1	—	176 (1.26)	0.41	0.42	0.96		5 (0.09)	0.02	0.01	1.86
2	0.47	223 (1.6)	0.52	0.57	0.92	0.02	11 (0.2)	0.03	0.03	0.94
3	0.57	246 (1.76)	0.63	0.58	1.09	0.03	9 (0.17)	0.04	0.02	2.02
4	0.70	355 (2.54)	0.81	0.81	1.00	0.05	22 (0.41)	0.06	0.04	1.56
5	0.93	500 (3.58)	1.11	1.06	1.05	0.08	55 (1.01)	0.11	0.12	0.88
6	1.32	822 (5.89)	1.59	1.64	0.97	0.14	100 (1.84)	0.19	0.18	1.04
7	1.91	1292 (9.26)	2.39	2.43	0.98	0.25	236 (4.35)	0.38	0.36	1.06
8	2.98	2143 (15.36)	3.88	3.85	1.01	0.56	634 (11.69)	0.94	0.99	0.95
9	5.00	3490 (25.01)	6.69	6.54	1.02	1.50	1494 (27.54)	2.64	2.60	1.02
10	8.75	4705 (33.72)	12.85	12.96	0.99	4.24	2858 (52.69)	8.39	8.04	1.04
Men										
1	—	196 (4.34)	0.40	0.42	0.95		7 (0.4)	0.02	0.01	2.32
2	0.43	200 (4.43)	0.45	0.46	0.98	0.03	18 (1.04)	0.03	0.04	0.83
3	0.47	239 (5.29)	0.49	0.49	1.00	0.04	21 (1.21)	0.04	0.04	1.07
4	0.51	221 (4.89)	0.53	0.48	1.11	0.05	22 (1.27)	0.06	0.05	1.10
5	0.56	254 (5.62)	0.58	0.54	1.08	0.06	36 (2.07)	0.07	0.07	1.04
6	0.61	339 (7.5)	0.65	0.71	0.92	0.08	51 (2.93)	0.10	0.08	1.28
7	0.69	345 (7.63)	0.76	0.70	1.08	0.12	83 (4.78)	0.16	0.15	1.06
8	0.84	488 (10.8)	0.97	0.92	1.06	0.20	180 (10.36)	0.29	0.30	0.96
9	1.16	796 (17.61)	1.54	1.47	1.05	0.40	378 (21.75)	0.67	0.66	1.02
10	2.11	1441 (31.89)	4.19	4.20	1.00	1.10	942 (54.2)	3.20	2.68	1.19

*10 year predicted risk %.

Figure 3.4: Predicted and observed fracture risk at 10 years for deciles of risk score in men and women. Originally published by Hippisley-Cox et al.[123] Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0).

Reclassification tables were included in two articles.[133,140] Such tables also include absolute risk information by a particular time-point, as groups of patients are categorised into risk groups using two comparative models to see how many are 'reclassified' based on the new model. An example of a reclassification table is given in Figure 3.5 in which risk groups were created for individuals with <5%, 5-15% and >15% risk of atrial fibrillation within 10 years.



A

		After reclassification with echocardiographic measurements			
		<5%	5–15%	>15%	Total
Before reclassification with echocardiographic measurments	<5%	59 (23%)	17 (7%)	2 (0·8%)	78 (30%)
	5–15%	14 (5%)	76 (29%)	20 (8%)	110 (42%)
	>15%	0	13 (5%)	59 (23%)	72 (28%)
	Total	73 (28%)	106 (41%)	81 (31%)	260 (100%)

B

		After reclassification with echocardiographic measurements			
		<5%	5–15%	>15%	Total
Before reclassification with echocardiographic measurments	<5%	3216 (65%)	232 (5%)	1 (0·02%)	3449 (70%)
	5–15%	244 (5%)	891 (18%)	98 (2%)	1233 (25%)
	>15%	0	68 (1%)	172 (3%)	240 (5%)
	Total	3460 (70%)	1191 (24%)	271 (6%)	4922 (100%)

Predicted risk of atrial fibrillation before and after reclassification with echocardiographic measurements for individuals who (A) did or (B) did not develop atrial fibrillation in 10 years

Figure 3.5: Example of reclassification table. Reprinted from the Lancet, 373, Schnabel et al.[140] Development of a risk score for atrial fibrillation (Framingham Heart Study): a community-based cohort study, 739-45, Copyright 2009, with permission from Elsevier.

The ways in which absolute risks were reported, as summarised above, included comparisons of observed and predicted risk as well as comparisons between development and validation datasets.

Absolute risk for prediction in new individuals

The previous section focused on how absolute risks were reported following the model development, and were primarily for the purpose of model checking and validation. This section now focuses on how authors reported absolute risk for the purposes of making predictions in new patients.

Eleven of the 25 articles presented the developed multivariable prediction model in a way that could be used by others to predict the risk of an event in new individuals. The most common method (six articles) was to assign points to each of the covariates in the model that can be summed to get the risk score. This is then accompanied by a table of risk scores and corresponding absolute risk up to a particular time point.[115,117,124,128,134,140] One of these articles included a figure of risk score by absolute risk in addition to a table, the risk score sheet and figure for absolute risk are shown in Figure 3.6.[124]

Two articles published a web address to an online risk calculator, which can be used to predict risk up to a specified time-point in new individuals.[125,140] Fox et al. included a snapshot of the online calculator in their article (Figure 3.7) which was referred to as a 'simplified nomogram'. An online risk calculator was published in addition to the point score system (discussed in the paragraph above) in the other article.[140] The four articles by Hippisley-Cox et al. did not include links to online risk calculators in the published articles (and are therefore not included in the 11 that published their model in a way that could be used for prediction), but these calculators have since been developed: QRISK2 for cardiovascular disease risk, QFracture for risk of osteoporotic fracture and QDiabetes for risk of type 2 diabetes.[119,123,138,139] QRISK2 predicts 10-year risk and the other two risk calculators can be set for between one and 10-year risk, at one year increments.

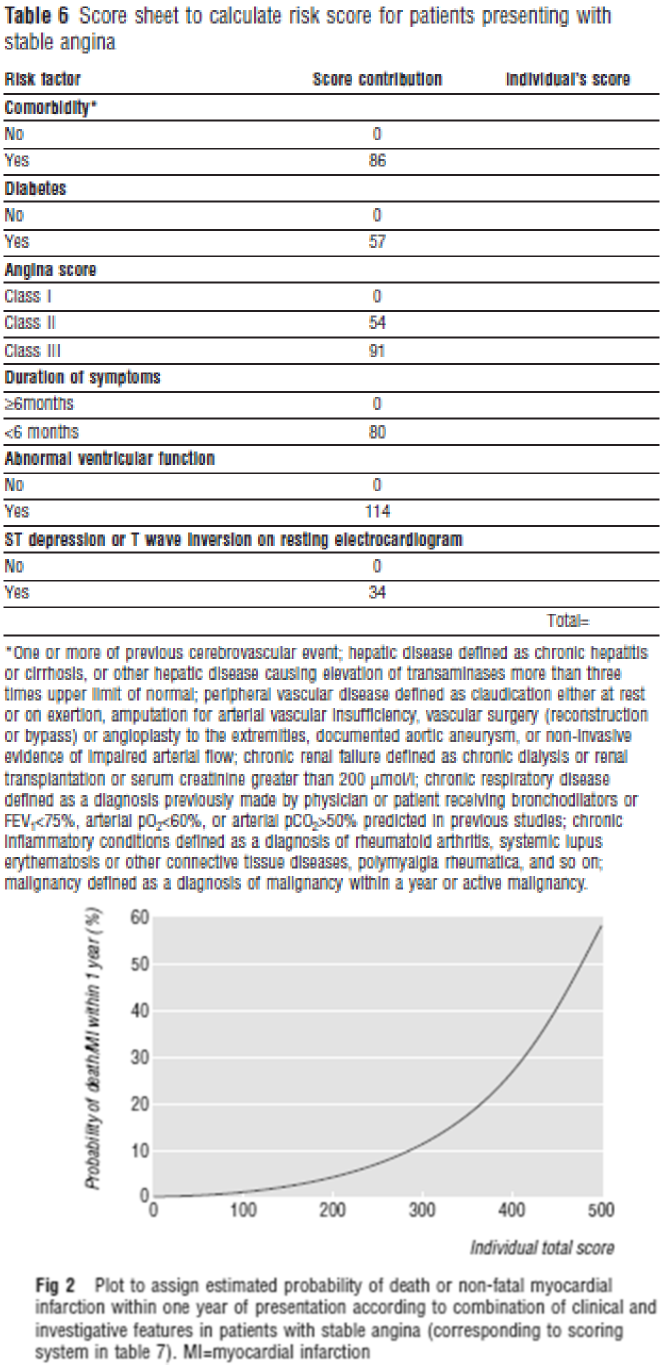


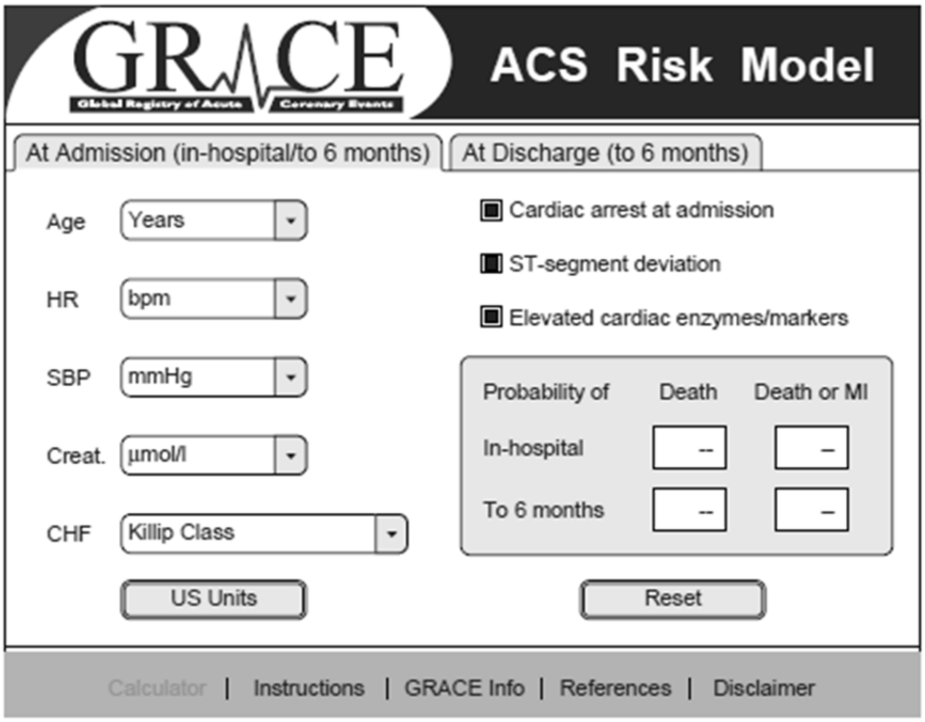
Table 6 Score sheet to calculate risk score for patients presenting with stable angina

Risk factor	Score contribution	Individual's score
Comorbidity*		
No	0	
Yes	86	
Diabetes		
No	0	
Yes	57	
Angina score		
Class I	0	
Class II	54	
Class III	91	
Duration of symptoms		
≥6months	0	
<6 months	80	
Abnormal ventricular function		
No	0	
Yes	114	
ST depression or T wave inversion on resting electrocardiogram		
No	0	
Yes	34	
		Total=

*One or more of previous cerebrovascular event; hepatic disease defined as chronic hepatitis or cirrhosis, or other hepatic disease causing elevation of transaminases more than three times upper limit of normal; peripheral vascular disease defined as claudication either at rest or on exertion, amputation for arterial vascular insufficiency, vascular surgery (reconstruction or bypass) or angioplasty to the extremities, documented aortic aneurysm, or non-invasive evidence of impaired arterial flow; chronic renal failure defined as chronic dialysis or renal transplantation or serum creatinine greater than 200 μmol/l; chronic respiratory disease defined as a diagnosis previously made by physician or patient receiving bronchodilators or FEV_1<75%, arterial pO_2<60%, or arterial pCO_2>50% predicted in previous studies; chronic inflammatory conditions defined as a diagnosis of rheumatoid arthritis, systemic lupus erythematosis or other connective tissue diseases, polymyalgia rheumatica, and so on; malignancy defined as a diagnosis of malignancy within a year or active malignancy.

Fig 2 Plot to assign estimated probability of death or non-fatal myocardial infarction within one year of presentation according to combination of clinical and investigative features in patients with stable angina (corresponding to scoring system in table 7). MI=myocardial infarction

Figure 3.6 Example of how a risk score can be reported and equated to risk estimates at a given time point. Reproduced from BMJ, Daly et al.[124] 332, 262-5, 2006, with permission from BMJ Publishing Group Ltd.


GRACE
Global Registry of Acute Coronary Events
ACS Risk Model
At Admission (in-hospital/to 6 months)
At Discharge (to 6 months)
Age
Years
HR
bpm
SBP
mmHg
Creat.
µmol/l
CHF
Killip Class
■ Cardiac arrest at admission
■ ST-segment deviation
■ Elevated cardiac enzymes/markers
Probability of
Death
Death or MI
In-hospital
To 6 months
--
US Units
Reset
Calculator | Instructions | GRACE Info | References | Disclaimer

GRACE risk calculator for death or myocardial infarction from admission to hospital to six months after discharge with the simplified model (www.outcomes.org/grace)

Figure 3.7: Example of an online risk calculator using a developed model. Reproduced from BMJ, Fox et al.[125] 333, p1091-4, 2006, with permission from BMJ Publishing Group Ltd.

A nomogram is a graphical form of a point score system that uses bars of varying length for each covariate in the model. The points corresponding to each risk factor can be summed for a total score which relates to a time specific risk or survival probability. Based on this definition, there was only one nomogram published.[142]

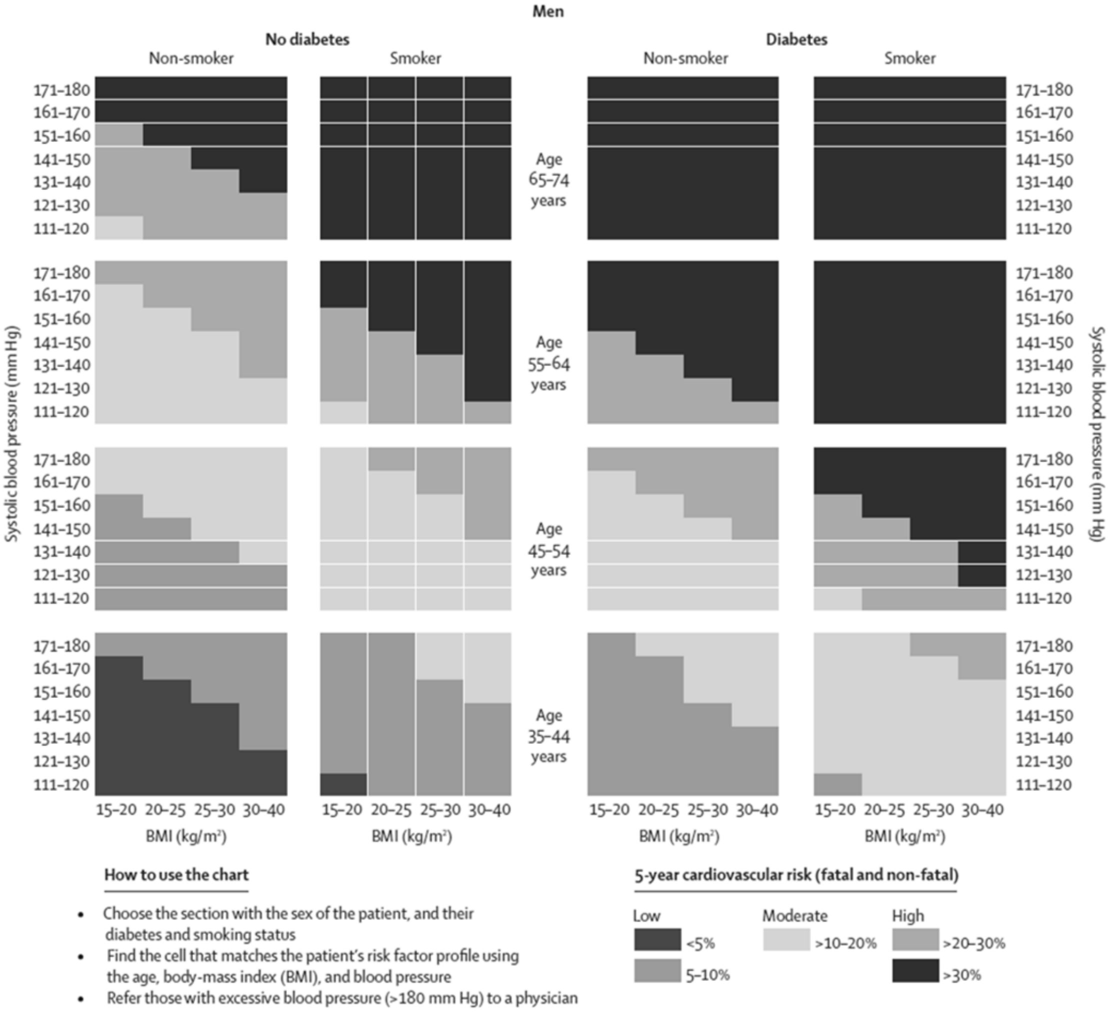
Two articles published an equation for absolute risk up to a specified time point (one was included in the appendix rather than the article itself).[134,136] The 10-year risk formula published by Ridker et al. was written out for 10-year cardiovascular risk which takes a similar form to the survival function given in Chapter 1 (Equation (1.9)).[136] The difference in the article was that rather than the equation being written for survival probability $S(t)$, it was written for percentage risk (=1 – $S(t)$ x100%) at a particular time point (t=10 years), taking the form

$$\text{Risk}(\%)=(1-S_0(t=10))^{e^{LP}}\times 100\%$$

The baseline survival term $S_0(t)$ can be estimated for a particular time point (here 10 years) as discussed in the section above on 'Modelling the baseline hazard'.

One article reported risk prediction charts for 5-year cardiovascular risk.[131] Separate charts were produced for men and women (see Figure 3.8 for the male risk prediction chart). This is a graphical method in which the individual's location on the chart is determined by the values of their risk factors and the different coloured regions correspond to the different risk levels.

One article developed a multiple-stage algorithm for estimating risk of breast cancer at five years for women having a mammogram.[118] Multiple stages were required to obtain firstly age and race specific incidence rates from an external database (SEER), then adjust for those dying from other causes over time (using US vital statistics), and combining with the HR estimates from a proportional hazard model before then making further adjustments. Though the algorithm is complex, the important observation for this thesis is that external data about incidence over time for different types of women is being used to translate the estimated model hazard ratios to absolute risk predictions for new women.



How to use the chart

- Choose the section with the sex of the patient, and their diabetes and smoking status
- Find the cell that matches the patient's risk factor profile using the age, body-mass index (BMI), and blood pressure
- Refer those with excessive blood pressure (>180 mm Hg) to a physician

Figure 3.8: Risk prediction charts for men. Reprinted from the Lancet, 371, Gaziano et al.[131] Laboratory-based versus non-laboratory-based method for assessment of cardiovascular disease risk: the NHANES I Follow-up Study cohort, 923-31, Copyright 2008, with permission from Elsevier.

All of the above methods for predicting risk in new patients were for predicting risk up to a specified time point. Two articles displayed Kaplan-Meier plots for the risk groups created from the score system.[117,128] These have been discussed in the section on 'Presenting absolute risks' as they compared Kaplan-Meier curves for risk groups but these could also be used with the risk score system to give an estimate of survival over time (see Figure 3.3 for example). However, such predictions are the average for the risk group rather than for the individual conditional on their covariate values.

3.3.7 Validation

Some form of validation was performed in 26 of the 31 articles (83.9%). Twenty-five of these 26 articles (96.2%) used internal validation (using the same data as model development),[113-115,117-119,123-126,128-142] two articles (7.7%) performed temporal validation (using more recent data for validation),[125,140] and seven articles (26.9%) performed external validation (using a different dataset to the development data).[112-114,117,125,128,142] Models could be validated using more than one type of validation in an article, for example, internally and externally validated.

Discrimination and calibration performance of the model

Discrimination measures were reported in 22 articles (84.6% of the 26 articles that included validation). This included the C-statistic or area-under-the-curve (AUC), reported in all 22 articles,[115,117-119,123-126,128-132,134-142] D statistics reported in five articles,[119,123,134,138,139] and ROC curves displayed in seven articles.[117,126,129-131,135,141]

Calibration was reported in 17 articles (65.4% of the 26 articles that included validation) and included the Hosmer-Lemeshow test in seven articles,[131,132,134-137,140] comparison of observed and predicted risk (graphically or tabulated) in eight articles,[115,117,119,123,138-140,142] expected/observed ratio in two articles,[118,138] and the Gronnesby and Borgan calibration test in one article.[133]

Other statistics for model checking

Other measures included the Brier score in three articles,[119,136,139] R^2 in five articles,[119,123,134,138,139] likelihood ratio tests in three articles,[133-135] and AIC and BIC in one article.[134] Another article also reported Entropy and Yates slope as global measures of model fit.[136] Two articles reported the *k* statistic for concordance (in comparing models),[126,134] eight articles (30.8%) reported reclassification improvement,[118,132,133,135-139] and four articles

(15.4%) reported integrated discrimination improvement.[132,133,136,137] The three genetic articles validated gene-signatures by using log-rank tests for the difference in survival between high and low risk groups as defined by the gene-signatures.[112-114]

Consideration of absolute risk in validation

Eleven articles considered validation of model performance in terms of absolute risk, mostly when checking calibration of the model (discussed above). Absolute risks were compared between the development and validation datasets by comparing observed and expected risk for deciles of predicted risk according to the prediction model at a specified time point (reported in five articles),[118,119,123,138,139] or by plotting Kaplan-Meier survival curves for risk groups in each dataset (reported in five articles, see Figure 3.3 for example).[112-114,117,128] Another article plotted Kaplan-Meier and predicted survival curves for risk groups in the validation dataset.[115] The survival curves were predicted using the fitted Cox model, but no detail was given on how the predicted curves were derived.

Comparison of absolute risk

In six articles, absolute risk was compared across different models (but not different datasets). Four of these articles compared absolute risk between different models using risk reclassification tables,[134-137] one article compared models using observed and expected risk,[134] another using Kaplan-Meier survival risk for risk groups to compare models,[126] and one other compared cumulative probability at specified time points for different models.[141]

3.4 Discussion

3.4.1 Main findings

This chapter has presented a review of the methods and reporting of multivariable prediction models in leading medical journals. The first of the primary aims was to identify if and how researchers are modelling the baseline hazard in the development of multivariable prediction models. The review found that 100% of the 31 articles used Cox proportional hazards modelling to develop a prediction model using time-to-event data. This finding is slightly higher than another review by Mallett et al. on prognostic models in cancer, which found that 94% of studies fitted Cox models.[84] Two other reviews (not focused specifically on modelling survival data) also found that the Cox model was the most commonly used approach in the studies that used survival data to develop the model.[144,145] The key feature of Cox modelling is that the baseline hazard function is not explicitly modelled; therefore none of the articles included in the review in this chapter, modelled the baseline hazard over time within their multivariable model, which meant that other approaches were required to estimate absolute risk from the fitted multivariable model. Altman also notes that the baseline hazard function is required for absolute survival estimates and that none of the articles in the review he conducted on models for breast cancer, modelled the baseline hazard.[11]

The most common method of presenting absolute risk was to produce Kaplan-Meier survival curves for risk groups, defined by categorising risk scores from the model into groups and then plotting the Kaplan-Meier curve for each group separately. The review by Altman found that 54% of studies produced a graph (Kaplan-Meier survival curves) to show the expected survival for risk groups.[11] Authors also produced tables or plots of risk scores and corresponding absolute risk estimates at a specified time point.

Absolute risk is important when making predictions for individuals and the authors most commonly used the point scoring system with corresponding risk estimates to make absolute risk predictions for individuals at a specified time point, for example 10-year risk. Other methods for predicting risk in individuals included online risk calculators, nomograms and colour-coded risk prediction charts, however all of these were for a single estimate of risk up to a specified time point. Two articles used Kaplan-Meier estimated survival/risk for risk groups and this was the only method for predicting a profile of risk over a whole time period from the prediction model. However, this is also limited as it only gives average estimates for groups of individuals rather than individual predictions. This highlights the limitation of not explicitly modelling the baseline hazard function in the Cox model. If the baseline hazard is modelled, as demonstrated in Chapter 2, it allows predicted risk for an individual over a whole time range, rather than being limited to a single predicted risk by a particular time point or an average predicted risk over time for categorised risk groups.[34] Modelling the baseline hazard therefore offers greater flexibility in how the model could be presented and used to enhance predictions for patients.

In general, even when baseline survival was reported at a particular time-point, there was very little information on how it was estimated. For example, Ridker et al. and Schnabel et al. reported baseline survival at a particular time point but did not state either in the article or appendices how this value was obtained.[136,140] Particularly, when new prediction models are developed for risk up to a particular time point (and are not the original survival models), it should be clear how the baseline survival or hazard has been calculated.

3.4.2 Other important findings

Aside from absolute risk, this review also looked at other statistical issues. Missing data was an issue amongst the articles included in the review, reported in 84% of articles. However, complete case analysis was the most common approach and few used any method of imputation. This leads to a reduction in sample size of the development dataset leading to imprecise model estimates and biased estimates if missingness is associated with the outcome.[12] This is an area in which the methods for imputation are available and can be implemented in many software packages, yet are often ignored for a simpler approach.[22] Five articles (16%) used multiple imputation, either for their primary analysis or as a sensitivity analysis. This is slightly higher than in a previous review of clinical prediction articles which found that only 8% of included articles used multiple imputation.[85] but comparable to a review of models for chronic kidney disease which found that 18% used multiple imputation, however the number of studies that this relates to was small.[145]

Variable selection methods remain a debatable topic with data-driven methods such as automatic selection strategies often being criticised.[22,146,147] In this review, 32% of articles used automatic selection procedures. Three articles used univariable results to select variables for inclusion in multivariable model selection. This approach in particular has been heavily criticised as it may result in potentially important variables being excluded.[44,83] Variables were pre-specified for a multivariable model in 36% of the studies, however it is not always possible to have prior information on variables that are clinically important and data-driven methods are sometimes necessary, but should be used with care. Also, at least 29% of articles had fewer than 10 events per variable (EPV) considered. This is lower than the 50% found by Bouwmeester et al. in a review of clinical prediction articles,[85] but similar to a review of risk prediction models for type II diabetes which found that 21% of studies had less than 10 EPV and that EPV could not be calculated in 33% of articles.[144] When using data-driven variable selection methods, it is important to ensure that there are enough events when developing a model. Studies into EPV have warned that a low ratio of EPV can result in poor accuracy and precision in regression coefficient estimates and their significance tests, leading to important predictors being missed and unimportant factors being deemed predictors.[43,148]

Hazard ratios with confidence intervals and p-values were the most common results reported for both univariable and multivariable analyses. Interpretation of results using hazard ratios means that it is done on the relative scale. As highlighted in Chapter 2, it is important to know what the hazard is relative to. If the baseline hazard is small, a large hazard ratio may not mean a large absolute difference in survival and likewise when the baseline hazard is large, a hazard ratio closer to the value of one does not necessarily mean a small difference in absolute survival probabilities. Particularly in prediction and prognostic modelling, it is important to model this baseline hazard and report absolute survival probabilities in addition to hazard ratios. Even if the study is only interested in identifying important variables, consideration of how they change absolute risk is informative.

The Hippisley-Cox et al. articles included in the review all used a large database with data from many practices.[119,123,138,139] So if the baseline hazard is modelled, they would need to consider if it is consistent from practice to practice. Currently they implicitly assume it is consistent as practice was not accounted for, however this may not be true. Authors would face a similar issue when developing a risk prediction model from multiple studies.[69,149] This issue is considered in detail in Chapter 5.

3.4.3 Limitations of the review

Extractions in this review are dependent on reporting standards and some information was difficult to extract. For example, it was often not stated whether the proportional hazards assumption was checked or whether non-linear functions were tested for continuous variables. This might indicate that it was not considered, but of course it may have been done but not reported. Several other reviews have also commented on poor reporting on many key aspects of model development and/or validation of prediction models,[11,54,83,84,144,145] which has led to the recent publication of the TRIPOD (transparent reporting of multivariable models for prediction or diagnosis) statement which provides guidelines for reporting of multivariable prediction models.[42,59]

The review in this chapter used a database of articles from a previous review.[3] Therefore, it was dependent on the previous search strategy. This was not necessarily a limitation in that the previous review was more inclusive than necessary for this review, it was also updating other (previous) reviews and used a published search strategy for prognostic models.[150] However, the review was limited to general medical journals which the authors acknowledge do not include the majority of prognostic models.

Another limitation of the review is that developing a model intended for absolute risk prediction was not always the main goal for authors and absolute risk was not discussed in all of the articles. For this reason, a reduced number of 25 articles were considered in the review when looking at reporting absolute risk. Also, these articles may not consider it necessary to validate the model. A further limitation is that in using an existing database of prediction models, the review is limited to clinical articles between 2006 and 2009. There is the possibility that since 2009 flexible parametric modelling or other methods in which the baseline hazard is modelled may have increased in popularity.

3.5 Recommendations

Based on the findings of the review, a number of core recommendations for improving the development of multivariable prediction models are made in Box 3.1. Mallett et al. already provide suggested improvements for reporting,[83] and TRIPOD has recently published their guidelines.[42,59]

Box 3.1: Recommendations for improvements in the development of multivariable prediction models using time-to-event data, based on findings of the review.

- Move away from Cox modelling and rather consider other approaches (such as flexible parametric models) that explicitly modelling the baseline hazard in order to:
  - Make absolute risk predictions for individuals directly from the model.
  - Alleviate the need to create risk groups and thus lose information.
  - Examine the impact of each predictor on absolute risk, not just relative risk.
  - Allow risk predictions *over* time rather than limiting predictions to one particular time.
  - Check model performance in terms of calibration of observed and predicted absolute risks, in the whole population and also subgroups (e.g. across different populations or practices with variation in case-mix)
- Check if the proportional hazards assumption holds or if time-dependent effects are required.
- Model continuous variables on their continuous scale, with potential consideration of non-linear functions.
- Compare complete case analysis with a multiple imputation analysis.
- Pay attention to sample size constraints (events per variable > 10).
- Do not base inclusion of a predictor on univariable results. Consider clinical judgement and previous studies to select predictors where possible.

3.6 Conclusion and next steps

The most important finding of this literature review of the development of prediction models is that Cox modelling is the common approach, and so the baseline hazard is not being explicitly modelled. This creates problems when trying to use the model to make absolute risk predictions, and typically forces researchers to create risk groups (and thus lose the ability to make individual-level predictions) and focus on predictions by one time-point, rather than over time. However, if parametric or flexible parametric models were fitted instead of the Cox model, the baseline hazard function would be explicitly modelled over time, and therefore individual risk profiles could be predicted at any time point.[32,34] To demonstrate this, the following chapter uses flexible parametric methods to develop a prognostic model for advanced pancreatic cancer using clinical trials data.

CHAPTER 4: DEVELOPING A PROGNOSTIC MODEL UTILISING THE BASELINE HAZARD AND DATA FROM CLINICAL TRIALS: AN EXAMPLE IN PANCREATIC CANCER

4.1 Introduction and objectives

Chapter 2 identified several statistical advantages of using Royston-Parmar models. One of the major benefits was explicitly modelling the baseline hazard function which can then be used for predictions. This is particularly important in the context of prognostic modelling, where a risk prediction model is needed to predict outcome risk in diseased patients. Despite this, the literature review in Chapter 3 highlighted that the baseline hazard function is not usually modelled by researchers when developing prediction models from survival data. Models were mostly developed using a Cox proportional hazards model and then risk groups were created by categorising patients based on their risk score (linear predictor) alone, thus ignoring the baseline hazard. Kaplan-Meier survival curves could then be estimated for each risk group, to obtain estimated survival probabilities over time. However this forces individuals to be in one of the risk groups, thereby collapsing the individual risk to be similar to others in the group. In contrast, Royston-Parmar models explicitly model the baseline hazard. This allows for smooth survival functions to be estimated for individuals as well as for risk groups and so Royston-Parmar models seem potentially very pertinent for prognostic model research.

To investigate this and demonstrate the advantages and limitations, this chapter explores the use of Royston-Parmar modelling of the baseline hazard in the development of prognostic models. The data considered come from advanced pancreatic cancer trials, and the use of trials data bring some particular issues and challenges, which will also be highlighted and explored.

The primary aims of this chapter are:

- To show the added benefit of modelling the baseline hazard in prognostic models, especially in terms of making individualised risk predictions.
- To compare the use of the baseline hazard when making predictions to the current typical method of using risk groups and ignoring the baseline hazard.
- To identify and illustrate issues in modelling the baseline hazard using randomised trial data for the purposes of prediction, specifically when there are multiple treatment groups and a time-dependent treatment effect.

4.2 Summary of available pancreatic cancer trials data

Data from clinical trials are increasingly being used to develop and validate risk prediction models. For example, the IMPACT database, which included data from eight randomised controlled trials (as well as three observational studies), was used to develop a prognostic model for outcome after traumatic brain injury.[151] Clinical trials provide a prospective cohort of patients, with a rich set of prognostic factors measured at baseline, good quality follow-up and complete follow-up for most patients. In this chapter, a prognostic model will be developed using data from two international phase III trials sponsored by British Biotech.

4.2.1 Details of the two trials

The primary aim of each trial was to compare the effect of treatments in patients with advanced pancreatic cancer where death was the primary end point. Pancreatic cancer was the 10th most commonly diagnosed cancer in the UK in 2010 and prognosis for patients with pancreatic cancer is poor with only around 4% surviving five or more years.[152] The inclusion criteria for patient selection were similar for both trials, including patients aged 18 or over with a histological or cytological diagnosis of non-resectable pancreatic cancer and Karnofsky performance status >50% in trial BB128 or >60% in BB193. Trial BB193 also specified ranges of values of laboratory parameters. Although the patients recruited to each trial were similar, the treatments differed between the trials. The first trial (BB128) included a total of 414 patients randomised to gemcitabine or one of three doses of marimastat, 5, 10 and 25mg. The second trial (BB193) randomised 239 patients to receive gemcitabine + placebo or gemcitabine + marimastat 10mg. The number and proportion of patients in each of the treatment groups along with the number of events and number of patients censored are given in Table 4.1. The median and maximum follow-up durations were 21.1 months (95% CI: 20.2 to 22.8) and 25.8 months respectively in BB128 and 20.3 months (95% CI: 18.9 to 21.3) and 23.3 months in BB193. Median follow-up was calculated using the reverse Kaplan-Meier method.[153]

Table 4.1: Number and proportion of events and censoring by treatment group and total number of patients receiving each treatment in both trials.

Treatment	N Events (% of treatment group)	N Censored (% of treatment group)	N Total (% of trial)
Trial BB128			
Gemcitabine (1000mg/m^2)	97 (94.2)	6 (5.8)	103 (24.9)
Marimastat (5mg)	98 (94.2)	6 (5.8)	104 (25.1)
Marimastat (10mg)	101 (96.2)	4 (3.8)	105 (25.4)
Marimastat (25mg)	96 (94.1)	6 (5.9)	102 (24.6)
Trial BB193			
Gemcitabine (1000mg/m^2) + placebo	107 (89.9)	12 (10.1)	119 (49.8)
Gemcitabine (1000mg/m^2) + marimastat (10mg)	113 (94.2)	7 (5.8)	120 (50.2)

Kaplan-Meier survival plots for each trial are shown in Figure 4.1. For trial BB128, as reported by Bramhall et al., there was no statistically significant difference in overall survival between marimastat 25mg and gemcitabine (log-rank, p=0.78); overall survival was lower in marimastat 10mg than gemcitabine but not statistically significant when using the Bonferroni adjustment for multiple testing (HR=0.76, 95% CI: 0.57 to 1.01, p=0.05), and survival was lower in marimastat 5mg compared to gemcitabine but again not statistically significant (HR=0.82, 95% CI: 0.62 to 1.09, p=0.16).[154] For trial BB193, Bramhall et al. reported no significant difference in overall survival between the two treatment arms, gemcitabine + placebo and gemcitabine + marimastat (HR=0.99, 95% CI: 0.76 to 1.30 , p=0.95).[155]

In summary, there is some evidence in the BB128 trial (though not statistically significant) of a possible difference between gemcitabine and marimastat. However in the BB193 trial, marimastat does not seem to add any value over gemcitabine when compared to gemcitabine with a placebo. Therefore in terms of prognosis, it seems more important to consider whether patients received gemcitabine or not.

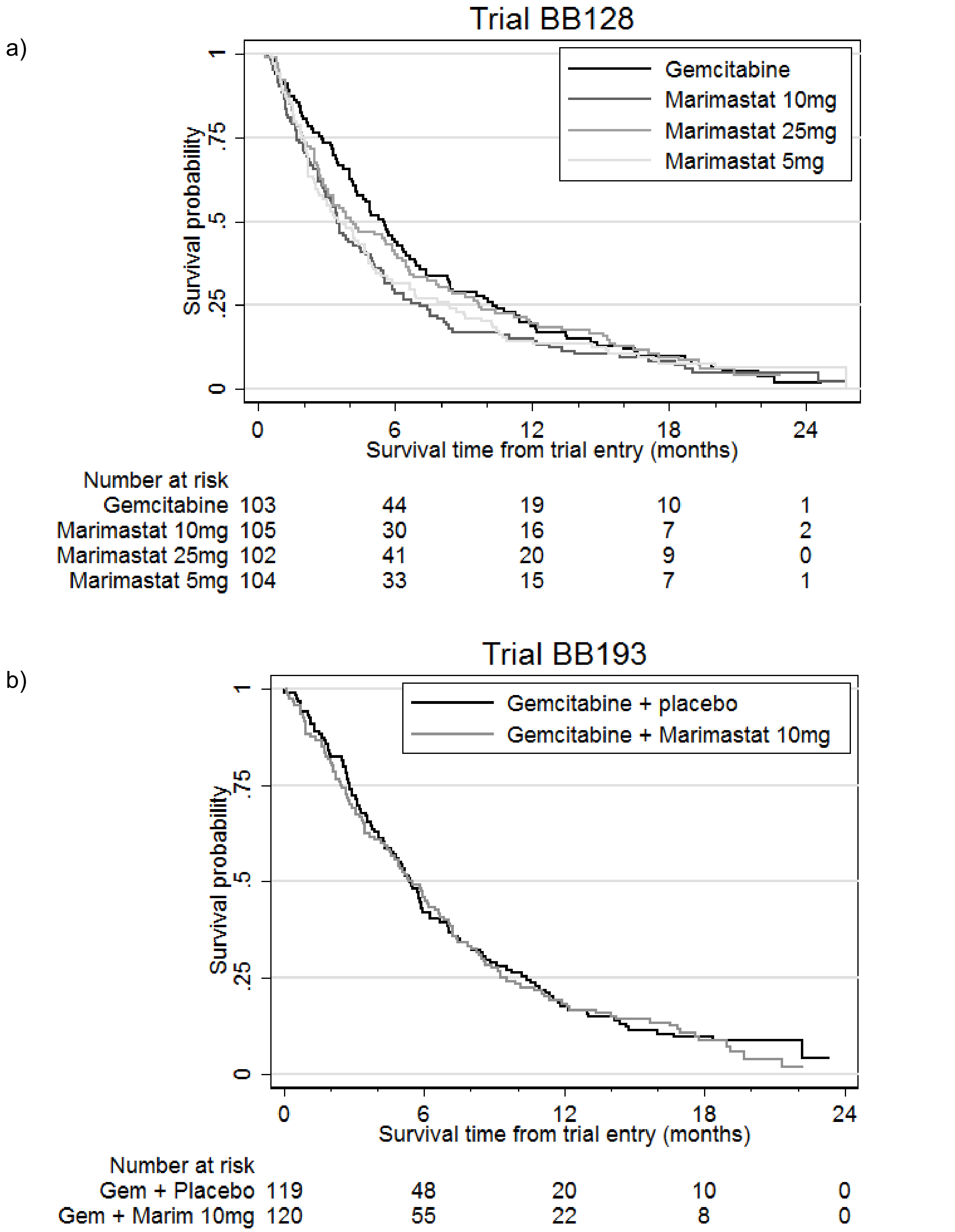


Figure 4.1: Kaplan-Meier survival plots for treatment in trial (a) BB128 and (b) BB193.

4.2.2 Comparison of gemcitabine and no gemcitabine patients

Information recorded for patients at baseline included demographics such as age, sex and race; haematology and serum chemistry variables such as haemoglobin, white blood cell count (WBC), aspartate aminotransferase (AST), bilirubin, alkaline phosphatase, albumin, lactate dehydrogenase (LDH), blood urea nitrogen (BUN) and CA19-9; and variables relating to the cancer such as stage (I-IV), tumour size and whether it had extended into nearby tissue (T0-T4), whether regional lymph node metastasis was present (N0, N1) and whether the cancer had metastasised (M0, M1). The summary statistics for these variables pooled across the two trials are given in Table 4.2 below. The pooling of the two trials and different treatment groups is discussed further in Section 4.4.1. The mean age and proportions for sex and race are similar across treatment groups. The median values for WBC and alkaline phosphatase are slightly higher in the gemcitabine treatment group compared to the no gemcitabine group with 8.1 g/dL versus 7.6 g/dL for WBC and 155IU/L versus 132.5 IU/L for alkaline phosphatase. Patients were followed up over time until death occurred or patients were censored and had a median survival time of 5.5 months in the gemcitabine group and 3.7 months in the no gemcitabine group.

There is a rule of thumb based on simulation studies that there should be at least 10 EPV considered for inclusion when developing a survival model.[43,148] This is to ensure that the sample size is large enough to avoid problems with precision and over-fitting, especially when using automatic selection procedures. Across the two trials, a total of 16 variables were considered and a very high event rate in pancreatic cancer meant that there were 612 events. Using this rule of thumb, the EPV is 38, which is sufficient for developing a prognostic model.

Table 4.2: Summary statistics for baseline characteristics by treatment group categorised as receiving gemcitabine or not receiving gemcitabine, with patients pooled across the two trials.

	No gemcitabine (n=311)	Gemcitabine (n=342)
Demographics	**Mean (SD)**	**Mean (SD)**
Age at entry (years)	62.62 (9.96)	61.62 (10.20)
	n (%)	**n (%)**
Sex		
Male	172 (55.31)	196 (57.31)
Female	139 (44.69)	146 (42.69)
Race		
White	272 (87.46)	318 (92.98)
Black	24 (7.72)	11 (3.22)
Oriental	6 (1.93)	1 (0.29)
Other	9 (2.89)	11 (3.22)
Missing	0 (0.00)	1 (0.29)
Cancer Information	**n (%)**	**n (%)**
Stage		
I	15 (4.82)	17 (4.97)
II	34 (10.93)	38 (11.11)
III	58 (18.65)	46 (13.45)
IV	200 (64.31)	239 (69.88)
Missing	4 (1.29)	2 (0.58)
Tumour		
T0	2 (0.64)	6 (1.75)
T1	82 (26.37)	76 (22.22)
T2	71 (22.83)	73 (21.35)
T3	126 (40.51)	154 (45.03)
T4	6 (1.93)	1 (0.29)
Missing	24 (7.72)	32 (9.36)
Nodes		
N0	115 (36.98)	128 (37.43)
N1	124 (39.87)	127 (37.13)
Missing	72 (23.15)	87 (25.44)
Distant Metastasis		
M0	97 (31.19)	97 (28.36)
M1	200 (64.31)	236 (69.01)
Missing	14 (4.50)	9 (2.63)

Table 4.2 continued…

	No gemcitabine (n=311)	Gemcitabine (n=342)
Laboratory Tests	**Median (IQR)**	**Median (IQR)**
Haemoglobin, g/dL **	12.40 (1.58)	12.37 (1.51)
WBC, 10^9/L	7.55 (6 – 9.55)	8.1 (6.5 – 10.45)
AST, IU/L	24 (17 – 40)	26 (18 – 40)
Bilirubin, μmol	13.68 (10.26 – 21.38)	13.68 (10.26 – 22.23)
Alkaline Phosphatase, IU/L	132.5 (91 – 236)	155 (97 – 260)
Albumin, g/L **	38.12 (4.16)	37.98 (4.35)
LDH, IU/L	161 (134 – 201)	168.5 (135.5 – 218)
BUN, mmol/L	9.29 (7.86 – 11.43)	8.57 (7.14 – 11.43)
CA19-9, KU/l	654 (84.5 – 4850)	800 (93 – 4500)
Follow-up		
Endpoint, n (%)		
Dead	295 (94.86)	317 (92.69)
Alive	16 (5.14)	25 (7.31)
Median follow-up length (months)	20.72	20.69
Median survival (months)	3.72	5.46

** Mean (SD) reported for normally distributed variables instead of median (IQR).

4.2.3 Missing data

All patients had a survival time recorded along with whether this was the time to death or censoring, therefore all patients could potentially be included in a survival analysis. Information was complete for age, sex and treatment. However, there was missing data for the other variables (Table 4.3). Information on nodes was missing for almost a quarter of patients, whereas other variables were missing in up to 9% of patients.

Table 4.3: Number of observations missing by variable.

Variable	Number (%) missing (N=653)
Age	0 (0.00)
Sex	0 (0.00)
Race	1 (0.15)
Treatment	0 (0.00)
Stage	6 (0.92)
Tumour	56 (8.58)
Nodes	159 (24.35)
Metastasis	23 (3.52)
Haemoglobin	41 (6.28)
WBC	41 (6.28)
AST	29 (4.44)
Bilirubin	24 (3.68)
Alkaline phosphatase	24 (3.68)
Albumin	25 (3.83)
LDH	32 (4.90)
BUN	33 (5.05)
CA19-9	47 (7.20)

4.3 Previous modelling using the advanced pancreatic trial data

The data from these trials has previously been used by Stocken et al. to develop a prognostic model for advanced pancreatic cancer.[156] The authors combined the two trials to form one dataset with the inclusion of a trial variable in the model. They also decided to create and use a new treatment variable based on whether or not patients received the experimental drug marimastat or not. A multivariable Cox proportional hazards model was used to develop the prognostic model for pancreatic cancer.[156,157] Variables were selected using the strategy proposed by Collett, which in summary first tests variables in a univariable analysis and then includes significant variables in a backward elimination process.[24] The variables that were not significant in the univariable analysis were then tested by adding them to the multivariable model one at a time. Stocken et al. developed the model using a complete case analysis but later explored the use of multiple imputation. The authors also included second degree fractional polynomial terms for CA19-9 due to non-linearity.

The Stocken model was developed using Cox proportional hazards regression and therefore did not explicitly model the baseline hazard. Stocken et al. used the final model to categorise patients into risk groups based on the risk score from the model. They then produced Kaplan-Meier survival curves displaying the average survival function for each of the risk groups (see Appendix Figure C1.1). This method of using Kaplan-Meier curves to report survival probabilities for patients after fitting a Cox model was used in several of the articles included in the literature review (see Chapter 3).

The published paper by Stocken et al. has been cited 20 times according to Web of Science (search date: 1 May 2015).

4.4 Prognostic model development

From here onwards, the data are now used to develop a new prognostic model using the Royston-Parmar modelling approach, and to thereby go beyond the Cox modelling approach of Stocken et al. to improve individualised predictions. The details of model development are now described, and along the way key issues are numbered and emphasised in bold.

4.4.1 Combining trials and treatment groups

Recall that in Section 4.2.2, patients were placed into two treatment groups rather than the six original groups from the two trials. The treatments were simplified into patients that received gemcitabine (includes gemcitabine in BB128, gemcitabine + placebo and gemcitabine + marimastat in BB193) and patients that did not receive gemcitabine (includes marimastat at 5, 10 and 25mg in BB128) **(CLINICAL TRIALS ISSUE 1: Dealing with multiple treatment groups in the development of a prognostic model)**. The decision to combine treatment groups into those receiving or not receiving gemcitabine was based on the published results of the two trials (see Section 4.2) which gave some suggestion (though not formally statistically significant) that patients receiving gemcitabine had a better survival in BB128 when compared to marimastat at any dose. This difference is more noticeable in the early months (Figure 4.1a). Furthermore, there was no noticeable difference (neither statistical nor visual) between gemcitabine + placebo and gemcitabine + marimastat in BB193, suggesting that marimastat did not improve the survival of patients already receiving gemcitabine.[154,155] The hazard functions for each treatment group were also estimated (using 3 d.f.) and compared. The shapes of the hazard functions for treatment groups in which patients received gemcitabine were similar. The hazard functions of treatment arms that did not include gemcitabine were also similar in shape to each other. This supports the decision to combine treatment groups based on receiving gemcitabine or not (Figure 4.2). Clinically, for the model to be relevant, it was also important to predict outcome for those receiving gemcitabine or not. It is also worth noting that combining treatment groups breaks the randomisation, but this is only a problem for prediction if the effect of the original trial treatment groups actually differs (otherwise predictions are the same for each group, and therefore including multiple groups is unnecessary).

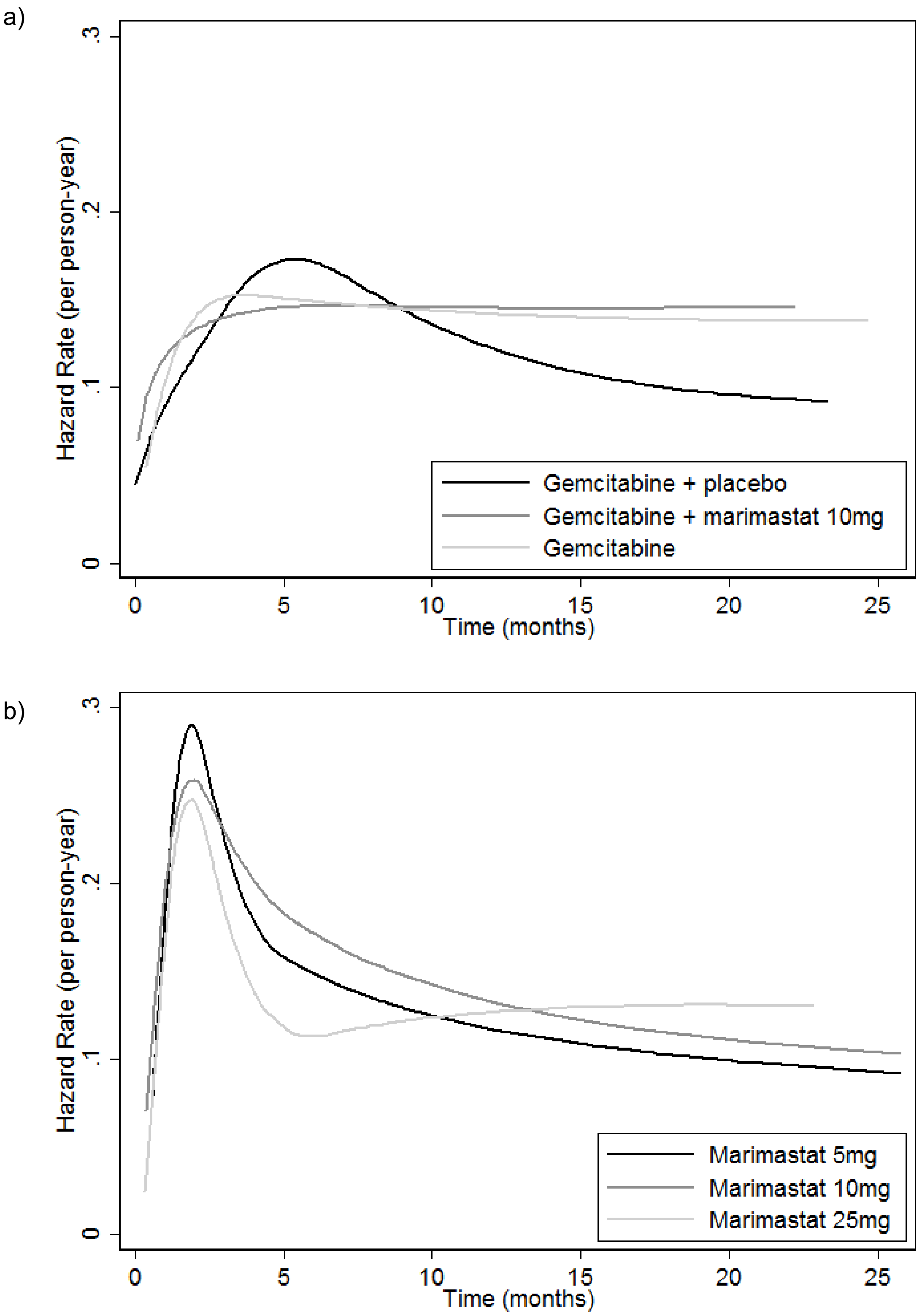


Figure 4.2: Estimated hazard functions for individual treatment groups, (a) patients receiving gemcitabine and (b) patients not receiving gemcitabine.

To begin with, data from the two clinical trials were combined to form a single dataset for model development as was done by Stocken et al. **(CLINICAL TRIALS ISSUE 2: Handling multiple trials in the development of a prognostic model)**. As discussed in Section 4.2, the patient populations are similar for both trials and therefore the baseline hazard should be similar. Using the combined data to develop a prognostic model also gives more power to detect genuine prognostic factors and develop a more clinically useful model. For the same reason, it was also decided not to include a trial variable in the model and thus force both trials to have the same baseline hazard. This is because a trial variable would not be meaningful if the model was used to predict outcome in future patients that did not belong to either of these trials. This assumption will however also be looked at later in this chapter (see Section 4.6.2). Note that to allow a separate baseline hazard for each trial, a random effect could be fitted on the intercept (baseline hazard), however, this cannot be reliably estimated when there are only two trials.

4.4.2 Dealing with missing data

To handle the missing data, multiple imputation (MI) was used **(CLINICAL TRIALS ISSUE 3: Dealing with missing data)**. Assuming data are missing at random, i.e. that the probability of a variable being missing depends on observed data rather than unobserved data, MI can be used to give unbiased regression estimates and standard errors and is generally considered more efficient than complete-case analysis.[158] When several variables have missing values, multiple imputation by chained equations (MICE) uses a set of imputation models to impute values for each variable in turn using the other variables in a regression model. This is restricted to individuals that have an observed value for the variable being imputed, and then used to predict values for individuals that are missing values for that variable. The missing values for the next variable are then imputed, this time including the already imputed values for the first variable in the imputation model. The missing values are imputed for each variable in turn until all missing values have been imputed. This is said to be a cycle. Several cycles are run to stabilize the results and produce one imputed dataset. The whole procedure is repeated multiple times resulting in several imputed datasets. Each imputed dataset is analysed separately and then estimates are combined across the datasets to give overall estimates using Rubin's rules.[45,46] The total variance for an estimate includes both the within-imputation variance and the between-imputation variance, therefore reflecting the uncertainty due to the missing data as well as the uncertainty in the parameter estimate itself.

The variable 'nodes' was not recorded for 24% of patients and therefore it was decided to exclude this variable from model development. If this missingness reflects the information often not routinely recorded for pancreatic cancer patients, including nodes in the prognostic model may make the model unusable in clinical practice. Harrell supports data reduction including eliminating variables that are missing in a large number of patients and likely to be missing in future patients as this also makes the model more likely to validate well in future data.[22]

Statisticians used to think that only a small number of imputed datasets (*M*) were required for statistical efficiency (*M*=3-5 commonly used).[45] However, as more research has been done in the area, suggestions have been made to use a larger *M* to reduce the loss of power.[159] Further studies have looked at the number of imputed datasets required for reproducibility of the results if the analysis was repeated. This has led to a rule of thumb, that the number of imputed datasets should be selected to be at least equal to the percentage of incomplete cases.[46] In the advanced pancreatic cancer data, 24% of patients are missing values for at least one of the variables considered in the model development, therefore *M*=25 was selected for analyses with multiple imputation.

The variables recorded in both trials were the same; therefore there was no problem in combining the datasets and imputing missing values. However, if a variable was only recorded in one trial and not the other, that variable could not be included in the analysis. Imputing values for a variable in a whole study would mean using the observed values from the other study, which may not be sensible as the distribution of values for that variable could be different in each study, though methodology work is ongoing to address this.[160]

Summary statistics are shown for the data available for each variable as well as for a complete case analysis (*n*=496) and using multiply imputed data (Table 4.4). The summary statistics are very similar in all three settings, suggesting no obvious bias due to the missingness of these variables in the complete case for any of the summary statistics. However, the estimates from the multiply imputed data are closer than the complete case estimates when compared to all available data. There is no difference in the median survival or follow-up between all available data and the imputed data because the outcome was complete for all individuals.

Table 4.4: Summary statistics for all available data, complete case and multiply imputed data.

	All available data	Complete case (*n*=496)	Multiply imputed data (*n*=653)
	Mean (SD)	**Mean (SD)**	**Mean (SD)**
Age at entry (years)	62.09 (10.09)	61.90 (10.02)	62.09 (10.09)
	Percentage	**Percentage**	**Percentage**
Sex			
Male	56.36	55.65	56.36
Female	43.64	44.35	43.64
Race			
White	90.49	89.31	90.46
Black	5.37	6.45	5.37
Oriental	1.07	1.21	1.08
Other	3.07	3.02	3.09
Stage			
I	4.95	5.44	4.98
II	11.13	10.89	11.16
III	16.07	16.53	16.09
IV	67.85	67.14	67.77
Follow-up			
Endpoint			
Dead (%)	93.72	93.55	93.72
Alive (%)	6.28	6.45	6.28
Median follow-up (months)	20.69	21.12	20.69
Median survival (months)	4.70	4.84	4.70

All available data:	Uses all available data for each variable.
Complete case:	Uses observations that are complete for all variables considered for model development.
Imputed data:	Uses all observations and averages across multiply imputed datasets.

4.4.3 Stratification factors

Several articles have been published recommending that randomisation factors included in the design of a trial should also be included in the analysis of data.[161-164] Not accounting for the randomisation factors in the analysis of trials data can lead to wider confidence intervals and larger p-values when testing the treatment effect.[163] Within each of the two trials, patients were assigned to treatment arms based on minimization as the randomisation method, using the following factors: pancreatic cancer staging, Karnofsky score, recurrent/newly diagnosed disease, gender and study centre. These factors were used for randomisation as they were considered to be prognostic. Although information on most of these factors was not available in the dataset provided for prognostic model development, stage and sex were available and so they were forced into any models fitted. **(CLINICAL TRIALS ISSUE 4: Dealing with trial stratification factors in the modelling)**.

4.4.4 Proportional hazards assumption

There were 17 candidate predictors available in the dataset. The assumption of proportional hazards was checked for each variable considered for inclusion in the multivariable model using plots of –ln(-ln(survival probability)) against log time (referred to as 'log-log' plots) and a test based on Schoenfeld residuals for non-proportional hazards in univariable Cox models. To produce log-log plots, continuous variables had to be categorised and this was done by dichotomising haematology and serum chemistry variables into normal and abnormal values as defined by clinical investigators. Age was dichotomised using the mean value of 63 years. When the assumption of proportional hazards holds, the lines in the log-log plot should appear approximately parallel.

A statistical test can be performed to check if hazards are not proportional by checking if there is a deviation from zero slope of scaled Schoenfeld residuals over time in a generalised linear regression.[27,29] The p-values of the proportional hazards test (Table 4.5) suggest that the assumption might not hold for the variables albumin, LDH, BUN, stage, tumour, metastasis and treatment (all p<0.05). Therefore this needed to be considered during model development.

Table 4.5: P-values for proportional hazards test of factors within a univariable model based on scaled Schoenfeld residuals.

Variable	Proportional hazards test p-value
Age	0.286
Sex	0.401
Race	0.723
Haemoglobin	0.609
WBC	0.142
AST	0.767
Bilirubin	0.580
Alkaline phosphatase	0.348
Albumin	<0.001
LDH	0.021
BUN	0.020
CA19-9	0.645
Treatment with gemcitabine	0.011
Stage	0.001
Tumour	0.014
Nodes	0.819
Metastases	<0.001

Although there were a few variables with significant p-values for the statistical test, most of the log-log plots were not too concerning. However, the log-log plot for treatment of gemcitabine or no gemcitabine (Figure 4.3) showed crossing lines. This suggests that a time-dependent effect may be required to model treatment correctly. For this reason, it was decided to start by fitting separate prognostic models to each treatment group and gradually work towards a more complex model involving both treatments. The full set of log-log plots can be found in Appendix Figure C2.1 to Figure C2.3.

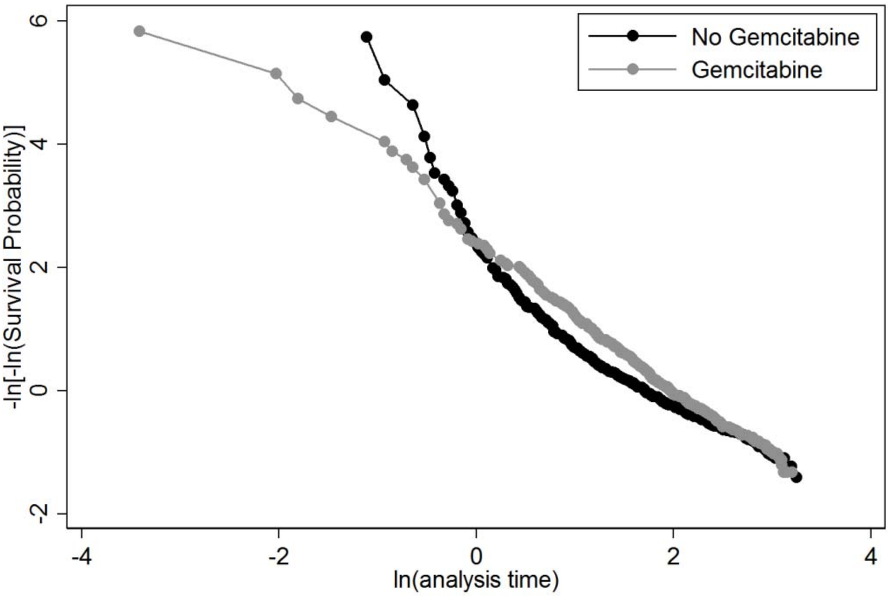


Figure 4.3: Log-log plot for treatment and death as the outcome.

4.5 Development and internal validation of treatment specific prognostic models

4.5.1 Development

Choosing d.f. for baseline hazard

For each treatment specific prognostic model, the two trials were combined for the purpose of model development and the trial variable was ignored as discussed earlier. To examine model fit, the baseline hazard functions from null Royston-Parmar models (model with no variables included) using between 1 and 4 d.f. were plotted and compared to the shape of the kernel smoothed function.[27] Using the hazard function plots (Figure 4.4) and the AIC and BIC as a guide, 2 or 3 d.f. seemed sufficient for each treatment group because it captured the same trend as larger d.f. Based on this, Royston-Parmar models with 3 d.f. for the baseline hazard function were fitted in the gemcitabine and no gemcitabine treatment groups.

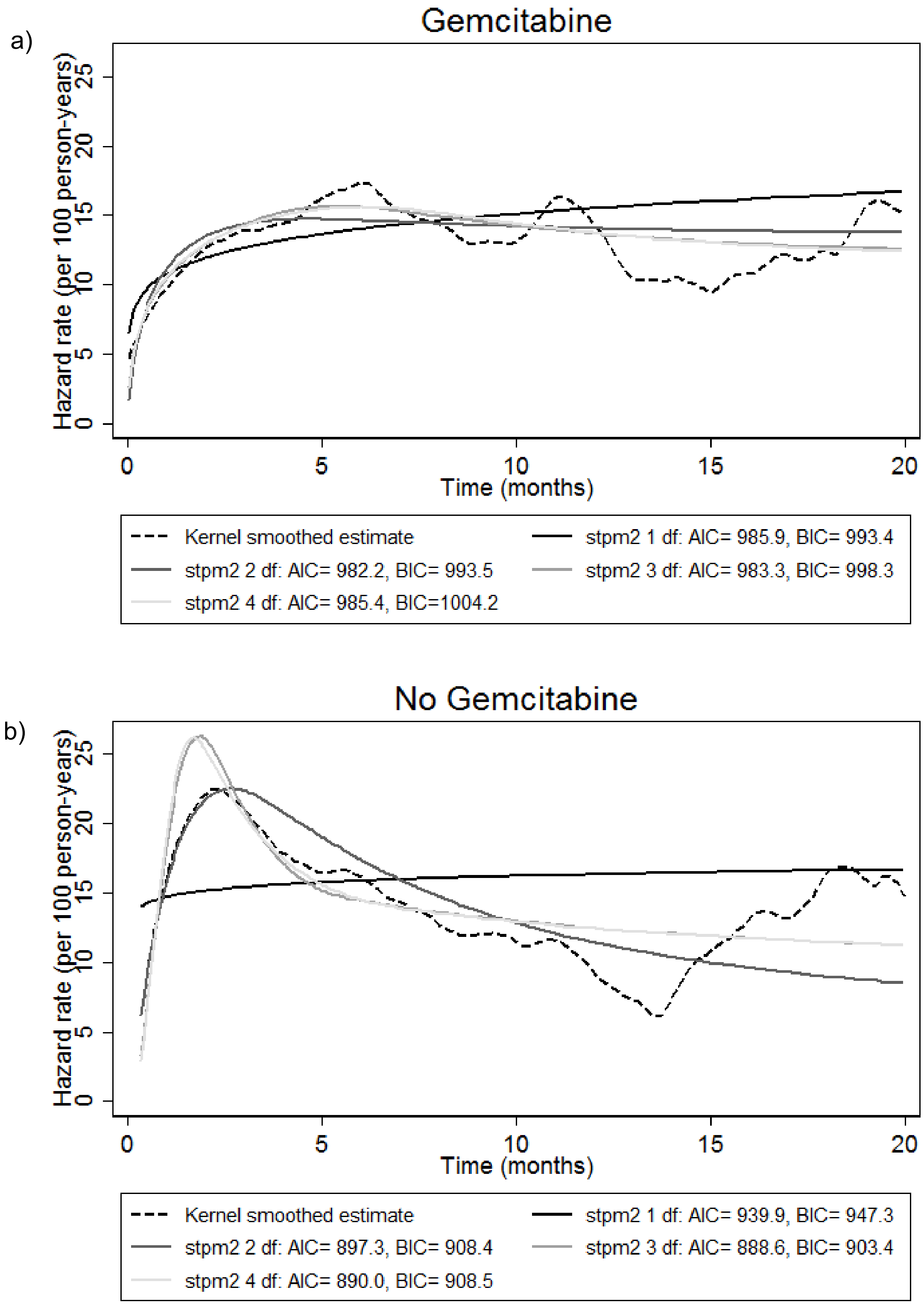


Figure 4.4: Hazard functions estimated by kernel smoothing and spline functions with between 1 and 4 degrees of freedom for (a) gemcitabine and (b) no gemcitabine treatment groups.

Modelling stage

Stage originally had four categories (stage I, II, III and IV) which were simplified into two categories (stage I/II versus III/IV). This decision was based on how Stocken et al. modelled stage in the previously published prognostic model,[156] and also because the dataset consists of patients with advanced stage pancreatic cancer, so there are small numbers in the stage I and II categories (approximately 5% and 11% of patients were stage I and II respectively), compared to stages III and IV.

Variable selection

The method used for selecting variables for inclusion in the multivariable model was an automatic selection procedure which considers non-linear functions for continuous variables. The variables were included in each treatment specific multivariable model if they were selected using the multivariable fractional polynomials for multiply imputed data (MFPMI) command in Stata using significance level α=0.157 for inclusion in the model. Backward elimination with α=0.157 is used as a proxy for all-subset model selection based on AIC.[165] Vergouwe et al. also used this method for model selection with imputed data.[166] In summary, multivariable fractional polynomials (MFP) does the following:[49,50]

- Determines the fitting order, by performing likelihood ratio tests for each variable comparing the full model (model including all variables included for model selection, fitted linearly for continuous variables) with a model excluding that variable. The fitting order is determined by ordering the variables based on the likelihood ratio test p-values from most to least significant.
- Using the fitting order and a specified significance level (e.g. α=0.157), each variable is considered in turn. For categorical variables, a likelihood ratio test is performed to test whether the variable should be dropped from the model. For continuous

variables, the deviance of a model using the 'best' fractional polynomial of second degree (FP2) for the variable under consideration is compared to the deviance of a model excluding that variable (χ_4^2 test). If this test was significant, the FP2 model is compared to a model in which the variable is modelled linearly (χ_3^2 test). If this test was significant, the FP2 model is compared to an FP1 model (χ_2^2 test).

- The procedure cycles through the variables in the same order until there is no change in the variables selected or the functional form of each variable.

Instead of testing for possible complex FP2 functions, MFP was restricted to FP1 for non-linear functions for simplicity and to reduce the number of tests performed. Rather than the first hypothesis test comparing the model with the FP2 term for a variable versus the model without the variable (null) for variable inclusion or exclusion, the first hypothesis is now FP1 versus null (χ_2^2 test). If the variable is not excluded based on this first test, a second test is then performed comparing FP1 to the linear form of the variable.

Rather than a single dataset being used for model development, there are multiple datasets here as a result of the multiple imputation strategy outlined in Section 4.4.2. A study comparing different methods for variable selection strategies using imputed data showed that stacking the imputed datasets and using weights according to the fraction of missing for each variable seemed a reasonable approach to use in this case.[167] The fraction of missing differed considerably from variable to variable (between 0.00 and 0.32) so rather than using an averaged fraction of missing in the weighting, the variable specific fraction of missing was used when considering inclusion of that variable. The weighting given to each variable in the stacked dataset is shown in (4.1) below,

$$w_i=(1-f_i)/M \tag{4.1}$$

where w_i is the weighting used for each observation in the stacked dataset when considering the inclusion or exclusion of variable X_i, f_i is the fraction of missing for variable X_i and M is the number of imputed datasets. MFPMI allows the MFP procedure to be implemented for model selection while using the stacking method with variable specific weightings described above.

Following variable selection using MFPMI, the selected variables (including any non-linear functions) were refitted using the MI package which uses Rubin’s rules to combine the estimates across datasets to obtain the final model estimates. A summary of the modelling strategy discussed above is given in Box 4.1.

Box 4.1: Steps taken in developing treatment-specific models for advanced pancreatic cancer.

Step 1: **Combine treatment groups**

Both trials were combined to form one dataset and the original trial treatment arms were combined to form two new treatment groups: gemcitabine and no gemcitabine. Models were then developed for each of these treatment groups separately as the new treatment groups did not have proportional hazards (see step 3).

Step 2: **Impute missing data**

Assuming data are missing at random, multiple imputation was used to generate 25 imputed datasets.

Step 3: **Check proportional hazards assumption**

'Log-log' plots and a test of scaled Schoenfeld residuals over time were used to assess whether the proportional hazards assumption was reasonable for each variable.

Step 4: **Decide d.f. for baseline hazard**

Null Royston-Parmar models were fitted for each treatment group using between 1 and 4 d.f. The decision was based on plots of hazard functions, as well as AIC and BIC statistics.

Step 5: **Select variables for inclusion in final models**

Imputed datasets were stacked and variable selection was performed using MFPMI. This uses backward elimination with α=0.157 and considers 1st degree fractional polynomial functions for continuous variables.

Due to multiply imputed data being used, a weighting is applied to each observation when considering inclusion/exclusion in the model. This weighting is based on the fraction of missing information for that variable and the number of imputed datasets.

Trial stratification factors were included in the model regardless of significance as they were already considered prognostic for advanced pancreatic cancer.

Step 6: **Obtain model estimates using Rubin's rules for multiply imputed data**

Refit variables selected in previous step including any non-linear functions using MI package to obtain final model estimates.

4.5.2 Results of model development

The models developed for each treatment group are given in Table 4.6. Twelve predictors were included in either one or both of the models. Sex and stage were forced into the models as they were stratification factors in the trials and clinically considered prognostic. However, neither stage nor sex were significant predictors in the treatment specific models. First degree fractional polynomial terms were selected for AST (in the gemcitabine model) and CA19-9, whereas the other continuous variables were modelled linearly. Alkaline phosphatase, albumin, LDH, ln(CA19-9) and metastasis were included in both models.

Interestingly, the variables age and WBC were included in the gemcitabine model but were not retained in the no gemcitabine model. Conversely, haemoglobin and BUN were retained in the no gemcitabine model but not the gemcitabine model. AST was modelled as an FP1 in the gemcitabine model but linearly in the no gemcitabine model **(CLINICAL TRIALS ISSUE 5: Using selection procedures to identify prognostic factors in a model)**.

Table 4.6: Hazard ratio estimates for variables selected using MFP in gemcitabine and no gemcitabine treatment groups modelled separately.

Variable	Gemcitabine		No gemcitabine	
	HR (95% CI)	P-value	HR (95% CI)	P-value
Age	1.019 (1.007 to 1.032)	0.002	-	-
Male	1	-	1	-
Female	0.824 (0.649 to 1.045)	0.110	0.842 (0.656 to 1.080)	0.176
Haemoglobin	-	-	1.058 (0.972 to 1.153)	0.194
WBC	1.065 (1.031 to 1.099)	<0.001	-	-
AST	-	-	0.996 (0.991 to 1.000)	0.044
$(AST/100)^{-0.5}$	1.686 (1.307 to 2.174)	<0.001	-	-
Alkaline phosphatase	1.002 (1.001 to 1.002)	<0.001	1.001 (1.001 to 1.002)	0.001
Albumin	0.951 (0.924 to 0.980)	0.001	0.902 (0.870 to 0.935)	<0.001
LDH	1.001 (1.0005 to 1.002)	0.002	1.004 (1.003 to 1.005)	<0.001
BUN	-	-	1.032 (0.998 to 1.067)	0.065
Ln(CA19-9)	1.084 (1.032 to 1.138)	0.001	1.169 (1.112 to 1.229)	<0.001
Stage I/II	1	-	1	-
Stage III/IV	0.930 (0.613 to 1.411)	0.734	1.014 (0.673 to 1.529)	0.945
No Metastasis	1	-	1	-
Metastasis	1.359 (0.968 to 1.907)	0.076	1.441 (1.053 to 1.972)	0.022

Interestingly, when the two models were re-fitted with the same set of variables forced to be included, the hazard ratios appeared similar across models for most of the variables (Table 4.7), suggesting the data may be combinable for the development of a single prognostic model (see Section 4.6).

Table 4.7: Comparable multivariable model hazard ratio estimates for gemcitabine and no gemcitabine treatment groups modelled separately.

Variable	Gemcitabine		No gemcitabine	
	HR (95% CI)	P-value	HR (95% CI)	P-value
Age	1.019 (1.006 to 1.032)	0.003	0.997 (0.985 to 1.010)	0.695
Male	1.000	-	1.000	-
Female	0.838 (0.648 to 1.083)	0.176	0.803 (0.613 to 1.053)	0.113
Haemoglobin	0.978 (0.888 to 1.077)	0.647	1.048 (0.959 to 1.145)	0.299
WBC	1.064 (1.031 to 1.098)	<0.001	1.032 (0.989 o 1.077)	0.146
$(AST/100)^{-0.5}$	1.653 (1.277 to 2.139)	<0.001	1.337 (1.033 to 1.732)	0.027
Alkaline phosphatase	1.002 (1.001 to 1.002)	<0.001	1.001 (1.001 to 1.002)	0.001
Albumin	0.955 (0.925 to 0.985)	0.004	0.908 (0.873 to 0.943)	<0.001
LDH	1.001 (1.0004 to 1.002)	0.002	1.004 (1.003 to 1.005)	<0.001
BUN	1.017 (0.982 to 1.054)	0.351	1.029 (0.993 to 1.065)	0.111
Ln(CA19-9)	1.086 (1.035 to 1.141)	0.001	1.167 (1.110 to 1.228)	<0.001
Stage I/II	1.000	-	1.000	-
Stage III/IV	0.911 (0.599 to 1.386)	0.664	1.009 (0.667 to 1.526)	0.966
No Distant Metastasis	1.000	-	1.000	-
Distant Metastasis	1.374 (0.977 to 1.931)	0.068	1.430 (1.043 to 1.961)	0.026

4.5.3 Internal validation using the same data

When a prognostic model is developed, it is important to validate its performance, first internally and then, if possible, externally. See Chapter 1 for an introduction to validation methods and statistics (Section 1.6). So here, validation of the two developed models is now considered.

Methods

Section 4.5.2 developed two prognostic models, one for each treatment group. For internal validation of each model, patients were categorised into four risk groups, based on quartiles of the risk score (using the linear predictor from the model, $\boldsymbol{x\beta}$), so risk group 1 had the best survival and risk group 4 had the worst survival. The calibration of each treatment specific survival model was internally validated by plotting the observed survival curves (Kaplan-Meier) and the mean predicted survival curves from the model for each treatment group. In order to do this, the risk score was calculated for each patient using the treatment specific model. Harrell's C-statistic for discrimination[168,169] was also calculated to assess model fit. This was done by using the model obtained using multiple imputation methods (using Rubin's rules for the parameter estimates) and looking at validation in each of the imputed datasets. Risk groups were derived using quartiles of risk scores in all of the imputed datasets combined to determine the cut-points, which were used within each of the 25 imputed datasets for comparability.

Results

The internal validation plots show that the treatment specific models (reported in Table 4.7) predict well on average for the risk groups within each treatment group. The calibration plots were created for each of the imputed datasets, Figure 4.5 shows the calibration between expected and observed survival probabilities over time within one of the imputed dataset as an example. There were slight differences between plots from the different imputed datasets due to the different values imputed for predictors, but Figure 4.5 is representative of what was generally seen across the imputed datasets. The expected survival functions lie close to the observed survival functions but there are slight deviations from this. Overall, the gemcitabine model fits marginally better than the no gemcitabine model. The no gemcitabine model slightly underestimates survival in the first seven months and then overestimates survival thereafter for patients in risk group 1 with the lowest risk of death.

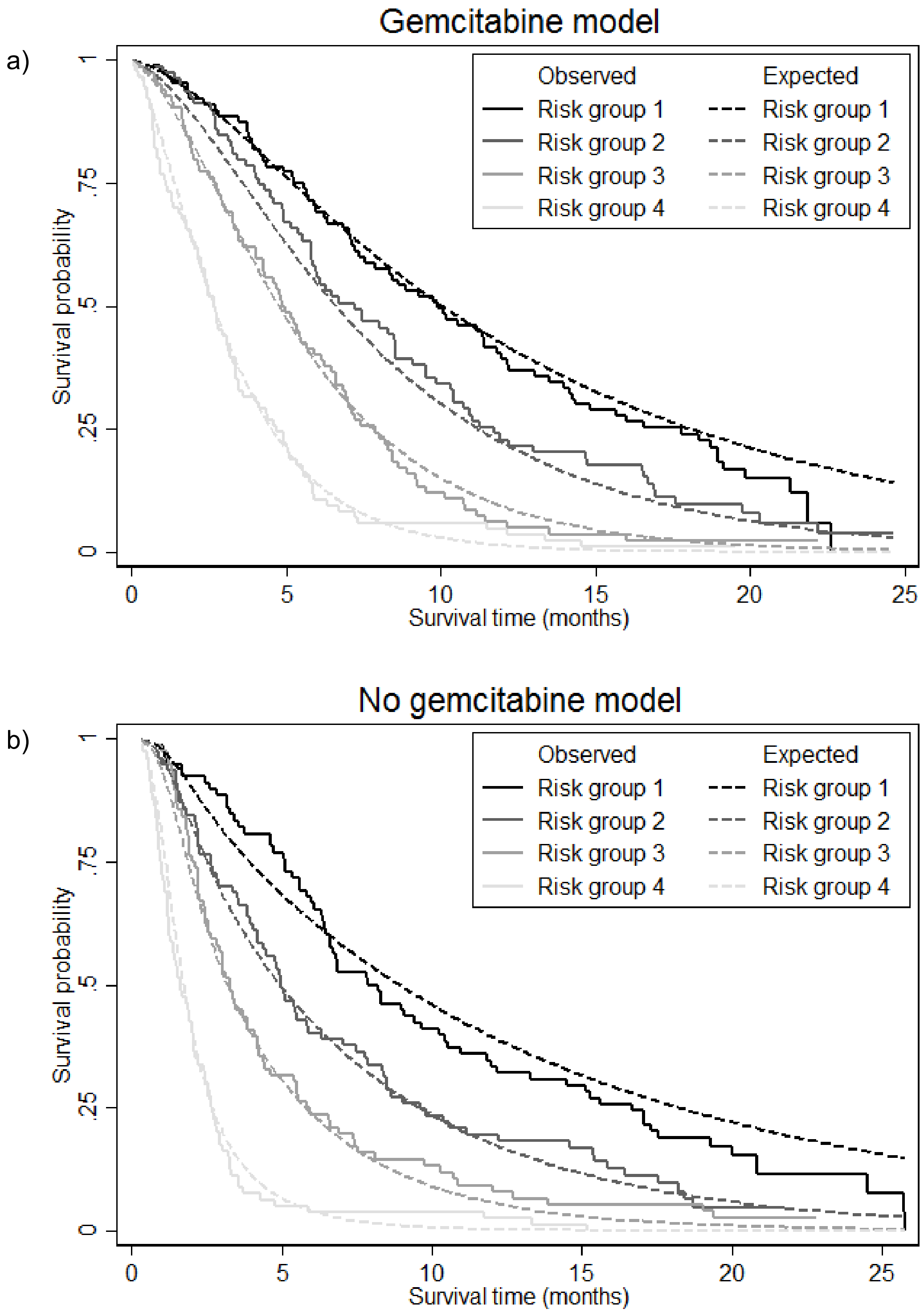


Figure 4.5: Kaplan-Meier (observed) and mean predicted (expected) survival for internal validation of each treatment specific model (a) gemcitabine and (b) no gemcitabine in one of the imputed datasets.

Validation statistics are reported for both treatment specific models in Table 4.8. The value of Harrell's C-statistic can range between 0.5 and 1.0, with values close to 1.0 suggesting good discrimination. The no gemcitabine model had better discrimination when validated within the same data used for development (average C-statistic=0.721) compared to the gemcitabine model (average C-statistic=0.688).

Table 4.8: Harrell's C-statistic for treatment specific models.

	Gemcitabine	No gemcitabine
Average C-statistic (across imputed datasets)	0.688	0.721
Range (across imputed datasets)	0.681 to 0.694	0.716 to 0.726

4.5.4 Internal-external validation using other treatment group

As two datasets were formed (one for modelling in each treatment group), an interesting option is to validate each model using the other dataset not used for its development. This is a sort of internal-external validation approach; internal as the same trial data are used, but external because a new group of patients on a different treatment are used for the validation **(CLINICAL TRIALS ISSUE 6: Use of trial data for model validation)**. Each model was developed in one treatment group and the cut-points for risk groups calculated as quartiles of risk score in the development data. The other treatment group was then categorised into risk groups for which mean predicted survival curves were created using the model and compared to the observed survival functions in the validation data (Figure 4.6).

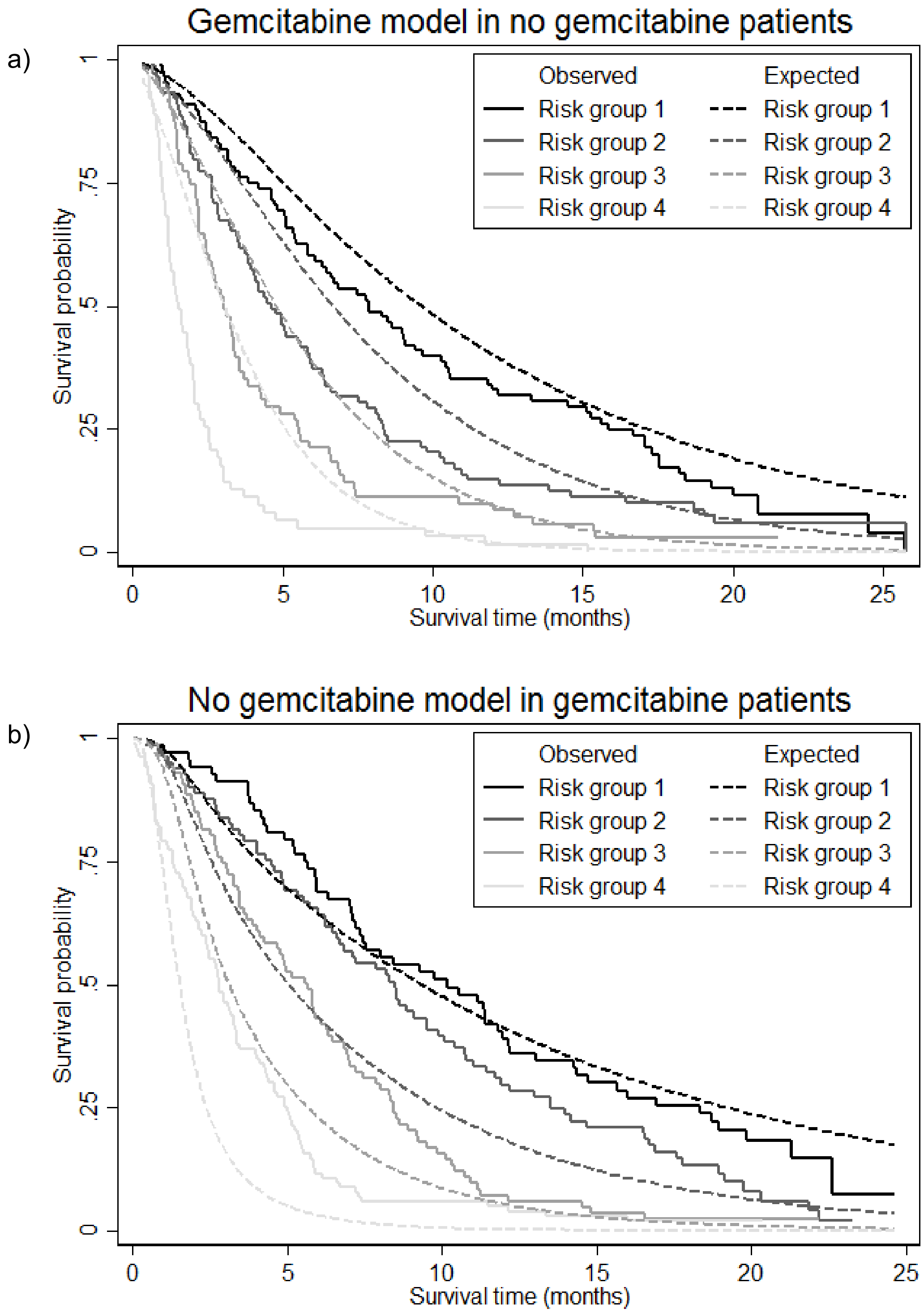


Figure 4.6: Kaplan-Meier (observed) and mean predicted (expected) survival using treatment specific models in the other treatment group, (a) gemcitabine model in no gemcitabine patients and (b) no gemcitabine model in gemcitabine patients. One imputed dataset used for illustration.

Discrimination

There still appears to be good discrimination between risk groups when the model is fitted in the other treatment group, as they remain well separated (seen in Figure 4.6). This suggests that the linear predictor (i.e. prognostic factor effects, HRs) is similar in both gemcitabine and no gemcitabine treatment groups. The similar discrimination is expected, as Table 4.7 showed that the HRs were very similar in both treatment specific models. However, the discrimination is slightly less than expected in the no gemcitabine model, and slightly more than expected in the gemcitabine group.

The internal validation (Section 4.5.3) showed that the models discriminated reasonably well with average C-statistics across imputed datasets of 0.688 and 0.721 for the gemcitabine and no gemcitabine models respectively. When each model was applied to the other treatment group in an internal-external validation, the average C-statistics were 0.691 and 0.680 for the gemcitabine and no gemcitabine models. These values suggest slightly better discrimination of the gemcitabine model in the other treatment group, whereas the no gemcitabine model performed slightly worse in the other treatment group.

Calibration

Figure 4.6 shows visually that calibration is poor. In particular, the baseline hazard is higher in the no gemcitabine treatment group and neither model validates well in the other treatment group. The gemcitabine model *over-predicts* survival probabilities over time in all risk groups. The prediction for a risk group can even exceed the observed survival curve of the next risk group. For illustration, Table 4.9 shows the observed and expected survival probabilities at six months in the imputed dataset used for Figure 4.6. The expected survival probability for risk group 3 is 0.384 which even exceeds the observed survival for risk group 2 of 0.371. In contrast, the no gemcitabine model *under-predicts* survival probabilities as would be expected, given that the gemcitabine model over-predicts survival.

Table 4.9: Observed and expected six month survival probabilities of treatment specific models tested in the other treatment group (in one imputed dataset).

Model	Risk group	Six month survival probability	
		Observed	Expected
Gemcitabine	1	0.602	0.691
	2	0.371	0.550
	3	0.211	0.384
	4	0.048	0.180
No gemcitabine	1	0.689	0.644
	2	0.654	0.434
	3	0.430	0.230
	4	0.116	0.031

4.5.5 Conclusion about the two developed models

Validation of the treatment specific models showed that they calibrated well within the same data used for development of the model, but did not discriminate particularly well in either treatment group (average C-statistic=0.688 and 0.721 for gemcitabine and no gemcitabine models respectively). Validation of each treatment specific model in the other treatment group (internal-external validation) highlighted quite consistent discrimination, but a different baseline risk of gemcitabine and no gemcitabine patients. This caused the gemcitabine model to over-predict survival in the no gemcitabine patients and the no gemcitabine model to under-predict survival in gemcitabine patients. Figure 4.6 showed that survival probabilities vary more between risk groups (greater discrimination) in the no gemcitabine treatment group than in the gemcitabine treatment group. The next section moves on to developing a single prognostic model for both treatment groups rather than having separate models.

4.6 Development and internal validation of a single prognostic model including treatment

The previous sections in this chapter show how treatment specific models can be developed and validated. However, the aim might be to develop one prognostic model that can be used for both treatment groups with the same set of predictors. In this situation, both the treatment groups should contribute to the same model development. This is complicated here as graphical checks indicated that the proportional hazards assumption does not hold for treatment (see Figure 4.3) and so separate baseline hazards may be necessary. To begin with, for simplicity a proportional hazards model was fitted before considering more complex models.

4.6.1 Assuming proportional hazards

The aim was to assess how well the model predicts and how much the failed proportional hazards assumption affects the predictions. Following variable selection using MFPMI and refitting the model using Rubin's rules, the hazard ratio estimates for the proportional hazards model with 3 d.f. for the baseline hazard are reported in Table 4.10. Age was not retained in the model which is a consequence of it being only significant in the gemcitabine model and not in the no gemcitabine model. Also, haemoglobin was not retained in the model as it was not significant (at even the 10% level) in either of the treatment specific models. Bilirubin was not retained in either of the treatment specific models but is significant ($p=0.019$) in the proportional hazards model with both treatment groups included and is modelled using a fractional polynomial term (power of -1). LDH, which was previously modelled linearly, is included in this model using the natural log function and is highly significant ($p<0.001$). The hazard ratios for the other variables are reasonably similar to the hazard ratios from the treatment specific models.

Table 4.10: Hazard ratio estimates from model assuming proportional hazards for treatment.

Variable	Hazard ratio (95% CI)	P-value
Male	1	-
Female	0.827 (0.699 to 0.979)	0.028
WBC	1.043 (1.018 to 1.069)	0.001
$(\text{AST}/100)^{-0.5}$	1.545 (1.274 to 1.873)	<0.001
$(\text{Bilirubin}/100)^{-1}$	0.973 (0.951 to 0.996)	0.019
Alkaline phosphatase	1.001 (1.0006 to 1.002)	<0.001
Albumin	0.934 (0.916 to 0.952)	<0.001
Ln(LDH)	1.851 (1.439 to 2.381)	<0.001
BUN	1.027 (1.003 to 1.051)	0.028
Ln(CA19-9)	1.111 (1.073 to 1.150)	<0.001
No gemcitabine	1	-
Gemcitabine	0.640 (0.541 to 0.756)	<0.001
Stage I/II	1.000	-
Stage III/IV	0.953 (0.715 to 1.272)	0.744
No distant metastasis	1	-
Metastasis	1.432 (1.136 to 1.807)	0.002

The C-statistic for the proportional hazards model is 0.705 when averaged across the imputed datasets with a range of 0.702 to 0.708.

4.6.2 Inclusion of a time-dependent effect for treatment

As discussed previously, there is evidence to suggest that the proportional hazards assumption does not hold for treatment (see Figure 4.3). The Royston-Parmar model assuming proportional hazards was given in equation (1.13). Writing the model out again, now using the spline variables created in Stata (*rcs_1* to *rcs_n*), the model presented in Section 4.6.1 can be written as

$$\ln\big(H_i(t)\big) = \gamma_0 + \gamma_1 rcs\_1 + \ldots + \gamma_n rcs\_n + \boldsymbol{\beta x_i} \qquad \textbf{(4.2)}$$

where γ_0 is the constant, γ_1 to γ_n are the coefficients for the spline variables, *n* is equal to the d.f. used for the restricted cubic spline function and $\boldsymbol{\beta x_i}$ is the linear combination of regression coefficients and covariate values (β_1**sex(female)*+ β_2**WBC+...*) for patient *i*. This model can be extended to include a time-dependent effect for treatment by including an interaction between treatment and the baseline hazard function by adding spline terms for treatment.

$$\ln\big(H_i(t)\big) = \gamma_0 + \gamma_1 rcs_1 + \ldots + \gamma_n rcs_n + \delta_1 rcs\_trt\_1 + \ldots + \delta_k rcs\_trt\_k + \boldsymbol{\beta x_i} \qquad \textbf{(4.3)}$$

where δ_1 to δ_k are the additional parameter estimates along with *rcs_trt_*1 to *rcs_trt_k* which are the additional spline variables created when *k* is the d.f. used for the time-dependent effect. When the time-dependent variable is binary as in the case of treatment, this effectively fits one baseline hazard function for patients not on gemcitabine and allows a different underlying hazard function for patients on gemcitabine. Royston-Parmar models were fitted using between 2 and 4 d.f. for the time-dependent effect within each imputed dataset, and the AIC and BIC for these models and the proportional hazards model were compared within datasets. In all of the 25 imputed datasets, the lowest AIC and BIC was observed for the model using 3 d.f. for the time-dependent effect of treatment (Table 4.11). There is also a clear improvement in AIC and BIC for the model using 2 d.f. for the time-dependent effect compared to the proportional hazards model reported in Section 4.6.1, suggesting improved model fit by including a time-dependent effect for treatment. Therefore, 3 d.f. seemed adequate to model the time-dependent effect for treatment.

Table 4.11: AIC and BIC for comparison of proportional hazards model (PH) and models with time-dependent effect for treatment (TD) with between 2 and 4 d.f.

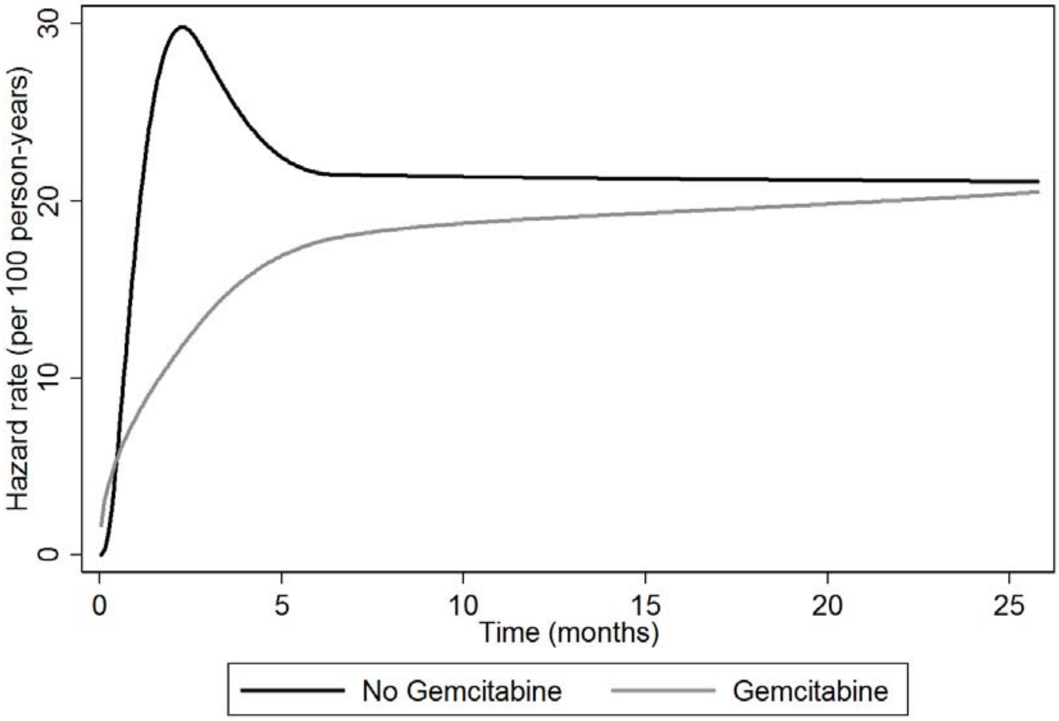
Imputed dataset	AIC				BIC			
	PH	TD 2 d.f.	TD 3 d.f.	TD 4 d.f.	PH	TD 2 d.f.	TD 3 d.f.	TD 4 d.f.
1	1678.9	1665.8	1660.8	1662.8	1750.6	1746.5	1745.9	1752.4
2	1669.8	1656.0	1649.7	1651.7	1741.5	1736.7	1734.9	1741.3
3	1673.9	1661.1	1655.4	1657.2	1745.6	1741.8	1740.5	1746.9
4	1677.6	1665.1	1659.1	1661.1	1749.3	1745.7	1744.3	1750.7
5	1665.5	1652.8	1647.4	1649.4	1737.3	1733.5	1732.6	1739.0
6	1678.6	1666.7	1660.2	1662.1	1750.3	1747.4	1745.4	1751.8
7	1662.8	1648.9	1643.2	1645.2	1734.5	1729.6	1728.3	1734.8
8	1678.9	1666.3	1660.8	1662.7	1750.6	1747.0	1745.9	1752.4
9	1665.1	1651.3	1645.9	1647.9	1736.8	1731.9	1731.0	1737.5
10	1672.7	1659.8	1653.8	1655.7	1744.4	1740.4	1738.9	1745.3
11	1676.0	1663.3	1657.3	1659.3	1747.7	1743.9	1742.5	1748.9
12	1671.1	1658.0	1652.0	1654.0	1742.8	1738.7	1737.2	1743.6
13	1668.2	1655.2	1649.2	1651.1	1739.9	1735.9	1734.3	1740.7
14	1672.1	1660.0	1654.7	1656.8	1743.8	1740.7	1739.9	1746.4
15	1674.7	1661.9	1656.4	1658.3	1746.4	1742.5	1741.5	1747.9
16	1674.6	1662.5	1656.9	1658.8	1746.3	1743.2	1742.0	1748.5
17	1666.1	1652.3	1646.8	1648.7	1737.8	1732.9	1731.9	1738.3
18	1680.5	1668.2	1661.6	1663.5	1752.2	1748.8	1746.7	1753.1
19	1678.8	1666.3	1660.9	1662.9	1750.5	1746.9	1746.0	1752.5
20	1677.0	1664.8	1658.2	1660.2	1748.7	1745.5	1743.4	1749.8
21	1670.0	1657.4	1651.6	1653.5	1741.7	1738.1	1736.7	1743.2
22	1675.2	1662.4	1656.4	1658.3	1746.9	1743.0	1741.5	1747.9
23	1677.8	1664.6	1658.4	1660.2	1749.5	1745.2	1743.6	1749.9
24	1672.2	1659.6	1653.7	1655.7	1743.9	1740.2	1738.8	1745.3
25	1674.0	1661.4	1655.5	1657.4	1745.7	1742.1	1740.7	1747.1

A likelihood ratio test comparing the time-dependent model (3 d.f.) to the model assuming proportional hazards for treatment (Section 4.6.1) gives a highly significant p<0.0001 in all of the imputed datasets, further evidence suggesting that a time-dependent effect is required to model the treatment effect. The hazard ratio estimates for the variables other than treatment in the time-dependent model are reported in Table 4.12 and are similar to the hazard ratio estimates from the proportional hazards model (Section 4.6.1).

Table 4.12: Hazard ratio estimates from the model including a time-dependent effect for treatment.

Variable	Hazard ratio (95% CI)	P-value
Male	1	-
Female	0.831 (0.702 to 0.983)	0.031
WBC	1.045 (1.020 to 1.071)	<0.001
$(AST/100)^{-0.5}$	1.556 (1.285 to 1.884)	<0.001
$(Bilirubin/100)^{-1}$	0.974 (0.952 to 0.997)	0.027
Alkaline phosphatase	1.001 (1.0007 to 1.002)	<0.001
Albumin	0.932 (0.914 to 0.950)	<0.001
Ln(LDH)	1.821 (1.414 to 2.346)	<0.001
BUN	1.027 (1.003 to 1.052)	0.026
Ln(CA19-9)	1.108 (1.070 to 1.148)	<0.001
Stage I/II	1	-
Stage III/IV	0.955 (0.716 to 1.273)	0.752
No distant metastasis	1	-
Metastasis	1.439 (1.141 to 1.815)	0.002

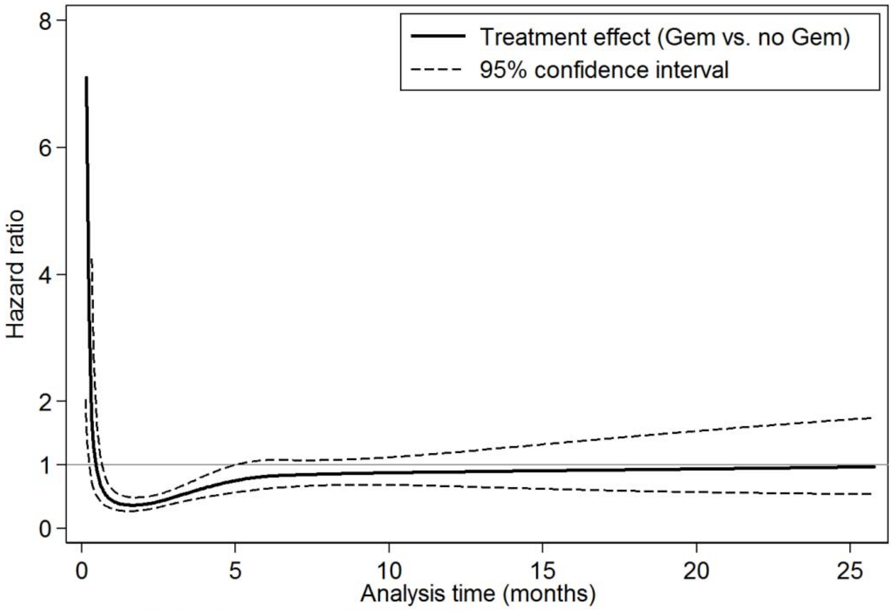
The time-dependent effect included for treatment means that the treatment effect is no longer represented by a hazard ratio estimate alone. This can therefore not be displayed in the table of hazard ratio estimates. It is easier to visualise by plotting the hazard functions of the two treatment groups to understand how they have been modelled as in Figure 4.7. The figure shows the baseline hazard function on either gemcitabine or no gemcitabine treatment, assuming mean values of covariates in the model, and male sex with stage III/IV cancer. This suggests that there is a difference between the hazard functions in the first six months but very little difference after this time.



Assumes male, stage III/IV with metastasis and mean values of other covariates

Figure 4.7: Baseline hazard functions modelled in time-dependent model.

Another way to visualise the time-dependent effect is to plot the hazard ratio for treatment over time. Figure 4.8 again shows that after around six months there is negligible difference between the treatments with a HR≈1. The hazard ratio for the first few days suggests a higher hazard of death for patients on gemcitabine compared to no gemcitabine with HR>1 and wide confidence intervals. However, Figure 4.7 showed that the hazard of death in this early period was very low for both treatment groups and therefore a relative risk may be large but the absolute risk difference is small (see section 2.6.5, pages 75 to 79 for a previous illustration of this). After 15 days, the hazard ratio reduces to less than one in favour of patients on gemcitabine until around six months where the difference in hazards has diminished.



HR at start of graph very high so cut around HR=7

Figure 4.8: Hazard ratio for treatment over time in time-dependent model.

Inclusion of a trial variable

As mentioned in Section 4.4.1, a decision was made not to include trial as a variable in the prognostic model. Here, this assumption was now checked by adding a trial variable to the model, to assess if the model fit significantly improved. Improvement was tested using a likelihood ratio test and also by comparing AIC and BIC between the two models (including and excluding the trial variable) within each imputed datasets. The likelihood ratio tests compares the two models with and without the trial variable within each imputed dataset and all p-values were greater than 0.05 (ranging from p=0.067 to p=0.173). The AIC was similar in models with and without trial and BIC was lower in models without trial, suggesting that trial does not improve the model fit significantly. The exclusion of the 'trial' variable thus appears justified, and allows for greater ease of use.

4.6.3 Internal validation of the time-dependent model

Due to the time-dependent effect fitted in the model, the two treatment groups effectively have different baseline hazard functions that are not proportional. This has implications when creating risk-groups, as the time at which the linear predictor is calculated affects which risk group a patient is categorised into. For example, when calculating the risk score (linear predictor) at 3 months, the hazard is much higher for patients not on gemcitabine, making them more likely to fall into a higher risk group. In comparison, at 25 months this is not the case as hazards in the groups are similar. To avoid this problem, risk groups were created within each treatment group separately, as the baseline hazard function is assumed to be the same for everyone on the same treatment.

Figure 4.9 shows the observed and expected survival curves for the treatment groups separately in one of the imputed datasets (others differed slightly but were similar). The expected survival curves are given for both the time-dependent model and the proportional hazards model from Section 4.6.1 for comparison. The same risk groups are used to compare both models and are based on quartiles of the risk score from the model including the time-dependent effect for treatment. If looking across the whole duration to maximum follow-up, the fit of the time-dependent model looks poor for most of the risk groups in both treatment groups. For some risk groups, the model *over-predicts* survival probabilities over time, such as risk group 3 in both treatments. In other risk groups, the model *under-predicts* survival, for example, risk group 2 of the no gemcitabine group.

The calibration of the time-dependent model appears to be better than the proportional hazards model as the expected survival curves are generally closer to the observed survival than for the proportional hazards model. Also, calibration looks better for all risk groups in the gemcitabine group than in the no gemcitabine group at earlier time points. Risk group 1 in the no gemcitabine group shows that observed survival is higher than predicted using the model until about 8 months.

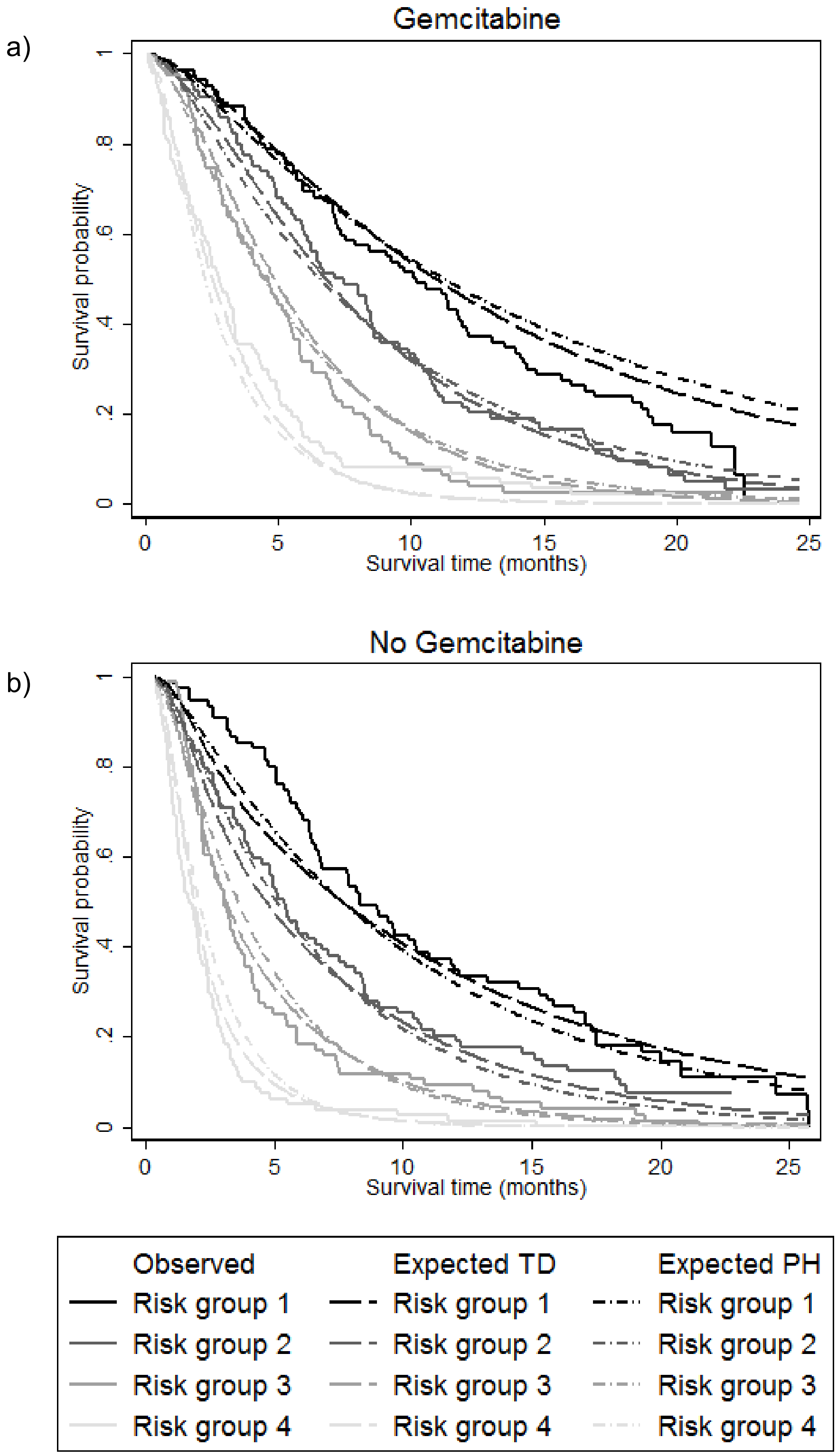


Figure 4.9: Kaplan-Meier (observed) and mean predicted (expected) survival using the time-dependent treatment effect model (TD) and proportional hazards treatment effect model (PH) for (a) gemcitabine and (b) no gemcitabine treatment groups in one of the imputed datasets.

The C-statistic for the time-dependent model is 0.655 averaged across imputed datasets with a range of 0.652 to 0.657, which is lower than for the proportional hazards model (average C-statistic=0.705). However, there is a trade-off between discrimination and calibration, as calibration was slightly better than in the proportional hazards model (Figure 4.9). Essentially, in the proportional hazards model, the larger discrimination is caused by the predicted probabilities being slightly too high in the upper group and slightly too low in the lower group, revealed by the poorer calibration in the proportional hazards model. A trade-off in discrimination and calibration has been illustrated previously, e.g. Debray et al. using logistic regression models.[149]

In conclusion, the time-dependent model performs better than the proportional hazards model in each of the imputed datasets as was seen by lower AIC and BIC values. The comparison of observed and expected survival curves (Figure 4.9) showed that calibration was slightly better in the time-dependent model than in the proportional hazards model, and consequently discrimination was slightly lower in the time-dependent model. Overall, the model only has moderate discriminatory ability and it may be necessary to investigate further prognostic factors to improve this before it could be used in a clinical setting, to better distinguish those who will or will not have a poor outcome. However, the model does appear to have good calibration and discrimination within the first six months for gemcitabine patients. Gemcitabine is often used as a first line therapy for advanced stage pancreatic cancer and NICE guidelines state that it is considered for patients with advanced or metastatic pancreatic cancer and a Karnofsky performance≥50.[170,171] Due to gemcitabine's common usage and the poor prognosis in advanced stage pancreatic cancer patients in general, a model for prognosis of patients on gemcitabine up to six months may still be useful. Therefore, the developed model may be important, and so external validation studies are necessary.

4.7 Predictions for individuals

Risk grouping has so far been used to help examine calibration, as observed versus expected risk can only be compared across groups of individuals. Calibration was quite good for risk groups of patients on gemcitabine, especially up to six months. However, once calibration is assessed at the group level, the availability of the baseline hazard allows individuals to have their own predicted risk over time, which may be somewhat different to the average in their risk group. This is a great advantage of the flexible parametric approach over the Cox model. Figure 4.10 illustrates this, by showing how different the predicted risk of two individuals within the same risk group can be. At six months, the predicted survival probabilities are 0.634 and 0.480 for patient 1 and 2 respectively, with a difference in survival probabilities of 0.154. By modelling the baseline hazard explicitly, individual predictions can be made rather than being restricted to grouping patients with potentially very different prognoses. If the baseline risk was unavailable (as from a Cox model), both patients 1 and 2 would be given the same predicted outcome risk from risk group 2 and thus their differences are hidden.

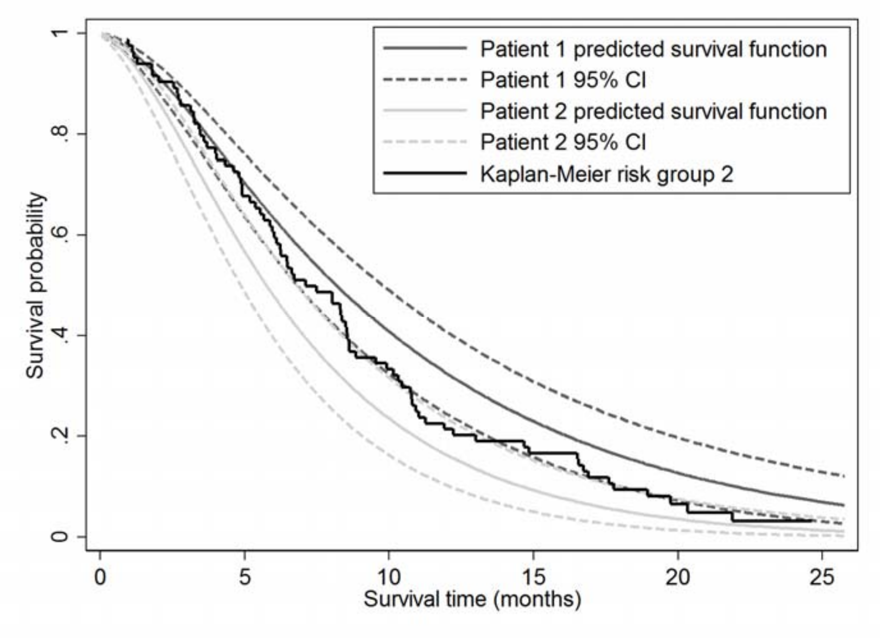


Figure 4.10: Predicted survival functions as derived from the developed time-dependent model for two individuals from risk group 2 of gemcitabine patients, compared to the observed Kaplan-Meier estimate of survival for risk group 2.

4.8 Discussion

A prognostic model developed for patients with advanced pancreatic cancer could be used to aid treatment decisions, by potentially identifying patients that could receive surgery or identifying patients suitable for future clinical trials or off-study treatments.[156] Royston-Parmar models are a novel approach, currently not often used to develop prognostic models (as highlighted in previous chapters), however a recent publication by Baade et al. which includes Royston as an author did develop a model using the Royston-Parmar approach (described in Section 1.4.6).[41] This chapter used Royston-Parmar modelling to develop a prognostic model for advanced pancreatic cancer that allows individualised predictions over time, and identified methodological issues when using clinical trials data for this purpose. The key findings and limitations are now discussed.

4.8.1 Summary and comparison to previously published model

A key aim of this chapter was to build upon the prognostic model published by Stocken et al. in which the clinical trials data were used to develop a prognostic model using a Cox proportional hazards model and a different categorisation of treatment.[156] In doing so, the model has now been extended by modelling the baseline hazard using Royston-Parmar models and by changing the categorisation of treatment to gemcitabine versus no gemcitabine rather than marimastat versus no marimastat; a decision that was based on the original trial publication findings, and because advanced stage pancreatic cancer patients often receive gemcitabine as a first line chemotherapy if their Karnofsky performance score ≥50 (recommended by NICE).[154,155,171] The final model was therefore very different to that developed by Stocken et al.[156]

Another reason for the difference in models is that Stocken et al. took a different approach to handling the functional form for continuous variables. The authors tested fractional polynomials in univariable analysis and selected second degree fractional polynomials for CA19-9, and modelled BUN, AST and alkaline phosphatase linearly. In contrast to this, the model developed in this chapter used MFP (restricted to first degree fractional polynomials) to test for non-linear functions in the same process as the variable selection. This meant that the functional form of each variable was tested with the other variables in the model in an iterative process (i.e. in the multivariable analysis rather than univariable analysis). This resulted in non-linear functions for AST (power: -0.5), bilirubin (power: -1), LDH (natural log) and CA19-9 (natural log).

4.8.2 Limitations and further work

The apparent discrimination of the final model produced with a time-dependent effect for treatment was 0.655 and was lower than for the proportional hazards model which had a C-statistic of 0.705, but the apparent calibration performance was better in the risk groups. Neither model discriminates between individuals with high and low risk of death particularly well. Model performance is usually somewhat optimistic when evaluated internally and therefore model discrimination is likely to be even lower when evaluated in external data.[58] Interestingly, Collins et al. also found that discrimination performance is generally higher when the external validation is done by the same authors that developed the model (mean C-statistic 0.78) compared to independent investigators (mean C-statistic 0.72), shown in Figure 1.8.[54]

If external data become available, the performance of both the original model developed by Stocken et al. and the one reported in this chapter should be evaluated and compared. The models were developed using the same data but use different treatment groupings and modelling strategies. External data would be required to judge if one model performs better than the other. However, given the generally low apparent discrimination performance, it is likely that additional predictors will be required to improve discrimination and the model therefore substantially updated before it can be used. For this reason (i.e. because the model needs to be improved upon before it can be recommended), shrinkage of the predictor effects (βs) in the final model and optimism-adjusted estimates of discrimination were not considered necessary here.[12,22,172] However, the predictors identified here and the modelling approach can be used as a starting point for further research. With more data, additional interaction terms might also be explored.

4.8.3 Issues in using clinical trials data

A review of prognostic models in cancer found 33% of the articles included used RCTs to develop prognostic models.[83] Another important aim of this chapter was to catalogue the issues faced when using RCT data to develop a prognostic model. These were highlighted throughout the chapter and are now discussed below.

(1): Dealing with multiple treatment groups in the development of a prognostic model

The chapter considered separate models for each of the two treatment groups and then a combined model. Building a single model that utilises all treatment groups is recommended to maximise data and minimise the degrees of freedom used in modelling treatment. In the pancreatic cancer data, keeping the original treatment categories (six categories) would have meant inadvertently modelling a trial effect as the treatment categories were different across the two trials. By modelling all treatment groups in the same analysis, the power to detect genuine predictors is increased and it is therefore more likely to generate a reliable model, even if the model predictions are ultimately only required for one of the treatment groups. One caveat is that non-proportional hazards, and potentially interaction terms, may be required to handle the multiple treatment groups as illustrated, and this increases complexity.

Of course, if treatments are ineffective or have similar effectiveness, then modelling can be made simpler by removing treatment,[8] or alternatively smaller treatment groupings can be identified. Exploring the shape of the baseline hazard function of each treatment category can aid the decision to combine categories if the shape of the underlying hazard function looks reasonably similar. In this example, this led to combining treatment categories in which patients received gemcitabine (regardless of whether they received marimastat or not). This decision should of course also consider clinical relevance and evidence of treatment effects.

(2): Handling multiple trials in the development of a prognostic model

The inclusion of a trial variable is problematic in a prognostic model, as the intention is to predict prognosis in new patients that do not belong to a trial. For this reason, in this chapter, the data from both trials were combined and steps were taken to ensure that differences between the two trials had been properly accounted for. The shape of the baseline hazard was plotted initially to check for similarity before combining and a test of including a trial variable into the final model was non-significant, suggesting that the variation between the trials had adequately been accounted for in the other variables and a trial variable was no longer necessary. If there are more than two trials in the study, a random effect could be considered for the baseline hazard function to enable a separate baseline effect per trial, but still producing an average baseline hazard to be used for prediction.[149] The issue of multiple studies in prediction research is considered again in Chapters 5 and 6.

(3): Dealing with missing data

Missing data is a common issue and one that is not specific to clinical trials. Due to the high proportion of missing data for nodes (24%), it was excluded from model development. If this proportion of missing data is representative of measurement of nodes in clinics, and the prognostic model included nodes as a prognostic factor, the model could not be used in almost a quarter of patients. In developing a prognostic model for survival in kidney cancer patients, Royston et al. made the decision to exclude any variables missing in more than 10% of patients and suggest that decisions regarding handling of missing data (as well as other issues) should be decided before development begins.[19]

Multiple imputation was used to account for missing data in the other variables considered for inclusion in the prognostic model. This resulted in added complexity in variable selection as there were 25 datasets rather than one dataset and therefore a stacked and weighted approach was used based on the literature. There were computational challenges in using MFP, multiply imputed data and Royston-Parmar models in combination. Some post-estimation commands were adapted to deal with the current data but for validation and Kaplan-Meier plots, imputed datasets had to be considered separately.

(4 and 5): Dealing with trial stratification factors and using selection procedures to identify prognostic factors in a model

Whenever possible, clinical knowledge should influence the variable selection rather than being driven by the data alone. Trial stratification factors should be included in the prognostic model as they were part of the trial design, having been considered clinically important in relation to the outcome of interest. This is akin to utilising external evidence that certain variables are prognostic, and therefore including them automatically in the model. For this reason, in this chapter, the stratification factors sex and stage were included in the model regardless of statistical significance.

In exploratory analyses of MFP (not reported), it was found variables that should be modelled linearly may be excluded when testing for more complex functions such as second degree FP functions due to the additional degrees of freedom used in the test. This can be a problem when there is not enough data. Use of backward elimination meant that haemoglobin was retained in one of the treatment specific models even with a large p-value. The use of automatic selection procedures should always be used with caution and is not an issue exclusive to using clinical trials data.

(6): Use of trial data for model validation

Validation in external data is critical for all prognostic models developed. When there are two or more trials or two or more treatment groups, the model could be developed in one and validated in the other. This was demonstrated using the two treatment groups in an internal-external validation in which the gemcitabine model was tested in no gemcitabine patients and vice versa. The treatment specific models showed moderate discrimination and did not calibrate well in the other treatment group, due to the different baseline hazard functions. However, as discussed removing data reduces power, and so generally it is not recommended to exclude data from model development unless the datasets are extremely large. This was evident by the single model (developed using all treatment groups) having more power to identify prognostic factors. Steyerberg et al. agree that data splitting is inefficient and instead recommend putting data altogether for model development, and using bootstrapping for internal validation purposes.[57]

Interestingly, when there are multiple trials, Royston et al. and Debray et al. recommend an internal-external cross-validation approach whereby each trial is omitted and then – if performance is always adequate regardless of the omitted trial – the final model proposed is based on all available data.[69,149] Chapter 5 extends that idea further.

A good example of RCT data being used to develop and validate a prognostic model is in traumatic brain injury.[151] The IMPACT database (consisting of eight RCTs and three observational studies) was used to develop and internally validate a prognostic model using internal-external cross-validation, each time excluding a study. The model was then externally validated using data from the CRASH trial.[151]

4.9 Conclusion

This chapter illustrated how clinical trials data can be used to develop a prognostic model for making absolute risk predictions over time and identified some of the challenges that researchers will face. A new pancreatic model was developed and internally validated, but due to its moderate discrimination, additional predictors are likely required before it can be useful for practice and further external validation is also necessary.

So far, this thesis has focused mainly on prediction model development, in terms of both application (to hip replacements and pancreatic cancer) and methodology issues such as flexible parametric modelling and using trials data. The emphasis in the next couple of chapters moves instead to the validation performance of a prediction model. All prediction models require external validation to check that the model predicts reliably in new data from similar (or even different) settings or populations. The next chapter proposes a new meta-analysis method for validation of a prediction model when multiple studies/datasets are available.

CHAPTER 5: VALIDATION OF A PREDICTION MODEL ACROSS MULTIPLE STUDIES

5.1 Introduction and background

The previous chapter showed the development and internal validation of a prognostic model using randomised clinical trials data. The final model's performance was only internally validated, as external data were not available. Internal validity can be considered as checking the 'reproducibility' of the model.[12] The model is expected to perform best in the data in which it was developed and over-fitting can often lead to very optimistic parameter estimates.[172] However, methods such as bootstrapping of the original dataset can be used to evaluate this over-optimism and adjust the validation performance statistics accordingly.[10] Ideally, however, an external dataset would be used to validate the model (external validation) to assess how the model performs in a different but related population to that used for model development. By testing the model in external data, the 'generalizability' of the model can be evaluated to see how well the model performs in other similar populations or settings.[12]

External validation is not possible when there is a single dataset available for model development *and* validation, as in the previous chapter. Randomly splitting a single dataset into two still produces samples from the same source, and so any testing of model performance is still considered to be internal validation.[20] Examples of prediction models developed and validated using data splitting by clusters are the Hippisley-Cox et al. prediction models included as part of the literature review (Chapter 3).[119,123,138,139] These models were developed using the QResearch database, which is a very large database containing data from multiple centres (approximately 530 general practices). The data were split by randomly allocating two thirds of the practices to the development set and reserving the other third for validation. The authors validated performance of each model across all practices in the validation set but did not assess how the model performed in individual practices or areas.

This chapter considers the scenario in which *multiple* studies (i.e. multiple datasets) are available for development and validation of a prognostic model. The multiple studies can be used in several ways to develop and validate a prognostic model. One approach is splitting the studies into two sets, one for development and the other for validation. This method ensures that the same data are not used for both development and validation. Another option for dealing with individual participant data (IPD) from multiple datasets (studies) is to develop the model using all of the available data (stratifying by study) and use internal-external cross-validation (IECV) to evaluate performance across the studies. The IECV approach was proposed by Royston et al.[69] and will be discussed in more detail below. Put simply, once a model has been developed using all of the studies it is then re-fitted several times (re-estimating the βs), each time excluding one study, and then validating the model 'externally' using the excluded study. Multiple validation results of model performance are therefore available following IECV, one for each excluded study. IECV is an extension to internal validation in that performance of the model is 'externally' validated in each excluded study, therefore also evaluating the generalisability of the model.[149]

This chapter proposes novel meta-analysis methods for summarising the performance of a model across multiple studies. Given multiple study estimates of the same validation performance statistic (e.g. C-statistic), one could just look at the **average** performance of a model across all studies. However, this chapter highlights why this is a missed opportunity as average performance is an incomplete picture, and additionally **heterogeneity** in model performance across the different studies (settings, clusters etc.) is also of importance. Also, some of the performance statistics (such as the C-statistic and calibration slope) are potentially correlated, and therefore multivariate meta-analysis methods are proposed to appropriately account for this correlation when summarising model performance.[173-175]

5.1.1 Aims and outline

The aim of this chapter is to extend the IECV approach by incorporating meta-analysis of validation performance statistics. In particular, to propose univariate and multivariate meta-analysis methods for summarising the IECV performance of a prediction model, and to help identify the best strategy for implementing the model in new populations. Two examples are used in this chapter. The first relates to a prediction model for deep vein thrombosis risk developed using logistic regression and IECV. The second example relates to a prognostic model for breast cancer mortality developed using Royston-Parmar models and IECV. It is worth noting that whereas previous chapters focused specifically on modelling time-to-event data, it is important to evaluate model performance regardless of the outcome type (binary or time-to-event). Therefore this chapter includes both logistic and Royston-Parmar models as examples.

The remainder of the chapter is outlined as follows. Firstly, the IECV approach is described in fuller detail and then the datasets are introduced. Univariate and multivariate meta-analysis methods are then described, with application to the datasets to illustrate their usefulness. Discussion is then given on the benefits and limitations of the work.

5.2 Internal-external cross-validation approach

Internal-external cross-validation (IECV) is a method that was proposed for situations when several datasets are available for development and validation of a model.[69] Development of prediction models will not be considered here as it has been discussed extensively in previous chapters. Instead, the predictors included in the model and the functional form of continuous predictors are assumed to be pre-specified and do not change within the validation process but the model parameters such as the *β*-coefficients are re-estimated.

5.2.1 IECV framework and notation

To understand how the IECV approach works, consider a set of *K* studies where *k*=1, …, *K.* Using the notation of Royston et al.,[69] let $\mathscr{S}_k$ be the *k*th study reserved for validation performance and $\mathscr{S}_{(k)}$ be the set of studies excluding study $\mathscr{S}_k$ (Table 5.1). The set of studies $\mathscr{S}_{(k)}$ are to be used to derive the model (estimate the *β*-coefficients), and an individual study within this derivation set will be referred to as study *j* to differentiate it from the excluded study *k* reserved for validation. In each cycle of the IECV, the model is fitted using the set of studies $\mathscr{S}_{(k)}$, (potentially stratifying by study),[149] and then validated in study $\mathscr{S}_{k.}$ Therefore in cycle one, a model is fitted using studies 2 to *K* ($\mathscr{S}_{(1)}$) and externally validated in study 1 ($\mathscr{S}_1$). In the next cycle of IECV, study 2 is excluded from the derivation set ($\mathscr{S}_{(2)}$) and used to validate the model ($\mathscr{S}_2$). This is repeated for *K* cycles, each time excluding a different study from the derivation set and reserving it for validation. This means that the *β*-coefficients of the developed model are allowed to vary from cycle to cycle and in this chapter re-estimating the *β*s will be referred to as 'model derivation' to distinguish it from model development which would be done prior to beginning the internal-external cross-validation of model performance. After proceeding through all cycles in the IECV approach, there are *K* estimates produced for each performance statistic of interest (e.g. C-statistic, calibration slope, etc.). The focus in this chapter is on how meta-analysis can be used to combine the *K* estimates of model performance for each performance statistic, to give an overall summary of model performance across the *K* excluded studies.

Table 5.1: Internal-external cross-validation approach for *K* studies.

Cycle	Studies used to derive model (estimate *β*s)	Studies used to evaluate model performance
1	$\mathcal{S}_{(1)}$ = Studies 2 to *K*	$\mathcal{S}_{1}$ = Study 1
2	$\mathcal{S}_{(2)}$ = Studies 1 and 3 to *K*	$\mathcal{S}_{2}$ = Study 2
⋮	⋮	⋮
K	$\mathcal{S}_{(K)}$ = Studies 1 to *K*-1	$\mathcal{S}_{K}$ = Study *K*

5.2.2 Summarising performance

Royston et al. propose summarising the performance statistics using a weighted average to obtain an overall summary of performance. For example, they consider the D-statistic which is a measure of separation defined in Chapter 1 (Section 1.6.3).[69] An overall D-statistic can be calculated by combining the D-statistics from all $\mathcal{S}_k$ (D_k's for k=1 to K) by

$$D_{IECV} = \frac{1}{K}\sum_{k=1}^{K} w_k\, D_k \qquad \textbf{(5.1)}$$

where w_k are the standardised weights,

$$w_k = \frac{w_k^*}{\overline{w^*}} = \frac{\dfrac{1}{s_k^2}}{\dfrac{1}{K}\displaystyle\sum_{l=1}^{K}\dfrac{1}{s_l^2}}.$$

The weights w_k^* are based on the inverse variance method ($1/s_k^2$ where s_k is the standard error for D_k and can be estimated using bootstrap resampling) and are then standardised by dividing by the average weight $\overline{w^*}$. The standard error for D_{IECV} which can be used to give confidence intervals, can be calculated as

$$SE(D_{IECV}) = \left[\frac{1}{K}\,\frac{1}{K\text{-}1}\sum_{k=1}^{K} w_k\;(D_k - D_{IECV})^2\right]^{1/2}. \qquad \textbf{(5.2)}$$

Although Royston et al. consider the D-statistic, their approach could also be used for other measures of model performance (such as the C-statistic, or calibration slope).

5.2.3 Between-study heterogeneity

The overall summary statistic of model performance proposed by Royston et al. in (5.1) is a weighted average of the individual estimates of model performance obtained from each study. Alternatively, this chapter now proposes that the performance statistics should be pooled using univariate and multivariate random-effects meta-analysis methods (described in Sections 5.4 and 5.5 below). Using a random-effects meta-analysis to pool statistics has the advantage of estimating the between-study heterogeneity (i.e. how much the model performance varies across studies). The between-study heterogeneity is important when considering consistency of performance across the populations in which the model will be implemented. An ideal model will have little or no heterogeneity, and consistently good performance. The further away from this ideal, the less reliable the model is. For example, if the model performs well on average but there is large heterogeneity, this would mean that in some settings the model performs poorly. If the average performance is poor and there is no heterogeneity, then the model performs consistently poorly.

Heterogeneity is also informative when deciding how to implement a model. Using the IECV approach, it is possible to look at different strategies for implementing the model in a new population. In this chapter the examples will consider three different implementation strategies when applying the model to a new study: (i) using the average intercept/baseline hazard from the derived model, (ii) using an intercept/baseline hazard from one of the studies included within the derivation data that has a similar outcome prevalence/incidence to the intended population, or (iii) estimating (recalibrating) the intercept/baseline hazard using new data from the intended population. These options for the model intercept were discussed by Debray et al.[149] Again, extension of the IECV approach to include meta-analysis allows model implementation strategies to be formally evaluated, with those strategies with lowest heterogeneity potentially preferred if average performance is acceptable.

5.3 Data

In this chapter, meta-analysis methods are applied to two datasets containing information on multiple performance statistics obtained using the IECV approach. The first dataset consists of performance statistics from a diagnostic prediction model that was developed using logistic regression for a binary outcome. The second dataset considers validation of a prognostic model where a Royston-Parmar survival model was used. The same predictors/functions were included in all cycles of IECV, allowing only the predictor effects (β-coefficients) to be re-estimated (referred to as model derivation in this chapter) and not the predictors included or functions to change. The baseline hazard function was estimated in each cycle of IECV, but then the shape was fixed in the validation study (the implementation strategy allowed a change in the constant not the shape of the function, detailed in Section 5.3.2). Note that the key focus here is on combining the performance statistics using meta-analysis, rather than how the set of included predictors were selected (e.g. in terms of handling of continuous predictors, etc.).

5.3.1 Dataset one: Deep vein thrombosis

The first of these datasets was obtained from Debray et al. and was referred to as case study three in their paper ('weak to moderate heterogeneity in predictor-outcome associations').[149] The original data contained IPD from 12 studies with study sample sizes ranging from 153 to 1768 patients. The studies were used by Debray et al. to develop a logistic regression model to predict the risk of having deep vein thrombosis (DVT) in patients that were suspected of having DVT. There were a total of 10014 patients across the 12 studies and 1897 (18.9%) of them had a true DVT.

The prediction model was developed using logistic regression and including the following pre-specified variables: sex (male, female), surgery (recent surgery or bedridden, no recent surgery or bedridden) and calf difference (≥3cm, <3cm). The model fitted within each cycle of the IECV approach can be written for individual *i* in study *j* as

$$\text{logit}(p_{ij})=\alpha_j+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_n x_{in} \qquad \textbf{(5.3)}$$

The parameter estimates for all derived models (fitted as part of the IECV approach) are given in Table 5.2. The estimated *β*-coefficients were quite consistent across derived models suggesting that, regardless of the study excluded, the models were similar.

Table 5.2: DVT model parameter estimates for study-specific intercept and predictors, as provided by Debray et al. on request.

Study (k) – excluded for validation	Study-specific intercept ($\hat{\alpha}_j$)												Sex ($\hat{\beta}_1$)	Surgery ($\hat{\beta}_2$)	Calf diff. ($\hat{\beta}_3$)
	Study 1	Study 2	Study 3	Study 4	Study 5	Study 6	Study 7	Study 8	Study 9	Study 10	Study 11	Study 12			
Study 1	-	-1.256	-2.528	-1.807	-2.511	-2.316	-3.061	-1.839	-2.165	-2.339	-2.038	**-2.788**	0.372	0.606	1.304
Study 2	-2.694	-	-2.555	-1.823	-2.538	-2.338	-3.080	**-1.857**	-2.192	-2.353	-2.059	-2.817	0.370	0.624	1.344
Study 3	-2.670	-1.253	-	-1.805	**-2.504**	-2.312	-3.062	-1.840	-2.157	-2.336	-2.033	-2.779	0.400	0.556	1.286
Study 4	-2.604	-1.191	-2.443	-	-2.426	-2.245	-3.005	**-1.785**	-2.082	-2.289	-1.967	-2.700	0.366	0.513	1.197
Study 5	-2.682	-1.265	**-2.538**	-1.814	-	-2.324	-3.070	-1.849	-2.173	-2.343	-2.045	-2.797	0.387	0.579	1.315
Study 6	-2.672	-1.255	-2.529	-1.811	**-2.508**	-	-3.066	-1.843	-2.161	-2.343	-2.041	-2.783	0.409	0.589	1.277
Study 7	**-2.672**	-1.255	-2.529	-1.810	-2.509	-2.319	-	-1.842	-2.162	-2.342	-2.040	-2.784	0.401	0.594	1.283
Study 8	-2.677	-1.260	-2.532	**-1.812**	-2.513	-2.321	-3.068	-	-2.166	-2.342	-2.042	-2.788	0.403	0.572	1.294
Study 9	-2.660	-1.245	-2.515	-1.797	-2.498	-2.304	-3.050	-1.828	-	-2.329	**-2.026**	-2.777	0.350	0.611	1.300
Study 10	-2.677	-1.260	-2.531	-1.809	-2.515	-2.318	-3.066	-1.844	-2.168	-	-2.039	**-2.791**	0.381	0.572	1.313
Study 11	-2.678	-1.261	-2.533	-1.812	-2.515	-2.321	-3.069	-1.847	**-2.167**	-2.342	-	-2.790	0.399	0.565	1.301
Study 12	-2.672	-1.256	-2.528	-1.808	-2.509	-2.317	-3.064	-1.841	-2.162	**-2.339**	-2.038	-	0.391	0.585	1.293

Note: Bold numbers represent the intercept used for validation in the excluded study for strategy 3, where the intercept from the study with the closest prevalence was selected.

The set of external performance statistics for the 12 studies were calculated for each of three different implementation strategies for using the model in practice (Table 5.3). For strategies 1 and 3, a model with study-specific intercepts was fitted (using a dummy variable for study) and a random-intercept model was used for strategy 2. The implementation strategies for the model intercept when validating in the excluded study are as follows:

Strategy 1: Estimate a new study-specific (recalibrated) intercept $\widehat{\alpha_k}$ in the validation dataset.

Strategy 2: Use the average intercept from the derived random-intercept model in the validation dataset.

Strategy 3: Select a study-specific intercept $\widehat{\alpha}_j$ from one of the studies included in the derivation dataset that had the most similar prevalence to the validation dataset.

The performance statistics measured in the validation dataset were the C-statistic, calibration slope, expected/observed number of events (E/O) and calibration-in-the-large. These performance statistics were introduced and defined in Section 1.6.3.

Estimates and standard errors of the four performance statistics are given in Table 5.3 for each study, and the within-study correlation of the estimates for each pair of performance statistics is shown in Table 5.4. Standard errors and within-study correlations were obtained by non-parametric bootstrapping with 100 samples. Calibration-in-the-large and log(E/O) have a near perfect negative correlation (-0.997 to -1.000), by definition, and calibration slope and C-statistic have very strong within-study correlations ranging from 0.90 to 0.98. These correlations were similar regardless of the implementation strategy used.

Table 5.3: Performance statistics (and standard errors, σ_k) from the DVT model using IECV approach with three different strategies for the intercept, provided by Debray et al. on request.

Study *k*	Strategy 1: Intercept estimated in external validation study				Strategy 2: Average intercept taken from derived random-intercept model				Strategy 3: Intercept from a study included in derivation with a similar prevalence			
	CITL	Calibration slope	Log(E/O)	C-statistic	CITL	Calibration slope	Log(E/O)	C-statistic	CITL	Calibration slope	Log(E/O)	C-statistic
1	-0.172	0.903	0.140	0.678	-0.440	0.905	0.349	0.678	0.114	0.905	-0.094	0.678
	(0.098)	(0.131)	(0.080)	(0.024)	(0.098)	(0.136)	(0.081)	(0.025)	(0.094)	(0.132)	(0.078)	(0.024)
2	-0.051	0.741	0.028	0.653	1.105	0.745	-0.709	0.653	0.583	0.736	-0.344	0.652
	(0.079)	(0.100)	(0.042)	(0.019)	(0.078)	(0.100)	(0.043)	(0.019)	(0.084)	(0.102)	(0.045)	(0.019)
3	-0.172	1.418	0.135	0.761	-0.292	1.396	0.223	0.756	-0.042	1.390	0.036	0.755
	(0.224)	(0.397)	(0.174)	(0.055)	(0.225)	(0.405)	(0.176)	(0.057)	(0.223)	(0.408)	(0.173)	(0.057)
4	-0.084	1.432	0.060	0.735	0.488	1.434	-0.367	0.736	0.031	1.434	-0.022	0.736
	(0.054)	(0.100)	(0.039)	(0.014)	(0.057)	(0.099)	(0.040)	(0.014)	(0.057)	(0.102)	(0.040)	(0.014)
5	-0.185	0.742	0.141	0.649	-0.267	0.744	0.201	0.649	0.016	0.749	-0.011	0.650
	(0.122)	(0.164)	(0.094)	(0.031)	(0.125)	(0.165)	(0.096)	(0.032)	(0.121)	(0.159)	(0.093)	(0.031)
6	-0.149	1.030	0.112	0.699	-0.055	1.044	0.042	0.701	0.187	1.035	-0.144	0.700
	(0.082)	(0.114)	(0.062)	(0.021)	(0.084)	(0.119)	(0.064)	(0.022)	(0.084)	(0.119)	(0.063)	(0.022)
7	-0.178	1.017	0.156	0.694	-0.877	1.020	0.732	0.694	-0.399	1.014	0.344	0.693
	(0.090)	(0.117)	(0.080)	(0.023)	(0.089)	(0.122)	(0.078)	(0.023)	(0.090)	(0.123)	(0.080)	(0.024)
8	-0.115	0.932	0.081	0.663	0.464	0.936	-0.340	0.663	-0.038	0.943	0.028	0.665
	(0.133)	(0.189)	(0.093)	(0.035)	(0.129)	(0.192)	(0.090)	(0.034)	(0.132)	(0.192)	(0.092)	(0.034)
9	-0.139	0.994	0.098	0.690	0.118	0.991	-0.084	0.689	-0.132	0.996	0.093	0.690
	(0.068)	(0.099)	(0.048)	(0.017)	(0.073)	(0.096)	(0.052)	(0.017)	(0.070)	(0.104)	(0.050)	(0.017)
10	-0.127	0.695	0.103	0.636	-0.081	0.693	0.066	0.635	0.447	0.699	-0.373	0.637
	(0.150)	(0.215)	(0.122)	(0.037)	(0.144)	(0.219)	(0.116)	(0.038)	(0.146)	(0.208)	(0.119)	(0.036)
11	-0.135	0.921	0.094	0.701	0.258	0.921	-0.185	0.700	0.132	0.916	-0.093	0.700
	(0.111)	(0.140)	(0.078)	(0.026)	(0.111)	(0.145)	(0.078)	(0.026)	(0.110)	(0.140)	(0.077)	(0.026)
12	-0.197	0.936	0.160	0.673	-0.570	0.923	0.440	0.671	-0.458	0.930	0.359	0.672
	(0.191)	(0.269)	(0.155)	(0.048)	(0.190)	(0.264)	(0.155)	(0.048)	(0.193)	(0.256)	(0.157)	(0.046)

Note: CITL refers to calibration-in-the-large and log(E/O) refers to log of the expected/observed number of events.

Table 5.4: Within-study correlations (ρ_{Wk}) between performance statistics for the DVT model, obtained through bootstrapping, provided by Debray et al. on request.

	Study *k*	CITL & calibration slope	CITL & log(E/O)	CITL & C-statistic	Calibration slope & log(E/O)	Calibration slope & C-statistic	Log(E/O) & C-statistic
Strategy 1: Intercept estimated in external validation study	1	-0.006	-1.000	-0.022	0.006	0.961	0.021
	2	-0.032	-1.000	-0.046	0.033	0.977	0.046
	3	-0.001	-0.999	0.009	0.001	0.955	-0.011
	4	0.118	-1.000	0.045	-0.117	0.919	-0.045
	5	0.046	-1.000	0.029	-0.046	0.983	-0.029
	6	-0.010	-1.000	-0.045	0.011	0.948	0.043
	7	0.047	-1.000	0.002	-0.047	0.912	-0.003
	8	0.071	-1.000	0.032	-0.072	0.953	-0.034
	9	-0.005	-1.000	-0.011	0.005	0.980	0.011
	10	0.108	-1.000	0.064	-0.108	0.900	-0.064
	11	0.000	-1.000	-0.025	0.002	0.956	0.026
	12	-0.035	-1.000	-0.029	0.036	0.980	0.030
Strategy 2: Average intercept taken from derived random-intercept model	1	-0.051	-1.000	-0.054	0.053	0.960	0.054
	2	0.046	-0.990	0.037	-0.054	0.976	-0.035
	3	-0.051	-0.999	-0.037	0.052	0.958	0.037
	4	0.062	-0.999	-0.009	-0.065	0.928	0.014
	5	0.069	-1.000	0.055	-0.069	0.981	-0.056
	6	0.031	-1.000	-0.013	-0.031	0.959	0.013
	7	0.018	-0.999	-0.025	-0.013	0.923	0.025
	8	0.105	-0.999	0.042	-0.106	0.951	-0.038
	9	0.074	-1.000	0.068	-0.074	0.976	-0.068
	10	0.035	-1.000	0.012	-0.034	0.895	-0.012
	11	0.053	-1.000	0.015	-0.052	0.953	-0.012
	12	0.000	-0.999	-0.001	-0.003	0.981	-0.003
Strategy 3: Intercept from a study included in derivation with a similar prevalence	1	0.005	-1.000	-0.012	-0.004	0.963	0.013
	2	-0.006	-0.997	-0.033	0.002	0.978	0.034
	3	-0.053	-0.999	-0.049	0.054	0.959	0.050
	4	0.082	-1.000	0.015	-0.082	0.925	-0.014
	5	0.034	-1.000	0.022	-0.033	0.982	-0.022
	6	0.063	-1.000	0.020	-0.065	0.956	-0.019
	7	-0.014	-1.000	-0.025	0.015	0.925	0.025
	8	-0.002	-1.000	-0.052	0.001	0.954	0.051
	9	-0.010	-1.000	-0.021	0.011	0.981	0.021
	10	0.020	-0.999	-0.043	-0.024	0.892	0.043
	11	0.038	-1.000	-0.001	-0.040	0.956	0.000
	12	-0.036	-0.999	-0.041	0.036	0.980	0.039

Note: CITL refers to calibration-in-the-large and log(E/O) refers to log of the expected/observed number of events.

5.3.2 Dataset two: Breast cancer

The second dataset consisted of IPD from several European studies. The studies included patients with breast cancer that were followed up over time for the outcome of death. The studies were grouped by country before developing a multivariable prediction model and using IECV to evaluate performance of the model. Therefore the cluster here is the country, and there were eight countries considered. The IPD from each country ranged in size from 69 patients to 3242 (total of 7435), with 2043 events that occurred across all countries. The maximum follow-up duration was 120 months and the median follow-up duration across all countries was 86.3 months (95% CI: 85.1 to 87.8 months).

The prognostic model was developed by Hua,[176] using Royston-Parmar flexible parametric modelling. The general form of the model fitted in each IECV cycle can be written on the log cumulative hazard scale for individual *i* in country *j* of the derivation set as

$$\text{Ln}(H_{ij}(t))=H_{0j}(t)+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_n x_{in} \qquad \textbf{(5.4)}$$

where $H_{0j}(t)$ is the country-specific baseline cumulative hazard function modelled using restricted cubic splines with 3 d.f. This baseline (cumulative) hazard function can be written as

$$H_{0j}(t)=\gamma_{0j}+\gamma_1 \ln t+\gamma_2 z_1(\ln t)+\gamma_3 z_2(\ln t) \qquad \textbf{(5.5)}$$

The terms relating to log-time $(\gamma_1 \ln t+\gamma_2 z_1(\ln t)+\gamma_3 z_2(\ln t))$ determine the shape of the function. This is adjusted by a constant amount (γ_{0j}) for each country *j* included in the IECV cycle (Table 5.5) giving a country-specific baseline hazard function. Table 5.6 and 5.7 give the estimated β-coefficients (log HRs) in each cycle for the following pre-specified predictors: age (in years), tumour type (invasive ductal, invasive lobular, colloid, tubular, medullary, papillary, other or unknown), histological grade (good, moderate, poor, unknown), nodal category (negative, 1-3, >3 to 10, >10), post-menopausal or age>65 (yes, no), pT score (pT1, pT2, pT3/pT4), adjuvant treatment (yes, no, unknown) and hormone receptor status (negatives or unknown, at least one positive).

The performance statistics available from the IECV approach were provided upon request by Hua,[176] and include Harrell's C-statistic, D-statistic and calibration slope. The C-statistic and D-statistic were defined in Chapter 1 (Section 1.6.3). The calibration slope for the flexible parametric survival model however is calculated by fitting a Royston-Parmar model in country *k* (excluded for validation),

$$\ln(H_{ik}(t))=\gamma_0+\gamma_1 \ln t+\gamma_2 z_1(\ln t)+\gamma_3 z_2(\ln t)+\widehat{\beta}\times LP_{ik}$$

where the spline function ($\gamma_0+\gamma_1 \ln t+\gamma_2 z_1(\ln t)+\gamma_3 z_2(\ln t)$) is taken from the model derived in the other countries and LP_{ik} is the linear predictor (linear combination of β's and covariate values from the derived model) calculated for patient *i* in validation country *k.* The estimate of β ($\widehat{\beta}$) in the calibration model above is the estimated calibration slope.

Table 5.5: Breast cancer model parameter estimates (and standard errors, σ_j) for country-specific baseline hazard, provided by Hua on request.

Country (k) – excluded for validation	Country-specific constant in baseline hazard γ_{0j}								Implementation strategy (i.e. value of γ_{0j} used for validation)	
	Netherlands	Ireland	Sweden	Slovenia	Austria	France	Switzerland	Denmark	Strategy 1: baseline hazard estimated in validation country	Strategy 2: average baseline hazard from derivation countries
Netherlands	-	-3.354 (0.418)	-4.332 (0.398)	-3.867 (0.411)	-3.017 (0.362)	-3.568 (0.328)	-3.548 (0.358)	**-2.847** **(0.366)**	-3.296 (0.029)	-3.492 (0.231)
Ireland	-3.207 (0.217)	-	**-4.260** **(0.246)**	-3.678 (0.333)	-2.892 (0.275)	-3.497 (0.218)	-3.475 (0.249)	-2.848 (0.230)	-3.244 (0.131)	-3.398 (0.188)
Sweden	-3.247 (0.222)	-3.305 (0.274)	-	-3.697 (0.336)	-2.926 (0.278)	-3.538 (0.223)	-3.552 (0.254)	**-2.929** **(0.235)**	-4.329 (0.083)	-3.304 (0.191)
Slovenia	-3.172 (0.217)	-3.221 (0.270)	-4.230 (0.247)	-	-2.860 (0.275)	-3.470 (0.218)	**-3.455** **(0.250)**	-2.824 (0.230)	-3.658 (0.258)	-3.321 (0.186)
Austria	-3.196 (0.222)	-3.247 (0.274)	-4.252 (0.251)	-3.690 (0.337)	-	-3.487 (0.223)	**-3.480** **(0.254)**	-2.847 (0.235)	-2.903 (0.186)	-3.441 (0.190)
France	-2.851 (0.248)	-2.865 (0.297)	-3.880 (0.275)	-3.354 (0.354)	-2.562 (0.297)	-	**-3.103** **(0.279)**	-2.478 (0.260)	-3.138 (0.060)	-3.001 (0.201)
Switzerland	-3.122 (0.218)	-3.152 (0.271)	-4.163 (0.248)	-3.621 (0.334)	**-2.854** **(0.276)**	-3.442 (0.218)	-	-2.758 (0.231)	-3.422 (0.115)	-3.292 (0.190)
Denmark	**-3.196** **(0.221)**	-3.239 (0.273)	-4.248 (0.250)	-3.655 (0.336)	-2.867 (0.279)	-3.495 (0.221)	-3.472 (0.254)	-	-2.855 (0.062)	-3.453 (0.192)

Note: Bold numbers represent the constant used for strategy 3: using the baseline hazard from the closest country based on proximity in the excluded country.

Table 5.6: Breast cancer model parameter estimates (and standard errors, σ_j) for categorical predictors, provided by Hua on request.

Country (k) – excluded for validation	β estimate (SE)												
	Tumour type*							Histological grade#			Lymph nodes positive†		
	Invasive lobular	Colloid	Tubular	Medullary	Papillary	Other	Unknown	Moderate	Poor	Unknown	Np1-3	Np3-10	Np10+
Netherlands	0.051 (0.153)	-0.661 (1.012)	-0.168 (0.713)	-0.897 (0.362)	-0.390 (0.729)	0.099 (0.222)	0.079 (0.146)	0.437 (0.221)	0.861 (0.226)	0.583 (0.268)	0.632 (0.104)	1.408 (0.108)	1.858 (0.135)
Ireland	-0.099 (0.095)	-0.307 (0.711)	-0.530 (0.502)	-0.774 (0.222)	-0.420 (0.506)	-0.076 (0.131)	0.023 (0.066)	0.426 (0.159)	0.786 (0.157)	0.608 (0.163)	0.715 (0.066)	1.241 (0.068)	1.716 (0.084)
Sweden	-0.122 (0.096)	-0.310 (0.711)	-0.539 (0.502)	-0.742 (0.222)	-0.419 (0.506)	-0.086 (0.131)	0.025 (0.066)	0.445 (0.162)	0.808 (0.160)	0.668 (0.167)	0.739 (0.068)	1.242 (0.070)	1.671 (0.088)
Slovenia	-0.101 (0.096)	-0.252 (0.711)	-0.529 (0.502)	-0.772 (0.222)	-0.662 (0.583)	-0.073 (0.131)	0.023 (0.066)	0.439 (0.161)	0.789 (0.158)	0.617 (0.165)	0.709 (0.065)	1.253 (0.067)	1.697 (0.084)
Austria	-0.119 (0.097)	-0.296 (0.711)	-0.496 (0.502)	-0.744 (0.227)	-0.257 (0.580)	-0.068 (0.133)	0.024 (0.066)	0.468 (0.170)	0.830 (0.166)	0.654 (0.172)	0.704 (0.065)	1.248 (0.067)	1.724 (0.084)
France	-0.121 (0.100)	0.260 (1.005)	-0.512 (0.502)	-0.763 (0.222)	-0.375 (0.507)	-0.113 (0.140)	0.010 (0.074)	0.301 (0.196)	0.652 (0.190)	0.469 (0.194)	0.803 (0.070)	1.324 (0.071)	1.786 (0.089)
Switzerland	-0.077 (0.099)	-0.285 (0.711)	-0.347 (0.502)	-0.733 (0.227)	-0.389 (0.507)	-0.043 (0.135)	0.023 (0.066)	0.423 (0.159)	0.748 (0.157)	0.552 (0.164)	0.689 (0.066)	1.231 (0.068)	1.683 (0.086)
Denmark	-0.100 (0.107)	-0.333 (0.711)	-1.060 (0.709)	-0.821 (0.262)	-0.422 (0.506)	-0.121 (0.140)	0.023 (0.066)	0.430 (0.159)	0.795 (0.157)	0.617 (0.163)	0.656 (0.068)	1.163 (0.070)	1.620 (0.087)

* Reference category for tumour type was ‘invasive ductal carcinoma’.

# Reference category for histological grade was `good’.

† Reference category for lymph nodes positive is ‘negative’.

Table 5.7: Breast cancer model parameter estimates (and standard errors, σ_j) for predictors modelled continuously, as provided by Hua on request.

Country (k) – excluded for validation	β estimate (SE)				
	Age	Menopausal status	pT score	Adjuvant treatment	Hormone receptor
Netherlands	0.015 (0.005)	-0.152 (0.114)	0.308 (0.060)	-0.113 (0.049)	-0.744 (0.077)
Ireland	0.012 (0.003)	-0.071 (0.077)	0.326 (0.038)	-0.122 (0.029)	-0.576 (0.053)
Sweden	0.011 (0.003)	-0.020 (0.078)	0.326 (0.038)	-0.106 (0.031)	-0.486 (0.055)
Slovenia	0.011 (0.003)	-0.052 (0.076)	0.328 (0.037)	-0.114 (0.029)	-0.573 (0.052)
Austria	0.011 (0.003)	-0.056 (0.076)	0.331 (0.037)	-0.113 (0.029)	-0.567 (0.052)
France	0.008 (0.003)	0.001 (0.082)	0.315 (0.039)	-0.144 (0.031)	-0.596 (0.056)
Switzerland	0.012 (0.003)	-0.059 (0.077)	0.317 (0.038)	-0.104 (0.029)	-0.572 (0.053)
Denmark	0.012 (0.003)	-0.079 (0.081)	0.340 (0.039)	-0.114 (0.030)	-0.538 (0.056)

Different strategies were implemented for the baseline hazard function to be used in the excluded country for validation. All three implementation strategies use the restricted cubic spline function estimated in the derivation set of countries, but use different constants (γ_{0j}). These implementation strategies were:

Strategy 1: Estimate a new baseline hazard for the validation country. This was done by re-estimating the constant part of the baseline cumulative hazard function (γ_{0j} in (5.5) now γ_{0k}) in the validation country.

Strategy 2: Use a weighted average of the baseline hazard functions from the countries included in the model derivation. This is done by taking a weighted average of the γ_{0j}'s.

Strategy 3: Use the baseline hazard function from one of the countries included in the derivation dataset, based on proximity to the validation country, i.e. γ_{0j} selected from the closest country to the validation country.

Estimates and standard errors of the performance statistics are given in Table 5.8 for each country, and within-study (within-country) correlations between the performance statistics are given in Table 5.9. Standard errors and correlations were obtained by bootstrapping with 1000 samples. There is a strong correlation between the C-statistic and D-statistic within each country (ranging from 0.61 to 0.84) as would be expected considering that they are both measures of discrimination. Within-study correlations between the D-statistic and calibration slope are also strong (and similar across implementation strategies) with values ranging from 0.57 to 0.83. The within-study correlations between the C-statistic and calibration slope are weaker, ranging from 0.16 to 0.47, but again are similar across implementation strategies.

Table 5.8: Performance statistics (and standard errors, σ_k) from the breast cancer model using IECV approach with three different strategies for the baseline hazard, as provided by Hua on request.

Country	C-statistic	D-statistic	Calibration slope		
			Strategy 1: Baseline hazard estimated in validation country	Strategy 2: Average baseline hazard	Strategy 3: Baseline hazard from closest country
Netherlands	0.697 (0.008)	0.493 (0.027)	0.977 (0.012)	1.049 (0.012)	0.805 (0.012)
Ireland	0.701 (0.036)	0.420 (0.117)	1.002 (0.057)	1.066 (0.057)	1.414 (0.056)
Sweden	0.715 (0.023)	0.106 (0.056)	1.026 (0.036)	0.578 (0.037)	0.405 (0.037)
Slovenia	0.735 (0.068)	0.326 (0.187)	0.991 (0.097)	0.870 (0.098)	0.919 (0.097)
Austria	0.666 (0.050)	0.238 (0.168)	0.946 (0.088)	1.168 (0.086)	1.184 (0.086)
France	0.682 (0.017)	0.182 (0.041)	0.969 (0.037)	0.896 (0.038)	0.951 (0.037)
Switzerland	0.781 (0.027)	0.280 (0.063)	1.054 (0.052)	0.996 (0.053)	0.794 (0.054)
Denmark	0.722 (0.016)	0.541 (0.058)	1.035 (0.030)	0.315 (0.029)	1.197 (0.030)

Table 5.9: Within-study correlations (ρ_{Wk}) between performance statistics for the breast cancer model, obtained through bootstrapping, provided by Hua on request.

Country	C-statistic & D-statistic	C-statistic & calibration slope (strategy 1)	C-statistic & calibration slope (strategy 2)	C-statistic & calibration slope (strategy 3)	D-statistic & calibration slope (strategy 1)	D-statistic & calibration slope (strategy 2)	D-statistic & calibration slope (strategy 3)
Netherlands	0.842	0.334	0.325	0.349	0.677	0.663	0.704
Ireland	0.827	0.303	0.307	0.324	0.662	0.664	0.668
Sweden	0.702	0.218	0.237	0.240	0.661	0.688	0.691
Slovenia	0.762	0.300	0.294	0.297	0.782	0.777	0.779
Austria	0.834	0.469	0.468	0.468	0.668	0.661	0.660
France	0.807	0.438	0.451	0.442	0.750	0.764	0.754
Switzerland	0.612	0.276	0.276	0.274	0.817	0.821	0.832
Denmark	0.812	0.199	0.166	0.183	0.625	0.573	0.599

Strategy 1: Baseline hazard estimated in validation country.
Strategy 2: Average baseline hazard from derived model.
Strategy 3: Baseline hazard from closest country included in model derivation.

5.4 Univariate meta-analysis of model performance

Methods for meta-analysis of validation performance statistics are now proposed. Meta-analysis was defined by Glass in 1976 as, ‘the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings’.[177] Meta-analysis is therefore the synthesis of several studies to summarise the effects (in this case, performance statistics) with a measure of the uncertainty for that pooled effect. By pooling the performance statistics obtained from each excluded study using the IECV approach, it is also possible to quantify the heterogeneity in model performance across studies.

In this section univariate meta-analysis is considered, which can be used for each performance statistic separately. There are two approaches commonly used for univariate meta-analysis, namely fixed-effect and random-effects which are detailed below.

5.4.1 Fixed-effect meta-analysis of model performance

A fixed-effect meta-analysis assumes that the true value of the performance statistic being pooled is the same in all studies. In other words, that there is one 'true' underlying value for the performance statistic that all of the studies are trying to estimate. The model can be written as

$$Y_k = \mu + e_k \qquad \textbf{(5.6)}$$

where Y_k is the performance statistic estimate (e.g. calibration-in-the-large, C-statistic, etc.) in study k, μ is the pooled performance statistic and the within-study errors are assumed to be normally distributed, $e_k \sim \mathrm{N}(0, \sigma_k^2)$, where σ_k^2 is the variance of Y_k and is assumed known. The fixed-effect model can also be written as

$$Y_k \sim \mathrm{N}(\mu, \sigma_k^2). \qquad \textbf{(5.7)}$$

The fixed-effect approach assumes that any differences in the performance statistic across studies are simply due to sampling error,[178] and the pooled estimate $\hat{\mu}$ is therefore interpreted as the best estimate of the underlying performance statistic. Maximum likelihood estimation of (5.7) shows that $\hat{\mu}$ is simply a weighted average, with the study weights inversely proportional to the study's variance (σ_k^2).[179] Thus the estimation method is also known as the inverse-variance method, with the weight for study k calculated as

$$w_k = \frac{1}{\sigma_k^2}.$$

As mentioned, the fixed-effect approach assumes that the underlying model performance is the same in each study excluded for validation. Using fixed-effect meta-analysis is very similar to the approach taken by Royston et al. (described in Section 5.2.2) which takes a weighted average of the performance statistics to obtain the overall summary of performance.[69] However, it seems reasonable that a prediction model could validate better (i.e. have better performance) in some studies than in others. This is referred to as *between-study heterogeneity* and therefore measures are needed to quantify and account for heterogeneity in performance.

The I^2 statistic

I^2 is a measure of the percentage of total variation across studies that is due to between-study heterogeneity rather than sampling error.[180-182]

$$I^2 = \left(\frac{Q - df}{Q}\right) \text{ x } 100\% \qquad \textbf{(5.8)}$$

where df is K–1, K is the number of studies and

$$Q = \sum_{k=1}^{K} w_k (Y_k - \hat{\mu})^2.$$

The w_k are the weights and $\hat{\mu}$ is the pooled result from a fixed-effect analysis. The Q-statistic is also a test for heterogeneity (compared to χ^2_{K-1}); however the I^2 statistic is generally preferred for assessing heterogeneity between studies as it quantifies the impact of heterogeneity,[182] although it is also open to criticism.[183]

5.4.2 Random-effects meta-analysis of model performance

A random-effects meta-analysis does not assume that there is one ‘true’ underlying effect, but rather that there is a distribution of ‘true’ study effects. This seems more reasonable for pooling performance statistics from different studies, as the prediction model would be expected to perform better in some studies than others due to differences in the patient populations. The random-effects meta-analysis model can be written as

$$Y_k = \mu + b_k + e_k \tag{5.9}$$

where b_k is the between-study error, assuming b_k~N(0,τ^2) and e_k is the within-study error, assuming e_k~N(0,σ_k^2) with σ_k^2 assumed known. The random-effects approach can equivalently be written as

$$Y_k \sim \mathrm{N}\left(\mu, \tau^2 + \sigma_k^2\right). \tag{5.10}$$

The approach allows the risk prediction model's performance to differ in each validation study and thus allows the between-study variance (τ^2) in performance to be estimated. The pooled performance statistic estimate $\hat{\mu}$ in a random-effects model should not be interpreted in the same way as for a fixed-effect model, but rather as the estimated *average* of the performance statistic; that is, the estimated average of the distribution of true validation performance across studies.[184]

There are several methods that can be used for estimation in a random-effects meta-analysis; these include the DerSimonian and Laird method,[185] maximum likelihood estimation (MLE) and restricted maximum likelihood (REML). Using REML, the likelihood is modified slightly to account for the data being used to estimate both the underlying mean and variance.[179] In this chapter, REML is used for estimation of all random-effects analysis models. This produces estimates for τ and μ. Rucker et al. suggest that τ^2 is preferred to the I^2 statistic.[183] However, τ^2 is also harder to interpret as it is a variance and therefore the unit of measurement needs to be considered. Prediction intervals (discussed below) are easier to interpret and increase in width as τ^2 increases (although they also incorporate within-study variance so do not only reflect between-study variability).

5.4.3 Confidence and prediction intervals for model performance

Following estimation of a meta-analysis model, a **100(1-*α*)% confidence interval** for the mean pooled performance statistic μ can be estimated as

$$\hat{\mu} \pm z_{\frac{\alpha}{2}} \widehat{\mathrm{SE}}(\hat{\mu}) \qquad \textbf{(5.11)}$$

using a normal approximation ($z_{\frac{\alpha}{2}}$) to determine the confidence limits. For a 95% confidence interval, $z_{0.025}$ = 1.96, and this is multiplied by the estimated standard error of the pooled performance statistic estimate $\left(\widehat{\mathrm{SE}}(\hat{\mu})=\sqrt{\widehat{\mathrm{Var}}(\hat{\mu})}\right)$. White proposed that $\widehat{\mathrm{SE}}(\hat{\mu})$ is inflated to account for the uncertainty in the estimated between-study variance, and this is implemented in this chapter.[186]

A **100(1-α)% prediction interval** for model performance in a new study (i.e. a study done subsequent to the meta-analysis) can also be derived following a random-effects meta-analysis using the formula

$$\hat{\mu} \pm t_{\alpha, K-2}\sqrt{\hat{\tau}^2 + \widehat{\mathrm{Var}}(\hat{\mu})} \qquad \textbf{(5.12)}$$

where $t_{\alpha, K-2}$ is the $100\left(1-\frac{\alpha}{2}\right)\%$ percentile of the t-distribution for K-2 d.f.[187] This is an approximate prediction interval that uses the t-distribution to account for $\hat{\tau}^2$ being an estimate itself and therefore having uncertainty that is otherwise not accounted for.

The 95% prediction interval gives the lower and upper bound for the predicted true performance of the model in a new external study that is similar to one of those included in the meta-analysis. Using 95% prediction intervals for model performance is a novel way to view the generalisability of a model across multiple populations or settings, and has recently been suggested for C-statistics by van Klaveren et al.[62] This is what external validation aims to achieve but is usually limited to validation in one or two external datasets. In using both the IECV approach and meta-analysis, the model is first developed using all available data and then essentially validated multiple times in studies not used to derive the model estimates. Using meta-analysis to produce 95% prediction intervals provides a new way to think about validation of a model in external populations as it provides a distribution for the performance statistic. A narrow prediction interval is reassuring that the model should perform well (for the statistic being considered, either a measure of discrimination or calibration) in a new population, which is ideal when the model is intended to be used in new populations. Narrow prediction intervals will relate to small $\hat{\tau}^2$, such that the estimated between-study heterogeneity is close to zero. The 95% prediction intervals can also be used to compare implementation strategies, where the best strategy would give an average pooled effect close to the ideal for that performance statistic (e.g. calibration slope=1) and have the narrowest prediction interval (smallest between-study heterogeneity).

5.5 Multivariate meta-analysis of model performance

Section 5.4 introduced univariate meta-analysis methods for combining each type of performance statistic separately. However, meta-analysis methods have also been extended for joint synthesis of multiple correlated statistics at the same time; so-called ‘multivariate’ meta-analysis methods. Jackson et al. discussed the advantages and limitations of using multivariate meta-analysis.[173] The advantages arise from using the correlations to gain additional information; this leads to better statistical properties and makes use of the relationship between the multiple performance statistics being analysed, whereas a univariate analysis assumes that there is no correlation between performance statistics. Multivariate analysis may not always be possible as within-study correlations of the multiple statistics may be difficult to obtain.[174] However, because IPD are available for the development and validation of the model using the IECV approach, correlations can be obtained through bootstrapping.[188] The within-study correlations for each study in the two examples were estimated using bootstrapping and presented earlier (Table 5.4 and Table 5.9), and are often large which emphasises why they should be considered.

The pooled estimates resulting from a multivariate random-effects meta-analysis are often similar to the univariate results; however when some studies are missing one or more of the statistics at random, the multivariate analysis ‘borrows strength’ and can produce smaller standard errors for the pooled estimates compared to those from univariate analyses.[174] Furthermore, even without missing data, the multivariate approach allows joint inferences across the multiple measures of interest, for which accounting for correlation is essential. For further details on the general benefits of multivariate meta-analysis, see Jackson et al. and Riley et al.[173,174]

As multiple performance statistics (such as C-statistic, D-statistic and calibration slope) for a prediction model may be correlated, and will ultimately be considered jointly to make overall judgements about model performance, it seems a natural extension to consider multivariate meta-analysis in this setting. The univariate methods from Section 5.4 are therefore now extended for a bivariate meta-analysis below. Trivariate or greater extensions follow naturally, and are briefly considered in Section 5.6.3.

5.5.1 Bivariate fixed-effect model

Two performance statistics from each study *k* (e.g. a measure of discrimination and calibration such as the C-statistic and calibration slope) are now assumed to follow a bivariate normal distribution and the bivariate fixed-effect model can be written as

$$\begin{pmatrix} Y_{k1} \\ Y_{k2} \end{pmatrix} \sim \text{MVN}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{k1}^2 & \rho_{Wk}\sigma_{k1}\sigma_{k2} \\ \rho_{Wk}\sigma_{k1}\sigma_{k2} & \sigma_{k2}^2 \end{pmatrix} \right) \qquad \textbf{(5.13)}$$

where $\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ is the vector of means for the performance statistics, $\begin{pmatrix} \sigma_{k1}^2 & \rho_{Wk}\sigma_{k1}\sigma_{k2} \\ \rho_{Wk}\sigma_{k1}\sigma_{k2} & \sigma_{k2}^2 \end{pmatrix}$ is the within-study covariance matrix assumed known, σ_{kq} is the within-study standard deviation for performance statistic q = 1, 2 and ρ_{Wk} is the within-study correlation between the two performance statistics in study *k*.

5.5.2 Bivariate random-effects model

The bivariate random-effects model can be written as

$$\begin{pmatrix} Y_{k1} \\ Y_{k2} \end{pmatrix} \sim \text{MVN}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{k1}^2 + \tau_1^2 & \rho_{Wk}\sigma_{k1}\sigma_{k2} + \rho_B\tau_1\tau_2 \\ \rho_{Wk}\sigma_{k1}\sigma_{k2} + \rho_B\tau_1\tau_2 & \sigma_{k2}^2 + \tau_2^2 \end{pmatrix}\right) \qquad \textbf{(5.14)}$$

where τ_q is the between-study standard deviation for performance statistic q = 1, 2, and ρ_B is the between-study correlation, with other terms as defined earlier.

Within-study and between-study correlations

In the bivariate random-effects model, the overall correlation between two performance statistics consists of within-study correlation and between-study correlation.[189] The within-study correlation is the correlation between the estimates of the performance statistics in a study, and is caused by the same patients' data informing the estimates of each performance statistic.[175] For example in the DVT dataset (described in Section 5.3.1), there is perfect correlation between the estimates of calibration-in-the-large and log(E/O) within each study, because the same data are used to estimate two statistics, related by definition. Calibration-in-the-large is a difference and log(E/O) is a (log) ratio of the observed and expected number with DVT (Table 5.4).

The between-study correlation describes the relationship between the true performance statistics across studies and is affected by differences in populations across the studies (e.g. treatments, characteristics, countries etc.).[175] For example, it may be that a study population with a higher than average calibration slope is also likely to have a higher than average C-statistic.

When conducting a multivariate meta-analysis, within-study correlations are used as input data along with the estimates and variances of the performance statistics from each study. The pooled values and between-study covariance matrix of model (5.14) are then estimated using for example REML.[186] In other applications of meta-analysis (other than for use with the IECV approach), it is not always possible to obtain estimates of within-study correlations as authors do not usually publish these in the original study publications. Riley et al. evaluated the effect of not having within-study correlations and proposed a method for use when the within-study correlations are unknown.[175,189] However this is unnecessary here, as with IPD, ρ_{Wk} can be estimated in each study using bootstrapping,[188] and this was used in the two examples as previously described.

In this chapter, in keeping with the univariate meta-analyses, all multivariate meta-analyses are performed using REML via the '*mvmeta*' package in Stata,[186] and the estimated standard errors of the pooled estimates were inflated to account for uncertainty in the estimated variance-covariance matrix. In the article, White states, 'The standard error provided for an REML analysis allows for uncertainty in estimating Σ by inverting the second derivative matrix of the restricted likelihood (1). This is not the standard approach (Kenward and Roger 1997), and its properties require further investigation.'[186]

5.5.3 Confidence and prediction ellipses

Following a multivariate meta-analysis, marginal 95% confidence and prediction intervals can be calculated for each performance statistic using (5.11) and (5.12). It is also possible to plot regions for joint confidence and prediction between two performance statistics following a bivariate analysis.[190,191] The formulae for the confidence ellipse are

$$\begin{aligned} \mu_1 &= \hat{\mu}_1 + s_1 * \mathit{Constant} * \cos(t) \\ \mu_2 &= \hat{\mu}_2 + s_2 * \mathit{Constant} * \cos(t + \arccos(r)) \end{aligned} \qquad \textbf{(5.15)}$$

where s_1 and s_2 are the standard error estimates for $\hat{\mu}_1$ and $\hat{\mu}_2$, and r is the estimated correlation between $\hat{\mu}_1$ and $\hat{\mu}_2$, all coming from the variance-covariance matrix of the estimates, $\widehat{Var}\begin{pmatrix}\hat{\mu}_1 \\ \hat{\mu}_2\end{pmatrix} = \begin{bmatrix} s_1^2 & rs_1s_2 \\ rs_1s_2 & s_2^2 \end{bmatrix}$. '*Constant*' is the boundary constant (see below) and t is varied between 0 and 2π to create the boundary of the ellipse.

The formulae for the prediction ellipse are

$$\begin{aligned} \mu_1 &= \hat{\mu}_1 + \sqrt{s_1^2 + \hat{\tau}_1^2} * \mathit{Constant} * \cos(t) \\ \mu_2 &= \hat{\mu}_2 + \sqrt{s_2^2 + \hat{\tau}_2^2} * \mathit{Constant} * \cos\left(t + \arccos(r_{pred})\right) \end{aligned} \qquad \textbf{(5.16)}$$

where the variance-covariance matrix used for estimates of the standard errors and correlation come from $\widehat{Var}\begin{pmatrix}\hat{\mu}_1 \\ \hat{\mu}_2\end{pmatrix}$ + the estimated between-study variance-covariance matrix, resulting in $\begin{bmatrix} s_1^2 + \hat{\tau}_1^2 & rs_1s_2 + \hat{\rho}_B\hat{\tau}_1\hat{\tau}_2 \\ rs_1s_2 + \hat{\rho}_B\hat{\tau}_1\hat{\tau}_2 & s_2^2 + \hat{\tau}_2^2 \end{bmatrix}$.[191] The correlation used for the prediction ellipse is now $r_{pred} = \frac{rs_1s_2 + \hat{\rho}_B\hat{\tau}_1\hat{\tau}_2}{\sqrt{s_1^2 + \hat{\tau}_1^2} \times \sqrt{s_2^2 + \hat{\tau}_2^2}}$.

The boundary constant ('*Constant*') gives the 100(1–α)% confidence or prediction region. There are different constants that can be used to produce the region, such as using the F-distribution ($F_{2,K-2}$) or χ_2^2 distribution. As K (number of studies) gets larger, the distribution

2*($F_{2,K-2}$) approaches the χ^2_2 distribution.[192] In this chapter, the F-distribution was used as K is small in both datasets and therefore an F-distribution that accounts for the number of studies seems more appropriate, giving a larger confidence or prediction region than the χ^2_2 distribution.

The 95% confidence ellipse is the region in which the pair of **mean** performance statistics is expected to lie. The 95% prediction ellipse contains the region in which the pair of **true** performance statistics is likely to fall in a new study. So for example, if the two performance statistics analysed in a multivariate meta-analysis are a discrimination statistic and a calibration statistic, the 95% prediction region gives a 95% prediction region for the joint discrimination and calibration performance in a new study population.

As the univariate predictive distribution is approximately a univariate t-distribution with K–2 d.f. to account for uncertainty in the estimate of τ,[187] it seems sensible to approximate the bivariate predictive distribution by a bivariate t-distribution with K–2 d.f. Using this, it is also possible to calculate the predicted probability of a model performing to specified criteria for the performance statistics following a multivariate meta-analysis. For example, the predicted probability of the calibration slope lying between 0.9 and 1.1, and C-statistic ≥ 0.7 in a new population. This is done by drawing a large sample from a bivariate t-distribution (with K-2 d.f.) using the mean vector and variance-covariance matrix from Equation (5.14). This can be performed using the '*randmvt*' module in SAS or the '*mnormt*' package in R software (see Appendix D1 for example R code). The joint probability is the proportion of paired samples that meet the criteria for both of the performance statistics.

5.6 Results

The univariate and multivariate meta-analysis methods were applied to the two datasets introduced in Section 5.3, and the key findings are now described.

5.6.1 Dataset one: Deep vein thrombosis

Recall that the DVT model was developed to predict the risk of having DVT in patients that were suspected of having DVT. There were 12 studies available in which a logistic model was developed for the binary outcome of having DVT or not. The model included three predictors: sex, surgery and calf difference. Validation of the model was done using the IECV approach with each of the following applied: a stratified intercept (for implementation strategies 1 and 3) or random-effect intercept (for implementation strategy 2). The performance statistics calculated were calibration-in-the-large, calibration slope, log(E/O) and the C-statistic (reported in Table 5.3).

Univariate results

Firstly, the DVT model performance statistics obtained using the IECV approach were pooled using univariate meta-analysis. Only results from random-effects models are reported here, as there was evidence of heterogeneity for most of the performance statistics, as determined by the estimated between-study heterogeneity $\hat{\tau}$ as well as looking at I^2 (Table 5.10).

The results of the random-effects model using REML as the estimation method are displayed in Table 5.10, for each implementation strategy (different choice of intercept) used to validate the derived prediction model in the excluded study.

Table 5.10: Univariate random-effects meta-analysis results for the DVT model performance statistics using different implementation strategies for the model intercept used in the validation dataset.

Strategy	Performance statistic	Pooled estimate (SE)	95% confidence interval	95% prediction interval	I^2 % (approx. 95% CI for I^2)	τ estimate (approx. 95% CI for τ)
Strategy 1: Intercept estimated in validation study	Calibration-in-the-large	-0.125 (0.027)	-0.178 to -0.072	-0.185 to -0.065	0	<0.001 (0.000 to <0.001)
	Calibration slope	0.972 (0.068)	0.839 to 1.104	0.551 to 1.392	63 (18 to 82)	0.176 (0.064 to 0.288)
	Log(Expected/Observed)	0.084 (0.019)	0.046 to 0.121	0.041 to 0.126	0	<0.001 (0.000 to <0.001)
	C-statistic	0.688 (0.010)	0.668 to 0.707	0.634 to 0.741	45 (3 to 73)	0.022 (0.004 to 0.040)
Strategy 2: Average intercept taken from derived random-intercept model	Calibration-in-the-large	-0.004 (0.158)	-0.314 to 0.306	-1.244 to 1.235	97 (91 to 98)	0.533 (0.302 to 0.765)
	Calibration slope	0.973 (0.068)	0.839 to 1.107	0.550 to 1.397	62 (18 to 82)	0.177 (0.065 to 0.289)
	Log(Expected/Observed)	0.022 (0.116)	-0.206 to 0.250	-0.888 to 0.932	97 (91 to 98)	0.392 (0.221 to 0.562)
	C-statistic	0.687 (0.010)	0.668 to 0.707	0.634 to 0.741	45 (2 to 73)	0.022 (0.004 to 0.039)
Strategy 3: Intercept from a study in derivation set with a similar prevalence	Calibration-in-the-large	0.046 (0.085)	-0.121 to 0.213	-0.584 to 0.677	89 (66 to 95)	0.270 (0.135 to 0.404)
	Calibration slope	0.970 (0.068)	0.837 to 1.103	0.550 to 1.390	62 (17 to 81)	0.176 (0.064 to 0.288)
	Log(Expected/Observed)	-0.028 (0.062)	-0.150 to 0.094	-0.485 to 0.428	89 (65 to 95)	0.195 (0.095 to 0.296)
	C-statistic	0.687 (0.010)	0.668 to 0.707	0.634 to 0.740	45 (3 to 73)	0.022 (0.004 to 0.040)

Calibration of the model

The average calibration slope is very similar for all three implementation strategies and close to one. This shows that regardless of the intercept used, on average the predicted risks of DVT are similar in the validation data compared to the derivation data across studies. The average calibration slope for the DVT model (using all three strategies) is around 0.97 (95% CI: 0.84 to 1.10), which is close to the ideal value of one (see Figure 5.1 for forest plot). The amount of between-study heterogeneity is also very similar when comparing the three implementation strategies ($\hat{\tau}$≈0.176). The 95% prediction interval for the calibration slope using strategy 2 (average intercept) is 0.55 to 1.40, and is very similar for the other two implementation strategies. This prediction interval is quite wide, due to the large heterogeneity of calibration performance in the different studies. A calibration slope of 0.6 or 1.4 would mean the model could severely over/under predict in a new population or setting. Therefore, although the model performs well on average across studies, there is concerning heterogeneity that may cause poor calibration in individual populations. This illustrates how the average performance is an incomplete picture; calibration slope is good on average, but could be poor in particular populations.

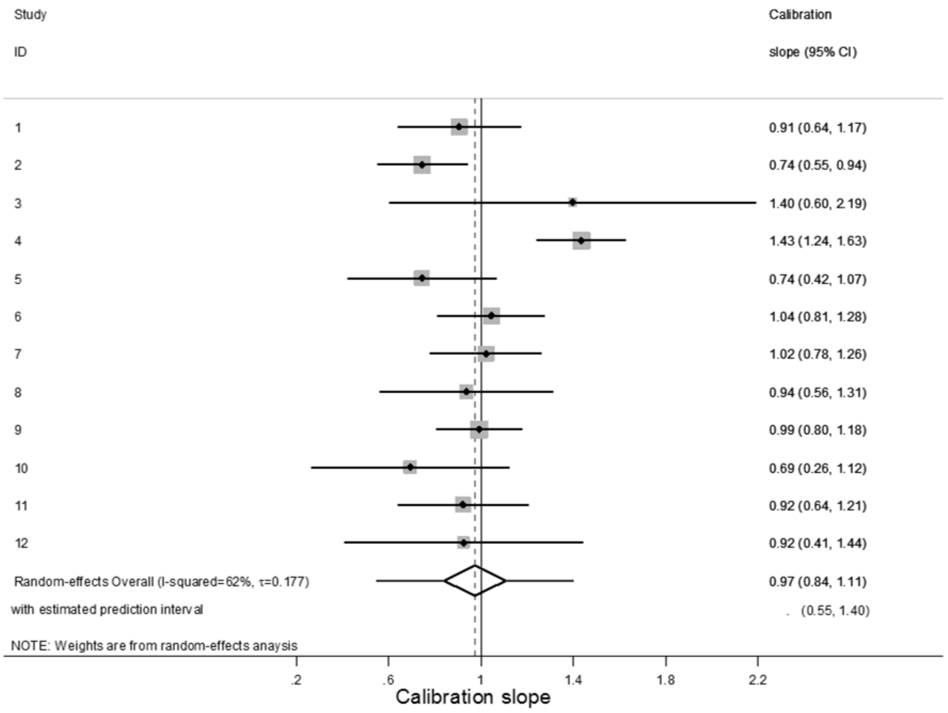


Figure 5.1: Forest plot from univariate random-effects meta-analysis of calibration slope (using strategy 2: average intercept) from the DVT model.

Another important measure is calibration-in-the-large. The pooled calibration-in-the-large is worst when the intercept is estimated in the validation study ($\hat{\mu}$= -0.125, 95% CI: -0.178 to -0.072), however there is almost no heterogeneity ($\hat{\tau}$<0.001). The average calibration-in-the-large is better when either the average intercept (strategy 2) or the intercept from a study with a similar prevalence (strategy 3) are used (-0.004 and 0.046 respectively), but there is a lot of between-study heterogeneity ($\hat{\tau}$=0.533 and 0.270 respectively), seen in Figure 5.2. It would be very unlikely to have data to estimate an intercept when applying the model to a new population or setting (strategy 1), therefore considering the other two strategies, using an intercept from a study with a similar prevalence would be better than using the average intercept as there is less heterogeneity. However, heterogeneity is still large, suggesting calibration-in-the-large may be poor in some study populations; this is shown by a 95% prediction interval of -0.58 to 0.68 (Table 5.10). The log(E/O) follows a similar pattern (when looking at heterogeneity and comparing implementation strategies) to the calibration-in-the-large as these two validation measures have a very strong negative correlation (Table 5.10). The 95% prediction interval for expected/observed for implementation strategy 1 suggests the overall agreement is likely to be reasonable in new populations (1.04 to 1.13), with the number of DVT cases over-predicted by between 4% and 13%. However, the 95% prediction interval is unsatisfactory for the other strategies; for example, it is 0.41 to 2.54 for strategy 2 indicating the number of predicted DVT cases in a new population could range from 59% too few up to 154% too many.

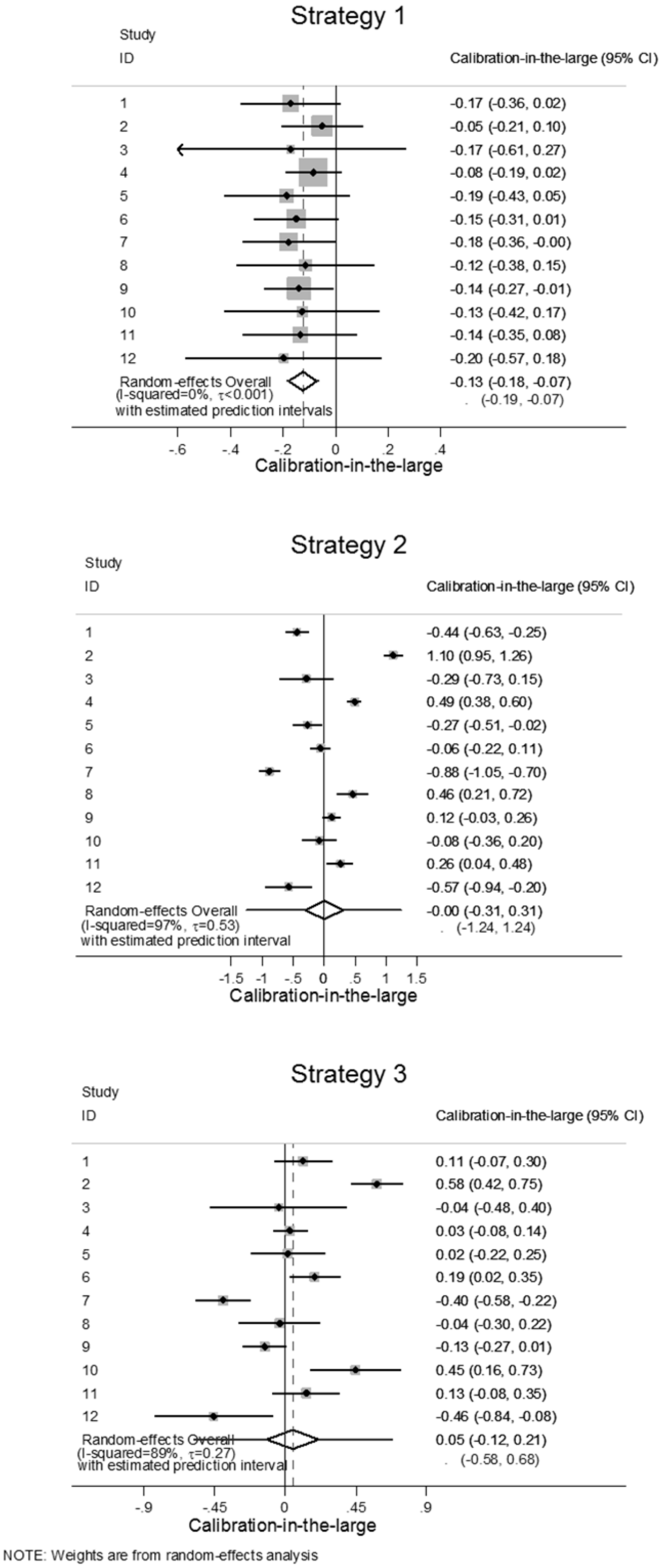


Figure 5.2: Forest plots from univariate random-effects meta-analysis of calibration-in-the-large from the DVT model, for each implementation strategy.

Discrimination of the model

The average C-statistic, used to measure discrimination of the model, is consistent regardless of which intercept is used (Table 5.10). This is seen in Figure 5.3 ($\hat{\mu}$=0.69, 95% CI: 0.67 to 0.71) and there is a small amount of between-study heterogeneity ($\hat{\tau}$=0.022). This suggests that the ability of the model to discriminate between individuals of very high and low risk is unaffected by the intercept, which is to be expected as the discrimination would only be affected by changes in the linear predictor rather than the intercept. The 95% prediction interval for the C-statistic ranges from 0.63 to 0.74, revealing consistent moderate discrimination across studies.

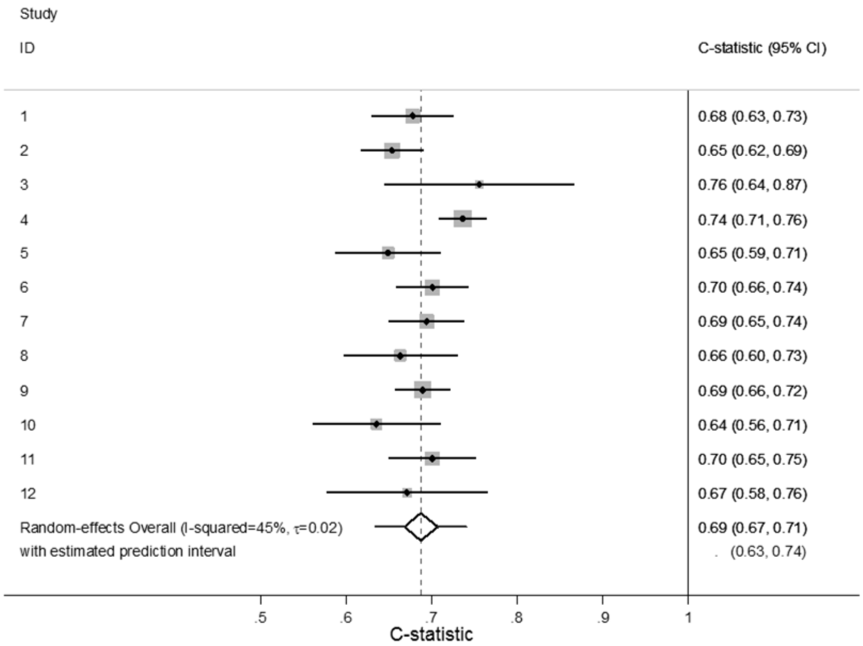


Figure 5.3: Forest plot from univariate random-effects meta-analysis of the C-statistic when the average intercept is used from the DVT model (strategy 2).

Bivariate random-effects meta-analysis

The overall correlation between performance statistics across studies is seen in Figure 5.4. In particular, there is a strong positive correlation between the calibration slope and C-statistic and these performance statistics were also robust against changes in which intercept was used. The correlation can be used in the analysis if the two performance statistics are analysed together in a bivariate random-effects meta-analysis (Equation (5.14)).

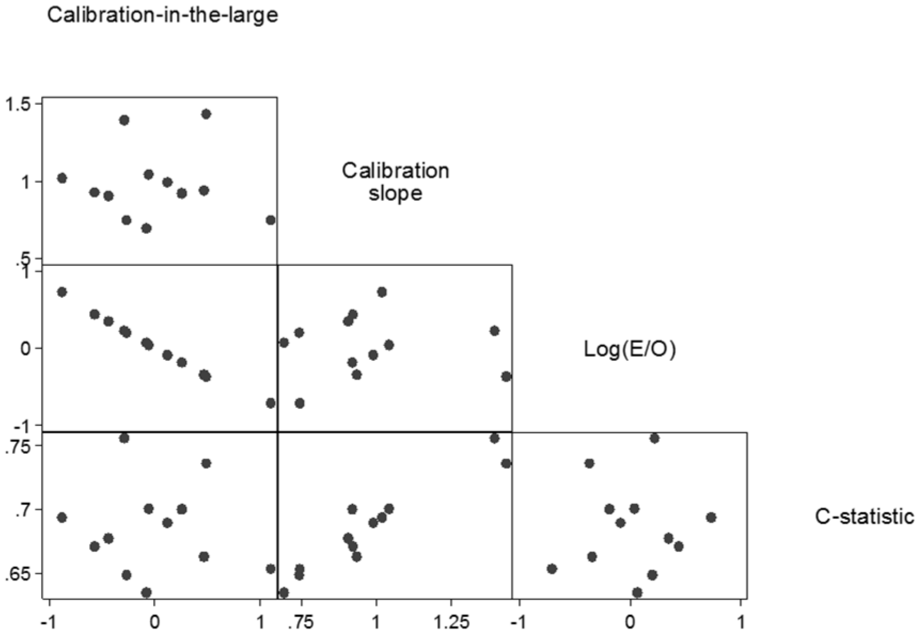


Figure 5.4: Scatterplot matrix for performance statistics from the DVT model, when using the average intercept in the validation data (strategy 2).

The results of the bivariate random-effects meta-analysis of the calibration slope and C-statistic are given in Table 5.11. The pooled estimates from the bivariate meta-analysis are very similar to those from the univariate analyses in this case. The 95% confidence intervals for both calibration slope and C-statistic are slightly narrower than the 95% confidence intervals obtained from the univariate analyses, potentially due to the utilisation of correlation gaining slightly more statistical precision.[174] The marginal 95% prediction intervals are also slightly narrower as the estimate of τ is slightly smaller than in the univariate analyses. The between-study correlation was estimated as +1 in all three analyses (different implementation strategies), which indicates that the between-study correlation is poorly estimated.[193] A between-study correlation is estimated iteratively but constrained between -1 and +1; therefore a value of ±1 often indicates that it has reached the end of its parameter space. This is likely to occur when the within-study variance is very large relative to the between-study variance (unless the number of studies is very large). Riley et al. suggest that pooled estimates remain unbiased in this situation,[193] however it may have more of an impact on joint predictive inferences.

Table 5.11: Bivariate random-effects meta-analysis of calibration slope and C-statistic from the DVT model, using different implementation strategies for the intercept.

	Performance statistic	Pooled estimate (SE)	95% confidence interval for pooled estimate	Marginal 95% prediction interval for pooled estimate	I^2 % (approx. 95% CI for I^2)	τ estimate (approx. 95% CI for τ)	Between-study correlation $\widehat{\rho}_B$ (95% confidence interval)
Strategy 1: Intercept estimated in validation study	Calibration slope	0.975 (0.062)	0.854 to 1.097	0.595 to 1.355	58 (16 to 78)	0.159 (0.059 to 0.259)	1.000 (cannot estimate CI)
	C-statistic	0.687 (0.009)	0.670 to 0.704	0.644 to 0.730	34 (2 to 63)	0.017 (0.004 to 0.031)	
Strategy 2: Average intercept taken from derived random-intercept model	Calibration slope	0.975 (0.063)	0.851 to 1.099	0.588 to 1.362	58 (16 to 78)	0.162 (0.061 to 0.262)	1.000 (cannot estimate CI)
	C-statistic	0.686 (0.009)	0.669 to 0.704	0.642 to 0.730	35 (3 to 63)	0.018 (0.004 to 0.031)	
Strategy 3: Intercept from a study included in derivation set with a similar prevalence	Calibration slope	0.972 (0.063)	0.849 to 1.094	0.589 to 1.355	57 (15 to 78)	0.160 (0.059 to 0.261)	1.000 (cannot estimate CI)
	C-statistic	0.686 (0.009)	0.669 to 0.703	0.642 to 0.729	34 (2 to 62)	0.017 (0.004 to 0.031)	

Joint confidence and prediction regions

An advantage of performing a bivariate meta-analysis is that the results can then be used to produce *joint* confidence and prediction regions for both performance statistics included in the analysis. Figure 5.5 shows the 95% confidence ellipse and 95% prediction ellipse for the calibration slope and C-statistic. The correlation (between estimates or between-study) affects how round or narrow the ellipses are. In the DVT data, both the 95% confidence and prediction ellipses are very narrow, resulting in several of the studies falling outside of the region. However, the correlation between the pooled performance statistic estimates was high (r=0.95 from $\widehat{Var}\begin{pmatrix}\hat{\mu}_1\\ \hat{\mu}_2\end{pmatrix}$) and the between-study correlation ($\hat{\rho}_B$=1.00) was poorly estimated (discussed in the previous section).

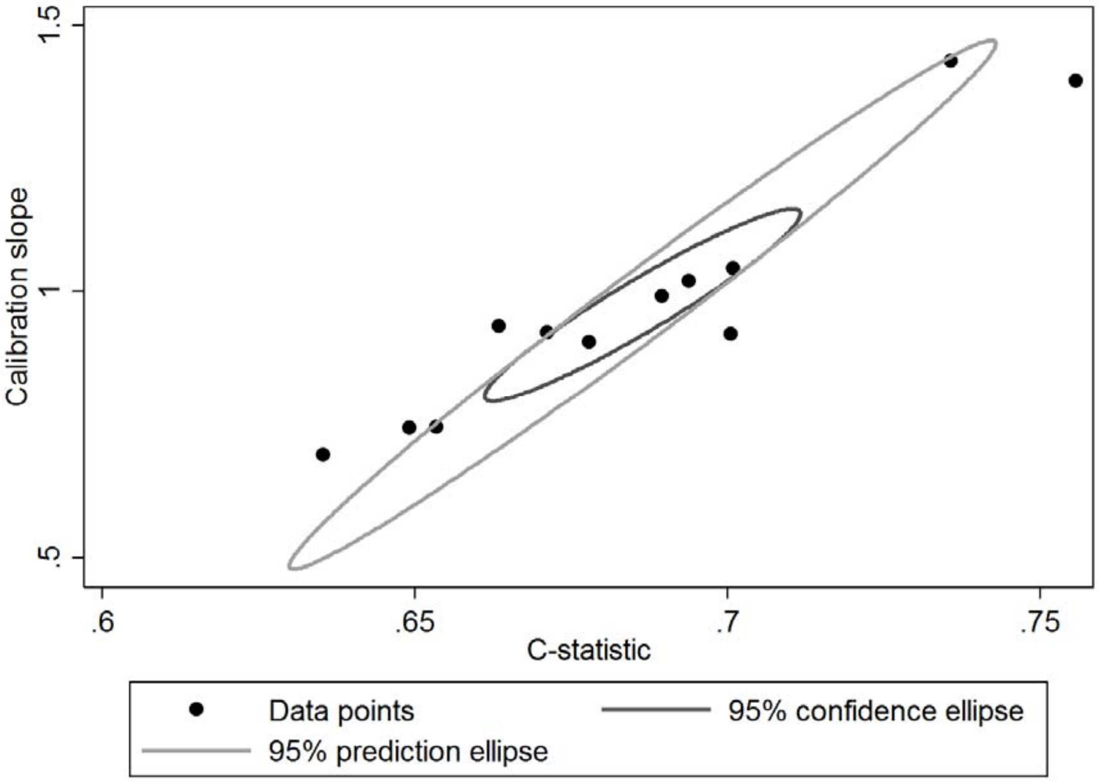


Figure 5.5: Joint 95% confidence ellipse and prediction ellipse for calibration slope and C-statistic from the DVT model using the F-distribution (strategy 2: average intercept used).

One way in which to quantify how the model and implementation strategy will perform in a new population or setting is to calculate the predicted probability of meeting criteria for the calibration slope and C-statistic (assessing model calibration and discrimination jointly), where the criteria for model performance should ideally be pre-specified. The predicted probability of the model performing with a *C-statistic*≥0.7 and 0.9≤*calibration slope*≥1.1 is only 2.4% when using implementation strategy 2: average intercept (see Table 5.12). This is because the prediction is for a higher C-statistic than was seen on average from the IECV approach (pooled C-statistic=0.69). If the criteria for model discrimination is relaxed to *C-statistic*≥0.65, then the predicted probability of jointly meeting the C-statistic criteria and a calibration slope of between 0.9 and 1.1 is 42% for the average intercept approach of strategy 2. If the calibration slope criteria is also relaxed to between 0.8 and 1.2 then there is a predicted probability of 72% for meeting the performance criteria in a new population. A perfectly calibrated model would have a calibration slope=1, therefore specifying that the calibration slope should be within 0.8 and 1.2 corresponds to the model under- or over-predicting by 20%.

Table 5.12: Joint predicted probabilities for performance of the DVT model (calibration slope and C-statistic) in a new population, using results from the bivariate analysis.

		Joint predicted probability		
Calibration slope bounds	Minimum C-statistic	Strategy 1: Intercept estimated in external validation study	Strategy 2: Average intercept taken from derived random-intercept model	Strategy 3: Intercept from a study included in derivation with a similar prevalence
0.9 and 1.1	0.70	0.027	0.024	0.024
0.8 and 1.2	0.70	0.145	0.140	0.138
0.9 and 1.1	0.65	0.426	0.420	0.420
0.8 and 1.2	0.65	0.728	0.720	0.723

Recommendations

Based on Table 5.12, all model implementation strategies have similar predicted performance. Given the low predicted probability for meeting even the relaxed C-statistic criteria, it is unlikely that any of the models are suitable for use for any of the implementations strategies. Note however, that joint inferences may be affected by the poorly estimated between-study correlation of +1, and so the aforementioned probabilities should be viewed as approximate. For example, if the between-study correlation was rather +0.5, then the predicted probability of the model performing with a *C-statistic*≥0.7 and 0.9≤*calibration slope*≥1.1 would be 9.6% rather than 2.4% for strategy 2. However, it is clear that the relatively low C-statistic indicates that the model is unlikely to be suitable, and further predictors are needed to improve discrimination and reduce heterogeneity in calibration.

5.6.2 Dataset two: Breast cancer

Recall that the breast cancer model was developed to predict the risk of death in patients with breast cancer. There were studies from eight countries that were included in the development and validation of the prognostic model using the IECV approach. Predictors included age, tumour type, histological grade, nodal category, post-menopausal or age>65, pT score, adjuvant treatment and hormone receptor status. Model performance statistics of interest are Harrell's C-statistic, the D-statistic and calibration slope. These were presented in Table 5.8.

Univariate results

The pooled estimates of the performance statistics from a random-effects meta-analysis are given in Table 5.13, along with 95% confidence intervals, 95% prediction intervals and estimates of the between-study standard deviation τ. In this dataset, 'study' relates to country as individual studies were combined by country for model development and validation. The performance statistics include the C-statistic and D-statistic as measures of model discrimination, and the calibration slope using the three implementation strategies for the baseline hazard in the validation dataset: (1) baseline hazard estimated in external validation country, (2) weighted average of baseline hazards from derivation countries, and (3) baseline hazard from country closest in proximity that was included in the model derivation. The discrimination (C-statistic and D-statistic) would not be affected by using different implementation strategies.

Table 5.13: Univariate random-effects meta-analysis results for the breast cancer model performance statistics using different implementation strategies for the baseline hazard used in the validation dataset.

Strategy	Performance statistic	Pooled estimate (SE)	95% confidence interval	95% prediction interval	I^2 % (approx. 95% CI for I^2)	τ estimate (approx. 95% CI for τ)
All strategies*	C-statistic	0.711 (0.012)	0.688 to 0.733	0.653 to 0.768	51 (0 to 85)	0.021 (0.000 to 0.047)
All strategies*	D-statistic	0.326 (0.063)	0.203 to 0.450	-0.082 to 0.734	88 (52 to 95)	0.154 (0.060 to 0.249)
Strategy 1: Baseline hazard estimated in validation country	Calibration slope	0.998 (0.016)	0.966 to 1.029	0.938 to 1.058	22 (0 to 68)	0.019 (0.000 to 0.051)
Strategy 2: Weighted average baseline hazard from derived model	Calibration slope	0.992 (0.080)	0.836 to 1.149	0.424 to 1.561	97 (89 to 99)	0.218 (0.099 to 0.337)
Strategy 3: Baseline hazard from country included in model derivation, closest in proximity	Calibration slope	0.957 (0.111)	0.740 to 1.173	0.156 to 1.757	99 (94 to 99)	0.308 (0.143 to 0.473)

* Results are the same for each of the three implementation strategies using different baseline hazards for validation of the model.

Calibration of the model

The calibration slope is closest to the ideal value of one when the baseline hazard is estimated in the validation data (average calibration slope=0.998). This has a narrow 95% confidence interval, and also narrow 95% prediction interval due to very little between-study heterogeneity ($\hat{\tau}$=0.02). So the calibration slope of the model in a new population is predicted (with 95% probability) to lie between 0.938 and 1.058. Estimating the baseline hazard in the country for which it is intended to be used is ideal but unlikely to happen in reality, considering the model could be implemented in countries where samples of individuals have not been collected that could be used for estimation. Therefore it is important to consider the other two implementation strategies as they are more plausible to use in practice. The calibration slope suggests slight miscalibration and large between-study heterogeneity when the baseline hazard from a nearby country is used (average calibration slope=0.957, $\hat{\tau}$=0.31). The model using the average baseline hazard calibrates better than the previously mentioned strategy, with an average calibration slope of 0.992. However, even though average calibration is good, there is large between-study heterogeneity ($\hat{\tau}$=0.22) suggesting that the model calibrates better in some countries than others. A 95% prediction interval for the calibration slope using strategy 2 is very wide (95% PI: 0.424 to 1.561) revealing potential for extremely poor calibration at the upper and lower limits, although the width of the interval also reflects uncertainty in $\hat{\tau}$.

Discrimination of the model

The pooled C-statistic is 0.711, which suggests that on average the model discriminates between individuals of high and low risk of breast cancer moderately well and similar to other prognostic models such as the pancreatic cancer model developed and validated in Chapter 4. There is also very little between-study heterogeneity in the C-statistic ($\hat{\tau}$=0.02), so the model discrimination is quite consistent across the different countries.

The D-statistic can be interpreted as a log hazard ratio when patients are separated into two equal groups based on their prognostic index. Therefore values further away from zero show greater separation between the two prognostic groups. The pooled D-statistic for the breast cancer model is 0.326, which equates to a hazard ratio of 1.385 (95% CI: 1.225 to 1.568) between two equal sized prognostic groups. There is some between-study heterogeneity in the D-statistic with $\hat{\tau}$=0.15.

Correlation between performance statistics

The overall correlations between performance statistics from the breast cancer model are displayed in a scatterplot matrix (Figure 5.6). Using the scatterplots as a visual guide, the performance statistics for the survival model are not as strongly correlated as the C-statistic and calibration slope in the logistic model for the DVT model (Section 5.6.1). However there are still some strong overall correlations between the D-statistic and calibration slope when the average baseline hazard is used and when the baseline hazard from the closest country is used. There is also a very strong overall correlation between the C-statistic and the calibration slope using the baseline hazard estimated in the external validation country. The relationship between the C-statistic and calibration slope using either the average or closest country baseline hazards are weaker. There also appears only weak correlation between the C-statistic and the D-statistic. These overall correlations can be quite different from the within-study correlations (Table 5.9).

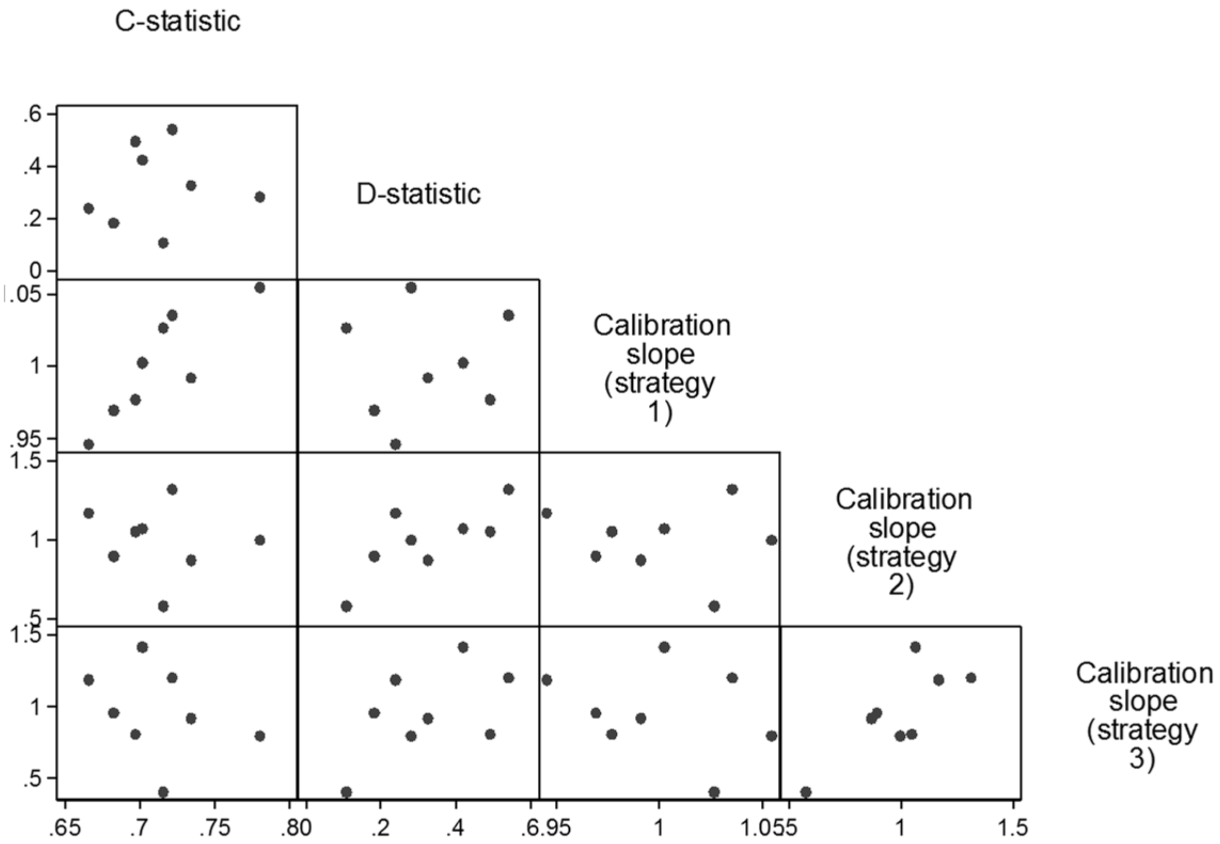


Figure 5.6: Scatterplot matrix for performance statistics of the breast cancer model.

Bivariate random-effects meta-analysis

Bivariate random-effects meta-analysis was performed for all pairs of C-statistic, D-statistic and calibration slope (using one of the implementation strategies at a time). The pooled estimates (Table 5.14) are very similar to those previously discussed from the univariate analysis, with little gain in precision as the 95% confidence and 95% prediction intervals are not much narrower than those estimated in the univariate analysis. However, the utilisation of correlation does enable more appropriate joint inferences in the multivariate meta-analysis (see below).

Table 5.14: Bivariate random-effects meta-analysis of performance statistics from the breast cancer model.

Strategy	Performance statistic	Pooled estimate (SE)	Marginal 95% confidence interval for pooled estimate	Marginal 95% prediction interval for pooled estimate	I^2 % (approx. 95% CI for I^2)	τ estimate (approx. 95% CI for τ)	Between-study correlation $\widehat{\rho}_B$ (95% confidence interval)
All strategies*	C-statistic	0.710 (0.011)	0.689 to 0.732	0.663 to 0.766	49 (0 to 85)	0.020 (0.000 to 0.047)	-0.131 (-0.878 to 0.801)
	D-statistic	0.327 (0.060)	0.210 to 0.444	-0.068 to 0.721	87 (52 to 95)	0.150 (0.060 to 0.240)	
Strategy 1: Baseline hazard estimated in validation country	C-statistic	0.712 (0.011)	0.690 to 0.734	0.662 to 0.769	51 (0 to 84)	0.020 (0.000 to 0.045)	1.000 (cannot estimate CI)
	Calibration slope	1.002 (0.017)	0.968 to 1.035	0.921 to 1.083	39 (0 to 76)	0.029 (0.000 to 0.0.063)	
Strategy 2: Weighted average baseline hazard from derived model	C-statistic	0.710 (0.012)	0.688 to 0.733	0.660 to 0.769	52 (0 to 85)	0.021 (0.000 to 0.048)	0.013 (-0.806 to 0.815)
	Calibration slope	0.994 (0.081)	0.836 to 1.153	0.416 to 1.572	98 (89 to 99)	0.222 (0.101 to 0.343)	
Strategy 3: Baseline hazard from country included in derivation, closest in proximity	C-statistic	0.709 (0.012)	0.686 to 0.733	0.655 to 0.771	55 (0 to 86)	0.022 (0.000 to 0.049)	-0.296 (-0.901 to 0.700)
	Calibration slope	0.957 (0.112)	0.738 to 1.177	0.145 to 1.769	99 (94 to 99)	0.312 (0.145 to 0.480)	
Strategy 1: Baseline hazard estimated in validation country	D-statistic	0.320 (0.058)	0.207 to 0.433	0.027 to 0.697	86 (48 to 94)	0.143 (0.055 to 0.231)	-0.741 (-1.000 to 1.000)
	Calibration slope	0.999 (0.015)	0.969 to 1.028	0.949 to 1.048	13 (0 to 67)	0.014 (0.000 to 0.050)	
Strategy 2: Weighted average baseline hazard from derived model	D-statistic	0.335 (0.062)	0.213 to 0.456	-0.044 to 0.746	88 (55 to 95)	0.156 (0.063 to 0.249)	0.866 (0.348 to 0.979)
	Calibration slope	0.995 (0.081)	0.835 to 1.154	0.412 to 1.577	98 (89 to 99)	0.224 (0.102 to 0.346)	
Strategy 3: Baseline hazard from country included in derivation, closest in proximity	D-statistic	0.337 (0.064)	0.212 to 0.462	-0.042 to 0.750	88 (54 to 95)	0.156 (0.062 to 0.251)	0.592 (-0.198 to 0.916)
	Calibration slope	0.962 (0.111)	0.744 to 1.180	0.154 to 1.770	99 (94 to 99)	0.311 (0.144 to 0.478)	

* Results are the same for each of the three implementation strategies using different baseline hazards for validation of the model.

Joint predictions

The joint 95% confidence and prediction ellipses for variables analysed together in the bivariate analyses are displayed in Figure 5.7. The 95% confidence and 95% prediction regions for the C-statistic and D-statistic are quite round as the correlation between these performance statistics is close to zero ($\hat{\rho}_B$ = -0.13). The confidence ellipse is the 95% confidence region for the mean of both performance statistics. The prediction ellipse is the 95% prediction region for the value of both performance statistics in a new population. All of the pairwise plots, apart from C-statistic versus calibration slope (using strategy 1: baseline hazard estimated in the validation dataset), contain all of the data points within the 95% prediction region. The prediction ellipse for the C-statistic and calibration slope (using implementation strategy 1) does not contain all of the data points and this is likely to be due to the poorly estimated between-study correlation (reaching the value +1).

The joint predicted probabilities for model performance in a new population are given in Table 5.15. The D-statistic of 0.3 corresponds to a hazard ratio of 1.35 if patients were separated into two equal groups based on their linear predictor. The predicted probability of the C-statistic≥0.7 and calibration slope between 0.9 and 1.1 is 0.15 for strategy 3 (baseline hazard from nearest country) and 0.21 for strategy 2 (average baseline hazard). However, the predicted probability is much higher at 0.67 for strategy 1. Considering whether the predicted probability may be higher because of the poorly estimated between-study correlation, the probability was predicted again assuming a between-study correlation of +0.5. The predicted probability remained at 0.67 suggesting that strategy 1 does perform better than the other two strategies and that this predicted probability is robust against poor estimation of the between-study correlation. The joint probabilities for D-statistic and calibration slope show similar results in that implementation strategy 1 performs better than strategies 2 or 3.

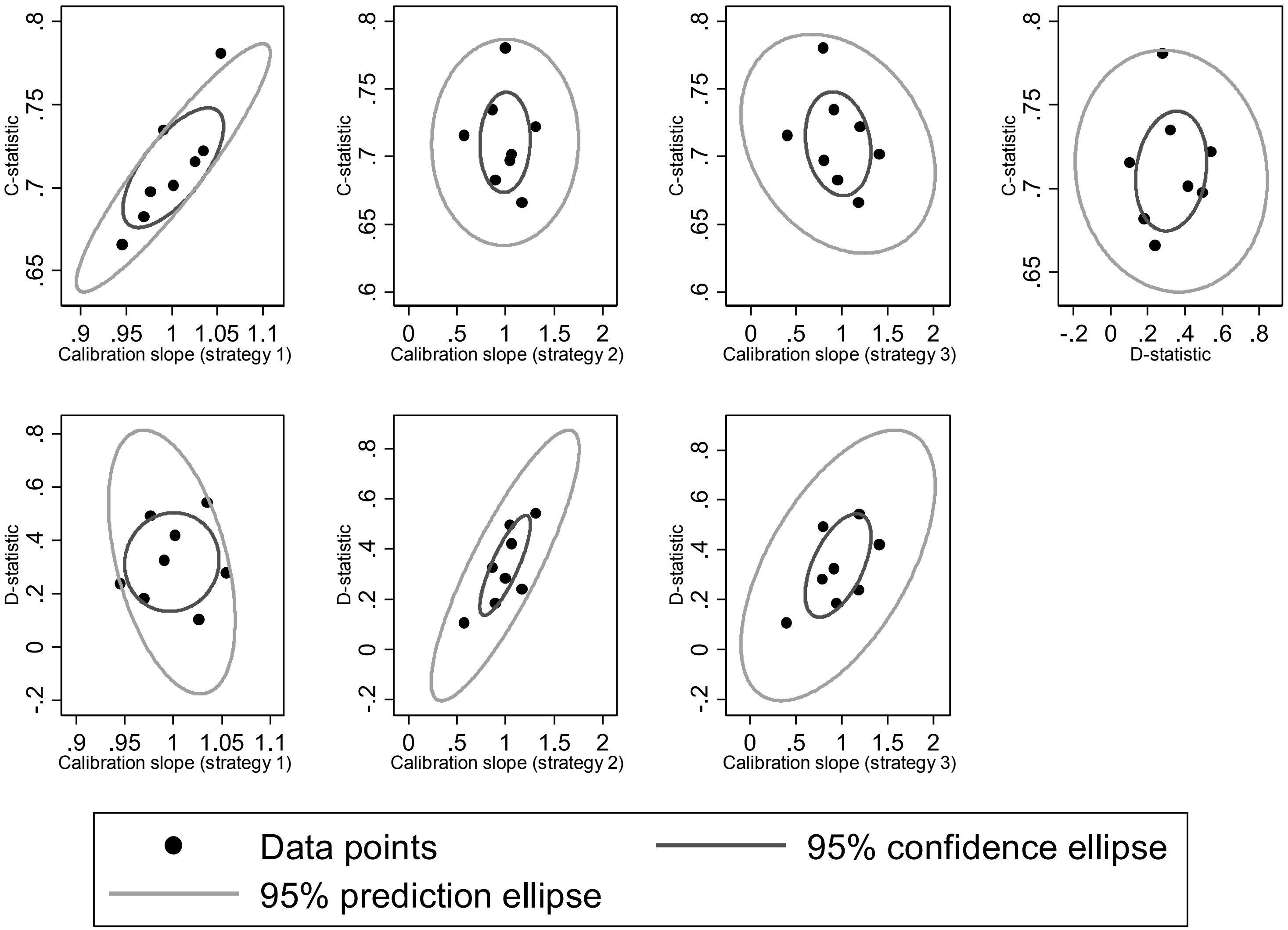


Figure 5.7: Joint 95% confidence and prediction ellipses for performance statistics from the breast cancer data.

Table 5.15: Joint predicted probabilities for specified values of performance statistics based on the breast cancer model in a new population.

Implementation strategy	Minimum C-statistic	Minimum D-statistic	Calibration slope bounds	Joint probability of meeting performance criteria
All strategies	0.7	0.3	-	0.368
Strategy 1: Baseline hazard estimated in validation country	0.7	-	0.9 and 1.1	0.674
Strategy 2: Weighted average baseline hazard from derived model	0.7	-	0.9 and 1.1	0.212
Strategy 3: Baseline hazard from country included in model derivation, closest in proximity	0.7	-	0.9 and 1.1	0.146
Strategy 1: Baseline hazard estimated in validation country	-	0.3	0.9 and 1.1	0.547
Strategy 2: Weighted average baseline hazard from derived model	-	0.3	0.9 and 1.1	0.205
Strategy 3: Baseline hazard from country included in model derivation, closest in proximity	-	0.3	0.9 and 1.1	0.144

Recommendations

These predicted probabilities confirm that using the baseline hazard of a country nearby (strategy 3) does not work well for the breast cancer model and that using the average baseline hazard (strategy 2) is only slightly better (although marginal with joint probabilities of meeting the performance criteria of 0.15 for strategy 3 compared to 0.21 for strategy 2). Both of these strategies perform poorly compared to estimating (recalibrating) the baseline hazard in the new data (strategy 1).

The source of heterogeneity may be identified in some cases. One of the countries (Sweden) was identified as being quite different to the other countries as calibration of the model was particularly poor for this country (calibration slope=0.578) when using implementation strategy 2 (weighted average baseline hazard). Therefore a sensitivity analysis was performed in which Sweden was removed and the IECV approach was repeated without it (see Appendix D2). The average calibration slope was closer to one (average=0.999) and the 95% prediction interval was narrower. The C-statistic did not change much but the joint probability for 'good' performance defined as C-statistic≥0.7 and calibration slope between 0.9 and 1.1 increased from 0.21 to 0.32. Of course this makes sense, as reducing the heterogeneity shrinks prediction intervals and increases predicted probabilities for 'good' model performance; however, it is important to understand why this heterogeneity exists and what makes Sweden different from the other countries. For example, it may be due to a very different diagnostic system, leading to breast cancer diagnosis at a substantially earlier (or later) stage of disease compared to other countries.

Overall, the breast cancer prediction model performs reasonably well in terms of discrimination (average C-statistic=0.71, 95% PI: 0.66 to 0.77) and calibration performance, especially when the baseline hazard is recalibrated by estimating it in the country in which the model will be applied (strategy 1 average calibration=1.002, 95% PI: 0.92 to 1.08). This implementation strategy resulted in the least heterogeneity of model performance across validation countries. If it is not possible to recalibrate the model to the population in which it will be applied, using a weighted average baseline hazard is a good second choice, especially if Sweden is considered too different from the other countries and excluded from the analysis. Using the baseline hazard from the closest country proved to be a poor implementation strategy and had the most heterogeneous model performance, therefore this strategy is not recommended.

5.6.3 Extension to trivariate random-effects meta-analysis

The between-study correlation was poorly estimated (+1) between the C-statistic and calibration slope in both the DVT and breast cancer datasets. This was the case for all three implementation strategies for the DVT data and implementation strategy 1 (baseline hazard estimated in validation study) for the breast cancer data. Trivariate random-effects meta-analysis was therefore considered to try and improve the estimation of between-study correlation between the C-statistic and calibration slope by incorporating further information from a third performance measure.

The pooled results, confidence intervals and prediction intervals did not change much when a trivariate random-effects meta-analysis was performed using the DVT data with the addition of either log(E/O) or calibration-in-the-large (see Appendix D3). Between-study correlations for the C-statistic and calibration slope could still not be estimated for implementation strategy 1 (intercept estimated in validation study) and were estimated between +0.993 and +0.999 for the other analyses (different implementation strategies and whether log(E/O) or calibration-in-the-large were included in the trivariate meta-analysis). This suggests that there is potentially a very strong between-study correlation between the

C-statistic and calibration slope in these studies and that the predicted probabilities for model performance (to specified criteria) would only change slightly if the trivariate meta-analysis results were used instead of the bivariate results to derive the predicted probabilities or prediction ellipses. For example, the predicted probability for strategy 2 (average intercept) meeting the criteria of C-statistic≥0.7 and calibration slope between 0.9 and 1.1 changed from 0.024 to 0.037.

In the breast cancer data, a trivariate random-effects meta-analysis of the C-statistic, D-statistic and calibration slope (for implementation strategy 1) gave similar pooled results for the C-statistic and calibration slope compared to the bivariate meta-analysis. The between-study correlation was estimated as +0.986, giving credence to a value close to +1 being genuinely plausible (i.e. not due to unreliable estimation as originally feared). As the between-study correlation and other estimates are similar to those from the bivariate analysis, the predicted probabilities and prediction ellipses from this trivariate meta-analysis are also very similar to those previously discussed for the bivariate analyses.

5.7 Discussion

Before a risk prediction model can be deemed clinically useful, it is important to assess how the model performs in external data that was not used for model development. This external validation is performed to evaluate the generalizability and transportability of the model to a range of plausibly related populations.[12] With this in mind, an important consideration is how the developed model should be implemented in different populations. Debray et al. published a framework that includes different implementation strategies for clinical prediction models.[149] These implementation strategies also require validation, and the availability of IPD from multiple studies is an ideal opportunity to do this. Selecting an intercept to use in a specific population is a way of recalibrating or updating the model to improve performance within that population. Several articles have mentioned that if a model does not perform particularly well when validated, the model should be updated (either by recalibrating or including additional predictors) rather than discarded and a brand new model developed.[3,59,194] Moons et al. also say that ideally IPD from the new situation are available.[194] This chapter supports this in that performance was best when the model was recalibrated using data from that study but also considered alternatives suggested by Debray et al. for when IPD are not available.[149]

This chapter has extended the IECV approach by proposing novel univariate and multivariate meta-analysis methods to combine performance statistics across the excluded studies, and thereby summarise overall model performance, and even predict model performance in a new population for a given implementation strategy. Royston et al. suggested a weighted average of the individual values to give an overall statistic (Equation (5.1)).[69] This overall performance statistic is similar to the pooled performance statistic from a fixed-effect meta-analysis. However, it seems preferable to assume that the performance of the model in each study could vary and therefore the random-effects meta-analysis approaches proposed are more appropriate. The advantage of using the random-effects meta-analysis approach is that the between-study heterogeneity is estimated, which allows inferences regarding model performance in new populations (rather than average performance across all populations). Further, the chapter has suggested multivariate meta-analysis approaches, to jointly synthesise multiple performance statistics and make joint inferences about their likely value in new populations. The work presented in this chapter has also been accepted for publication in the Journal of Clinical Epidemiology.

5.7.1 Related research in this field

The development and validation of prediction models using IPD from multiple studies is receiving growing attention. Ahmed et al. conducted a review of articles developing and validating risk prediction models using IPD meta-analysis.[195] The authors identify heterogeneity in baseline risk across studies as a methodological challenge and recommend the use of Debray et al.'s framework of using IECV and selecting an appropriate intercept term relating to the population in which the model would be applied.[149]

Other studies have used meta-analysis to combine performance statistics, such as Pennells et al., where the authors assessed the predictive performance of models developed from IPD.[61] However, they did not use the IECV approach but instead considered developing the model in each study, pooling βs using meta-analysis as well as the performance statistics. The authors comment on comparing discrimination across studies, and that the study-specific discrimination depends on the distributions of risk predictors within each study. Therefore, they caution against interpreting discrimination across studies when there are large differences in distributions of the risk predictors.[61] A further point for debate is the weighting scheme used in the meta-analysis. The weighting of each study in the proposed random-effects models of this chapter is an inverse-variance. Pennells et al. recommend weighting by the number of events when the aim is to produce a weighted average of discrimination performance; however, they agree that inverse-variance weighting using random-effects models is more appropriate when the predicted performance in new populations is of interest.[61] The random-effects approach is also more naturally extended to multivariate meta-analysis, as demonstrated here.

Another study has looked at discrimination within clusters when a model is developed using clustered data, using meta-analysis to combine the cluster-specific discrimination statistics.[62] The authors recommended using random-effects meta-analysis for the pooling of cluster-specific C-statistics across clusters, to assess the variability in discrimination across clusters. They also checked standardised residuals for the C-statistic on the probability and log-odds scale to check for normality when pooling C-statistics in the meta-analysis and found that the residuals fitted the standard normal distribution better on the probability scale than on the log-odds scale.[62]

5.7.2 Summary of key findings from this chapter

The DVT and breast cancer models were used to illustrate the meta-analysis methods, but their usefulness warrants discussion. Different implementation strategies were assessed by using different intercepts or baseline hazards for validation in the excluded study. Strategy 2 differed in the two examples, as one used prevalence to select an intercept and the other used the closest country. This chapter uses the performance measures of models developed by different authors, and evaluates the strategies that these authors selected. The measures of discrimination were not dependent on the intercept used, unlike some of the calibration measures.

The DVT model considered in this chapter was developed by Debray et al. to illustrate methodology rather than being intended for clinical use.[149] However, estimating the intercept within the population was the best strategy as it had the best average performance and least heterogeneity across studies when jointly considering the C-statistic and calibration slope. Calibration slope did not change depending on the implementation strategy used but calibration-in-the-large was least heterogeneous across studies when the intercept was estimated in the excluded study (strategy 1). If it is not possible to recalibrate the model to the intended population, strategy 3 is recommended (selecting intercept from study with similar prevalence level) as although C-statistics and calibration slopes were very similar for strategies 2 and 3, there was less heterogeneity in calibration-in-the-large across studies for strategy 3.

For the breast cancer model, the best model performance was achieved if the baseline hazard could be estimated for the population in which the model will be used. Using the IECV approach, the calibration slope was least heterogeneous when the baseline hazard was estimated in the excluded country (strategy 1). However, this approach of estimating the intercept or baseline hazard in the population for which it is intended is unlikely to be useful in practice. Using an intercept from a study with a similar prevalence level was a better approach than using the average intercept in the DVT model. In the breast cancer data, using the baseline hazard from the closest country did not work as well and therefore using the average intercept gave better validation performance. Even so, these implementation strategies had large between-study heterogeneity, indicating that model performance may be poor in many of the individual populations that the model could be used in, even if average performance across all of the excluded studies was good. A sensitivity analysis also showed that Sweden was quite different to other countries included and including it resulting in heterogeneity in model performance. This indicates that the model may also perform poorly in other countries outside of those included and would therefore need to be tested before the model is used in that country.

Perhaps the most important and novel element of this chapter is the multivariate meta-analysis proposal for jointly summarising multiple performance statistics. The bivariate (or trivariate) meta-analysis was demonstrated for combinations of performance statistics. There was very little difference between the pooled results of the multivariate meta-analysis and the univariate meta-analysis. The benefit of multivariate meta-analysis is more evident in situations where there are missing values for a study.[174] However, jointly synthesising multiple performance statistics has the advantage of estimating the correlation between performance statistics which enables joint predictions to be made for combinations of performance statistics such as a measure of calibration and discrimination. This can be helpful when comparing competing models or, as in this chapter, when comparing different implementation strategies by ranking them by their probability of meeting specified criteria for what is deemed to be a 'good' prediction model. Arbitrary values were selected for chosen statistical criteria in this chapter to illustrate the methods; however, criteria for a 'good' prediction model may include other statistical or clinical criteria.

Another key finding is that there may be issues with the estimation of between-study correlation in multivariate meta-analysis when the within-study variances are large relative to the between-study variances, as discussed by Riley et al.[193] Importantly, the authors noted in their simulation study that the pooled results were not biased, even when the between-study correlation was estimated as -1 or +1. However, this chapter showed that poor estimation of the between-study correlation can affect the joint confidence and prediction ellipses as they become too narrow, which was evident from the prediction ellipse of the C-statistic and calibration slope of the DVT model. Using additional information from a third performance statistic in a trivariate meta-analysis, the between-study correlations would often converge close to +1.

The multivariate meta-analysis methods encourage the joint synthesis of discrimination and calibration performance of a prediction model, as in the literature, model calibration is often ignored.[54] This could be due to ongoing debate on how best to measure calibration and when is a model well calibrated. One issue discussed by Vach is that a calibration slope of one does not necessarily indicate good calibration of a model,[196] and this warrants further consideration in further research.

5.7.3 Limitations and further work

In this chapter, (multivariate) normality was assumed for the distribution of true model performance across studies. This is a common assumption in meta-analysis but departures from this assumption would have implications on inferences made from the 95% prediction intervals and joint prediction regions. More flexible distributions have been proposed to allow for skewness and could be considered in further work.[197] Normality has also been assumed for the within-study sampling distribution of each performance statistic. However, expected/observed was log transformed as it was considered more likely to be skewed. Van Klaveren et al.[62] compared different scales for the C-statistic and found that residuals looked better for meta-analysis on the probability scale than on the log-odds scale, however this assumption will be interrogated further in Chapter 6.

Further development of the multivariate approach could involve the use of multivariate meta-regression for pooling performance statistics. Meta-regression combines the meta-analysis of the performance statistics with modelling study-level covariates to explain some of the heterogeneity in model performance across studies.[198] This is different to including the additional covariates in the developed prediction model (which would account for differences between patients) as the covariates are included at the study level (which would account for between-study heterogeneity). A sensitivity analysis of the breast cancer model revealed one study as different to the others as the model performed far worse in Sweden than other countries. Meta-regression may be one way of identifying potential differences in studies/countries to help explain between-study heterogeneity in performance statistics and identify where the model will perform better and where it could perform worse. Potential problems with the meta-regression approach can include obtaining study-level covariate information for each study included in the meta-analysis and also having a large enough number of studies to estimate the regression coefficients precisely and draw robust conclusions from the meta-regression.[199]

5.8 Conclusions

IECV is a useful approach for the development and validation of a prediction model when multiple studies are available as it maximises the amount of data used towards the development of the model as well as externally validating the model multiple times. This chapter has proposed using (multivariate) random-effects meta-analysis to pool performance statistics from IECV, in order to summarise average performance and the heterogeneity in performance across studies, which is important when considering the application of the model in specific populations. The real examples showed how prediction intervals from a random-effects meta-analysis help reveal how the model is expected to perform in new populations, and how a multivariate approach allows the joint synthesis of multiple correlated performance statistics. This is especially useful for making joint predictions of how a model will perform in a new population for combinations of two performance statistics (e.g. discrimination and calibration), and thereby helps identify the best reliable implementation strategy when using the model.

A key assumption made in this chapter when using random-effects meta-analysis is that the ‘true’ performance statistic (e.g. C-statistic) is normally distributed. However, there is no conclusive evidence to suggest that this is reasonable for the C-statistic,[62] or any other performance statistic considered in this chapter. The next chapter aims to evaluate this normality assumption for some of the performance statistics in different settings through simulation.

CHAPTER 6: A SIMULATION STUDY TO EVALUATE THE DISTRIBUTIONS OF PERFORMANCE STATISTICS

6.1 Introduction and aims

In Chapter 5, (multivariate) random-effects meta-analysis was proposed for synthesising estimates of prediction model performance from multiple validation studies. This provides a pooled estimate of the average performance of the model as well as an estimate of the heterogeneity of performance across the validation studies. Recall that a univariate random-effects meta-analysis of the C-statistic, as proposed in the previous chapter and in related work by van Klaveren et al.[62] can be specified as follows:

$$\begin{gathered} \widehat{C}_i \sim \text{Normal}\left(C_i, S_i^2\right) \\ C_i \sim \text{Normal}(\mu, \tau^2) \end{gathered} \qquad \textbf{(6.1)}$$

The C-statistic estimated in each study i ($\widehat{C}_i$) is assumed to be normally distributed with mean C_i and variance S_i^2. This assumption of normality in the sampling estimate is a reasonable one if the sample size and number of events are large, as one can appeal to the Central Limit Theorem. The C_i's (i.e. the true C-statistics in the studies) are also assumed to follow a normal distribution at the study level. This assumption is common in the meta-analysis field,[184,187] and is especially important when making predictions for the potential C-statistic in a new study similar to one of those included in the meta-analysis. Therefore, if assuming the true C-statistics follow a normal distribution is inappropriate, this would make model (6.1) and subsequent prediction intervals invalid. Thus further research of this issue is needed to ascertain whether the normality assumption is reasonable and, if not, whether a different scale may be preferable. Similarly, the between-study distribution of other performance statistics considered in Chapter 5 (such as the calibration slope) warrants investigation.

Van Klaveren et al.[62] also comment on this normality assumption for meta-analysis of performance statistics from prediction models, and attempted to look at the distribution of true C-statistics in clustered data. Assuming the cluster-specific C-statistics are normally distributed (both in terms of sampling estimates within-studies and true values across studies), the authors calculated standardised residuals and assessed the normality assumption using a normal probability plot and Shapiro-Wilk test applied to the residuals. They conclude that the C-statistic scale is appropriate, and indeed preferable to the logit scale. They state:

> *When assessing the discriminative ability of risk models used to support decisions at cluster level we recommend meta-analysis of cluster-specific c-indexes. Particularly, random effects meta-analysis should be considered.*
>
> *The normality assumptions for derivation of a prediction interval were better met on the probability than on the log-odds scale.*[62]

However, by looking at residuals from random-effects models the authors do not separate between the sampling error (within clusters) and the between-cluster heterogeneity in their study; in other words the normality distribution of the errors is some (weighted) amalgamation of the sampling distribution and the between-study distribution. Therefore, it is not possible to draw firm conclusions about whether the between-study distribution is approximately normal from their work. Furthermore, they only consider example datasets, rather than examination through statistical theory or simulation, and so their recommendations may not be generalizable.

The aim of this chapter is therefore to explore the distributions and possible transformations of the C-statistic and other performance statistics to evaluate how reasonable the between-study normality assumption is for pooling performance statistics using the univariate random-effects meta-analysis models proposed in Chapter 5. To achieve this, an in-depth simulation study is conducted using logistic regression for the underlying true prediction model, from which the C-statistic, expected/observed number of events (E/O), calibration slope and calibration-in-the-large are calculated as measures of model performance. The distributions of these performance statistics across studies are then evaluated under different levels of between-study heterogeneity in either the intercept or predictor effect in the logistic model. The simulation uses large study samples to reduce sampling error to almost zero, thereby allowing the true between-study distributions of the performance statistics to be revealed under different conditions, and thus avoiding the amalgamation of the within-study and between-study distributions that potentially affects the work of van Klaveren et al.[62]

The aims of the chapter are summarised as follows:

1. **Evaluate if true performance statistics are normally distributed**

   Evaluate the distributions of the true C-statistic, calibration slope, calibration-in-the-large and expected/observed proportion of events to assess if normality of between-study performance statistics is a reasonable assumption when there is heterogeneity in either the baseline risk (intercept) or predictor effect (beta). This will also be evaluated in different scenarios such as in data with particularly high or low baseline risk and for different strengths of the predictor effect.

2. **Consider appropriate transformations for performance statistics**

   If any of the performance statistics are not approximately normally distributed when there is heterogeneity in the intercept or predictor effect, transformations of the statistic will be considered to see if the distributions of the transformed statistic are closer to a normal distribution.

3. **Make recommendations for the scale on which to combine performance statistics in a meta-analysis**

   Based on the findings of this simulation study, recommendations will be made for the scale most appropriate for pooling performance statistics in the random-effects meta-analysis models proposed in Chapter 5.

6.2 Methods

6.2.1 Specifying the 'true' prediction models and nine base scenarios for simulation

In all the simulations, a 'true' logistic regression prediction model for a binary outcome such as diagnosis of DVT is considered (building on the DVT modelling work in Chapter 5), with just a single predictor of age included in the model for simplicity. Different scenarios and settings will be considered, covering applications where this 'true' model is correct, and others where this 'true' model is incorrect due to missing predictors or unexplained heterogeneity. In all situations examined, the 'true' model is applied to simulated patient data from a range of studies and the performance statistics (e.g. C-statistics) of interest calculated in each study. The between-study distributions of the calculated performance statistics are then examined for normality.

The simplest scenarios for the 'true' logistic model, which are termed the 'base scenarios' are now defined. Information from Oudega et al. is used as a starting point for the parameter values chosen in the true model and for the distribution of age.[200] The authors developed a simple diagnostic rule to safely rule out DVT in patients with suspected DVT, without the need for referral. In the article, age had a mean of 60 years and a standard deviation of 17.6 years. These values are used here within an assumed normal distribution for age when it is included as a predictor in the simulation study. The assumed 'true' logistic regression prediction model can be written as

$$\text{logit}\left(p_{ij}\right)=\mu_{\alpha}+\mu_{\beta}\times\text{age}_{ij} \qquad \textbf{(6.2)}$$

where *i* represents the individual in study *j* and the age (in years) of individuals in each study is sampled from a $\text{Normal}\left(60,\, 17.6^2\right)$ distribution. Note that unrealistic ages sampled from the $\text{Normal}\left(60,\, 17.6^2\right)$ such as negative ages, very young and very old ages are not considered a problem as it is simply a 'variable' with a specified distribution in this simulation study. However, restricting the range of values to plausible age values is considered further in the extensions to the simulation study (Section 6.2.6).

Oudega et al. reported a univariable odds ratio for age (per year increase) of 1.01 which approximately relates to μ_{β} = 0.01.[200] They also reported a DVT prevalence of 0.22 which, given μ_{β} = 0.01, can be achieved by generating data with α = -1.274. If data are generated using the model with these parameter values for μ_{α} and μ_{β}, the C-statistic of the model is approximately 0.55. This represents one base scenario for the simulation study. Eight further base scenarios were also considered, by varying model parameters to give combinations of different outcome prevalence levels and discrimination (C-statistic). This ranged from prevalence levels of 0.22 (as reported by Oudega et al.[200]), to a low prevalence of 0.05, and then a high prevalence of 0.9. The C-statistics ranged from around 0.55 (poor discrimination, not much better than chance), to around 0.7 (moderate discrimination), and then 0.9 (excellent discrimination), covering the range of poor to excellent C-statistic/AUC levels of discrimination, as defined by Hosmer and Lemeshow.[201] A C-statistic of 0.7 is fairly representative of the discriminatory ability of a prognostic prediction model (such as in Chapter 4), and a C-statistic of 0.9 may be akin to the discrimination of an excellent diagnostic prediction model.

The nine base scenarios are defined in Table 6.1 in terms of μ_α (representing the intercept) and μ_β (representing the predictor effect) for the assumed 'true' model to be used in the simulations.

Table 6.1: Parameter values of assumed prediction model in the nine base scenarios considered in the simulation study.

Scenario	μ_α	μ_β	Prevalence*	C-statistic*
1	-1.274	0.010	0.22	0.55
2	-2.957	0.010	0.05	0.55
3	2.210	0.010	0.90	0.55
4	-1.425	0.045	0.22	0.7
5	-3.215	0.045	0.05	0.7
6	2.440	0.045	0.90	0.7
7	-2.386	0.145	0.22	0.9
8	-5.133	0.145	0.05	0.9
9	3.987	0.145	0.90	0.9

*The μ_α and μ_β values are selected to give the corresponding average prevalence and C-statistic (average from 100 large samples each of 1000000 patients) when there is no heterogeneity in α or β (σ_α^2=0 and σ_β^2=0).

6.2.2 Specifying seven settings that allow for heterogeneity

The nine base scenarios defined in Table 6.1 above are the basis for the simulation study, and the given μ_α and μ_β values indicate the intercept and predictor effect values of the assumed 'true' model in each scenario given by (6.2).

As the aim of this chapter is to check the between-study distributions of true performance statistics when there is heterogeneity in true model performance, it was necessary to consider each base scenario in a range of seven different settings (Table 6.2) that introduce heterogeneity for either α or β when generating the patient-level data. This was done by sampling study-specific true intercept and predictor effect values, α_j and β_j values respectively for each study j from α_j~Normal$(\mu_\alpha,\sigma_\alpha^2)$, β_j~Normal$\left(\mu_\beta,\sigma_\beta^2\right)$. Thus, there is heterogeneity in α if $\sigma_\alpha > 0$ and heterogeneity in β if $\sigma_\beta > 0$. In such situations, the assumed 'true' model of (6.2) is incorrect, as it ignores heterogeneity and thus, when it is applied to a range of populations, will induce heterogeneity in its performance. If there is no heterogeneity, all the true study-specific intercepts are identical (=μ_α), as are the true study-specific predictor effects (=μ_β). Thus, in such situations the assumed true model (6.2) is correct in all populations.

In this chapter, heterogeneity is considered for either α or β but not for both at the same time, for simplicity. Across the settings, three values were selected for σ_α and three values for σ_β. The values were selected to give small, moderate and large variation in either α or β but are fairly arbitrary as the amount of variability in the performance statistic also depends on the α or β values themselves. The SDs for β (σ_β's) were chosen to be values of about half of each of the β values used in defining the scenarios. Therefore the largest value of σ_β will be extreme for scenarios 1 to 3 and large for scenarios 4 to 6. The selected values of SDs for α or β are given in Table 6.2, for each of the seven different settings covered.

In summary, the total number of simulations to be considered is therefore 63 (= 9 base scenarios x 7 settings), which covers a range of values for α and β, and the amount of heterogeneity present.

Table 6.2: Defined settings for simulation, with heterogeneity in either *α* or *β* across studies.

Simulation setting	Standard deviation for *α* and *β* in each scenario	
1: No heterogeneity	$\sigma_\alpha = 0$	$\sigma_\beta = 0$
2: Little heterogeneity in *α*	$\sigma_\alpha = 0.1$	$\sigma_\beta = 0$
3: Moderate heterogeneity in *α*	$\sigma_\alpha = 0.5$	$\sigma_\beta = 0$
4: Large heterogeneity in *α*	$\sigma_\alpha = 1.0$	$\sigma_\beta = 0$
5: Little heterogeneity in *β*	$\sigma_\alpha = 0$	$\sigma_\beta = 0.005$
6: Moderate heterogeneity in *β*	$\sigma_\alpha = 0$	$\sigma_\beta = 0.020$
7: Large heterogeneity in *β*	$\sigma_\alpha = 0$	$\sigma_\beta = 0.070$

6.2.3 Choice of sample size and number of studies

To examine the between-study distribution of *true* performance statistics, sampling error of performance estimates (e.g. C-statistic) needed to be removed as otherwise the observed between-study distribution would be a mixture of sampling error and the true between-study distribution. To do this, ideally studies of infinite sample sizes were needed, but as this was not practical, various choices of 'large' sample sizes were trialled to reduce within-study sampling errors to a very small value (i.e. close to zero). Studies of size 10000, 50000, 100000, 500000 and 1000000 were considered for scenario 1, and the smallest sample size that still gave negligible observed sampling error was chosen. This was studies of 500000 (see section 6.3.1 and Appendix E1 for detailed justification).

It was also important to examine the number of studies that needed to be generated in each simulation, to adequately reveal the shape of the between-study distribution for each performance statistic. Again using scenario 1, the numbers of studies considered were 100, 500, 1000 and 2000 and the smallest number of studies that still showed the distribution clearly was chosen. This was 1000 (see section 6.3.1 and Appendix E1 for detailed justification).

6.2.4 Generating patient-level data and obtaining the distribution of true performance statistics

For each scenario in each setting, data (i.e. an age and binary outcome response) needed to be generated for 500000 patients in each of the 1000 studies, so that the ‘true’ model could be applied to the data and performance statistics then calculated.

The patient data were generated for each study j by first sampling α_j and β_j values from the distributions α_j~Normal$\left(\mu_\alpha,\sigma_\alpha^2\right)$, β_j~Normal $\left(\mu_\beta,\sigma_\beta^2\right)$. Age was sampled for each patient from age$_{ij}$ ~ Normal$\left(60,\ 17.6^2\right)$, the linear predictor was then calculated for each patient as $LP_{ij}=\alpha_j+\beta_j\times\text{age}_{ij}$ and from this the outcome probability was calculated, $p_{ij}=\frac{\exp(LP_{ij})}{1+\exp(LP_{ij})}$. Using a Bernoulli distribution with probability p_{ij}, a binary outcome was sampled for each patient (outcome = 0 or 1).

For each of the nine simulation scenarios (defined in Table 6.1) within each of the seven simulation settings (defined in Table 6.2), the distributions of the following true performance statistics were of interest: C-statistic, expected/observed number of events (E/O), calibration slope and calibration-in-the-large. For the simulated data in each generated study, these performance statistics were estimated as described in Chapter 1 (Section 1.6.3) for the ‘true’ model assumed for that scenario as specified in Table 6.1.

The distribution of the obtained performance statistics across the 1000 studies was then summarised using the mean, SD, median, minimum, maximum and interval containing 95% of values. The coefficient of skewness and coefficient of kurtosis were also calculated. Summary statistics were then compared to those expected if the distribution was normal, for example an equal mean and median, skewness of 0 and kurtosis of 3.

The distributions of performance statistics was also considered graphically by plotting histograms (with normal distribution lines using mean and SD from data overlaid). The symmetry and shape of the distribution could then be compared to a normal distribution.

Transformations

In situations where the between-study distribution of performance statistics did not appear approximately normal, transformations of the statistic were also calculated, plotted and summarised to ascertain if they made improvements. Transformations considered were natural log, logit, arcsine (considered for the C-statistic) and square root (considered for E/O). Arcsine and square root transformations were included as previous work by Trikalinos et al. recommended variance stabilizing transformations for meta-analysis of proportions and rates.[202]

6.2.5 Step-by-step guide to simulating data and summarising performance statistics

For added clarity, the previous sections are now summarised in Box 6.1 to give a step-by-step process of the simulation study.

Box 6.1: Outline of the steps taken in simulating data and then examining between-study distributions of ‘true’ model performance statistics.

Step 1: Define the assumed true prediction model (i.e. select one of the scenarios 1-9, Table 6.1).

Step 2: Define the amount of between-study heterogeneity in either the study-specific intercepts (α_j) or predictor effects (β_j) (i.e. select one of the seven settings in Table 6.2).

Step 3: Specify 1000 studies (study denoted by *j*) and, within each study, specify 500000 individuals (individuals in a study are denoted by *i*) for which to generate data.

Step 4: For each study *j*, sample a true intercept (α_j) and predictor effect (β_j) from α_j~Normal$(\mu_\alpha,\sigma_\alpha^2)$ and β_j~Normal$\left(\mu_\beta,\sigma_\beta^2\right)$.

Step 5: Within each study *j*, generate patients *i* =1, …, 500000 and generate the age value for each patient by sampling from age_{ij} ~ Normal$(60, 17.6^2)$.

Step 6: Within each study *j, g*enerate the binary outcome variable for each patient by first calculating the linear predictor for each individual, $LP_{ij}=\alpha_j+\beta_j\times\text{age}_{ij}$ and use this to calculate the probability of the event occurring by $p_{ij}=\frac{\exp(LP_{ij})}{1+\exp(LP_{ij})}$. The binary outcome is then sampled from a Bernoulli distribution, outcome_{ij}~Bernoulli(p_{ij}).

Step 7: Estimate the performance statistics of interest in each study, by taking the assumed ‘true’ prediction model, for the scenario chosen in step 1, $\text{logit}\left(p_{ij}\right)=\mu_\alpha+\mu_\beta\times\text{age}_{ij}$ and fitting it to the generated data from steps 5 and 6.

Step 8: Summarise the distribution of the obtained performance statistics across the 1000 studies by calculating summary statistics (mean, SD, median, minimum, maximum and range containing 95% of values), plotting histograms and calculating coefficients of skewness and kurtosis.

Step 9: Repeat steps 7 and 8 to ascertain if distributions are more normally distributed if transformations such as natural log, logit, arcsine or square root are applied to the calculated performance statistics.

Programming of simulations

The simulation was conducted using Stata 13. An example of the Stata code used to generate the data and evaluate the distributions of the performance statistics is given in Appendix E2. Simulations typically took approximately 1 hour to run but up to 3 hours in some settings such as those in ‘Extension 3’ described below.

6.2.6 Extensions: more realistic heterogeneity inducing mechanism

Further, extended simulation settings were also considered as follows.

Extension 1: Limiting the age range to between 18 and 100 years

The values of age sampled for patients were restricted to between 18 and 100 years to be more realistic. Therefore, if an age<18 years or age>100 years was sampled for a patient, age for that patient would be classed as missing and another value would be sampled until an age within the specified range was found.

Extension 2: Varying the distribution of age values across studies

Previously the distribution from which age was sampled was $\text{Normal}(60, 17.6^2)$ and this remained the same across studies. In this additional simulation setting, the distribution from which age was sampled was allowed to vary across studies for both the mean and SD of age in each study. The mean and SD values for age were therefore also sampled assuming normal distributions. It was assumed that $\text{age}_{ij} \sim \text{Normal}\left(\mu_j, \sigma_j^2\right)$, where $\mu_j \sim \text{Normal}(60, 10^2)$ and $\sigma_j^2 \sim \text{Normal}(17.6, 4^2)$.

Extension 3: Including an additional predictor and interaction

Additional simulation settings were also considered that involve generating data from a model that includes a second predictor and an interaction between age and the additional predictor. However, the ‘true’ model evaluated for performance in each study still only included age. Thus, these additional simulations reflect a situation where the ‘true’ model to be used in clinical practice is incomplete (i.e. it misses important predictors), and are therefore a potentially more realistic alternative to those settings described previously. For simplicity, no heterogeneity in the intercept or predictor effects was considered in this extended setting and simulations were also restricted to scenarios 4 to 6 where the predictor effect was moderate rather than weak (scenarios 1 to 3) or strong (scenarios 7 to 9). Scenarios 4 to 6 were considered ideal as the original predictor age, could discriminate reasonably well between patients that have the event and patients that do not, but with room for improvement in the model if a further predictor and interaction were added.

The model for generating data in this extended setting can now be specified as follows:

$$\text{logit}\left(p_{ij}\right) = \alpha + \beta_1 \text{age}_{ij} + \beta_2 \text{pred}_{ij} + \beta_3 (\text{age}_{ij} \times \text{pred}_{ij}) \qquad \textbf{(6.3)}$$

This extended setting was considered for both a continuous and a categorical predictor (pred$_{ij}$). For settings in which pred$_{ij}$ was continuous, the original distribution of age was used for pred$_{ij}$ (not restricting values or allowing the distribution to vary across studies), so values were sampled from Normal$(60, 17.6^2)$ and the predictor effect assumed to be weak (β_2=0.01). Alternatively, when the additional predictor was categorical, a prevalence of 0.36 was assumed (using sex as an example from Oudega et al.[200] with β_2=0.1), and a correlation of +0.5 assumed between age and pred so that they were not independent.

Different strengths of interaction effects were considered, depending on whether the additional predictor was continuous or categorical. The values of β_3 were decided by comparing what the probability of the event would be with and without the additional predictor and interaction between the two predictors (Appendix E3).

Table 6.3: Defined simulation settings for model with additional predictor and interaction between age and additional predictor.

Simulation setting	Additional predictor OR (OR=exp(β_2))	Interaction OR (OR=exp(β_3))
Extension 3(i)	Continuous, OR=1.01	1.0010
Extension 3(ii)	Continuous, OR=1.01	1.0005
Extension 3(iii)	Continuous, OR=1.01	1.0001
Extension 3(iv)	Categorical, OR=1.1	1.0300
Extension 3(v)	Categorical, OR=1.1	1.0100
Extension 3(vi)	Categorical, OR=1.1	1.0050

The data for the 500000 patients for each of the 1000 studies was generated, using the new model (6.3) with specified parameter values, in a similar manner to the steps outlined in Box 6.1. Once more, the assumed 'true' prediction model only included a single predictor; that is, the additional predictor and the interaction were not included in the assumed 'true' prediction model for which performance was evaluated. The assumed value of the single coefficient, β_1 in the 'true' prediction model would, in reality also account for some of the variation in the other terms not fitted. Therefore, to calculate its assumed value, a large sample of five million patients was generated to estimate α and β_1 and these estimates taken as the 'true' values for evaluating the prediction model in each study generated (Table 6.4). This affected the intercept values and the β_1 values were slightly larger only when the missing predictor was continuous.

Table 6.4: Parameter values of assumed 'true' prediction model for extended simulation settings when data are generated including an additional predictor and interaction (original simulation setting 1 included for comparison).

Scenario	Simulation setting	Assumed 'true' α	Assumed 'true' β_1
4	Setting 1: original base scenario	-1.425	0.045
	Extension 3(i): missing continuous predictor & interaction	-1.127	0.050
	Extension 3(ii): missing continuous predictor & interaction	-1.116	0.050
	Extension 3(iii): missing continuous predictor & interaction	-1.125	0.050
	Extension 3(iv): missing categorical predictor & interaction	-1.393	0.045
	Extension 3(v): missing categorical predictor & interaction	-1.390	0.045
	Extension 3(vi): missing categorical predictor & interaction	-1.389	0.045
5	Setting 1: original base scenario	-3.215	0.045
	Extension 3(i): missing continuous predictor & interaction	-2.904	0.050
	Extension 3(ii): missing continuous predictor & interaction	-2.905	0.050
	Extension 3(iii): missing continuous predictor & interaction	-2.904	0.050
	Extension 3(iv): missing categorical predictor & interaction	-3.184	0.045
	Extension 3(v): missing categorical predictor & interaction	-3.176	0.045
	Extension 3(vi): missing categorical predictor & interaction	-3.183	0.045
6	Setting 1: original base scenario	2.440	0.045
	Extension 3(i): missing continuous predictor & interaction	2.719	0.050
	Extension 3(ii): missing continuous predictor & interaction	2.672	0.050
	Extension 3(iii): missing continuous predictor & interaction	2.718	0.050
	Extension 3(iv): missing categorical predictor & interaction	2.486	0.045
	Extension 3(v): missing categorical predictor & interaction	2.488	0.045
	Extension 3(vi): missing categorical predictor & interaction	2.455	0.045

6.2.7 Summary of elements considered in the simulation study

To examine the normality assumption for the performance statistics, the distributions are evaluated under different conditions to ensure robust conclusions are drawn. A summary of the different elements considered to be important in this simulation study are given in Table 6.5 below.

Table 6.5: Summary of elements that could affect the distribution of performance statistics and if/how they have been considered in this simulation study.

Element	If/how considered in this simulation study
Values of prevalence	Used to select intercept values in defined base scenarios (3 prevalence values)
Values of C-statistic	Used to select the predictor effects in defined base scenarios (3 values of the C-statistic)
Heterogeneity in intercept (*α*)	Intercept allowed to vary across studies with specified variances defined in simulation settings (3 values for SD of *α*)
Heterogeneity in predictor effect (*β*)	Predictor effect allowed to vary across studies with specified variances defined in simulation settings (3 values for SD of *β*)
Distribution of intercepts across studies	Assumed to be normally distributed across studies **Further work:** include different distributions such as log-normal
Distribution of predictor effects across studies	Assumed to be normally distributed across studies **Further work:** include different distributions such as log-normal
Distribution of predictor values	Age of patients assumed to come from same normal distribution for all studies Mean and SD later varied across studies in extension 2
Complexity of model	Original scenarios only include one predictor Additional predictor and interaction (excluded from prediction model) in extension 3
Type of model	Logistic regression model used for prediction model **Further work:** Use survival model as prediction model

6.3 Results

6.3.1 Performance statistics when there is no heterogeneity (setting 1)

When there is no heterogeneity in α and β, the assumed true prediction model specification is correct in all studies; therefore, any variability across studies in the simulated estimates of model performance can only be due to sampling error. Recall that in Section 6.2.3, it was emphasised that 500000 individuals were chosen per study in order to keep sampling error minimal. To illustrate this, data were generated for each of the nine scenarios defined in Table 6.1 for simulation setting 1, which had no heterogeneity in α or β. The summary statistics obtained for the performance statistics are reported below in Table 6.6. The means and medians are very similar because the distributions appear symmetric. More importantly, the range of values between the minimum and maximum are very small for most performance statistics in most scenarios. Therefore, although sampling error still exists when using 500000 individuals in each study, it is extremely small, and thus crucially will be swamped by the genuine between-study distribution when heterogeneity is introduced for α or β in settings 2 to 9 below. In other words, the observed distribution of performance estimates across the 1000 studies will be dominated by the true between-study distribution, therefore enabling the normality assumption to be examined as desired. Perhaps the only exception to this is for the calibration slope in scenarios 1 to 3, in which the predictor effect is only weak. However, the discriminatory performance of these models is poor and therefore these models are less likely to be of interest in reality if considering prediction intervals for model performance in a new setting.

Table 6.6: Summary statistics for the distribution of performance statistics in all scenarios when there is no heterogeneity in *α* or *β* (setting 1).

Performance Statistic	Scenario	Mean (SD)	Median	95% Range	Min to max
C-statistic	1	0.5493 (0.0010)	0.5494	0.5473 to 0.5513	0.5458 to 0.5531
	2	0.5495 (0.0019)	0.5494	0.5458 to 0.5532	0.5434 to 0.5555
	3	0.5494 (0.0014)	0.5495	0.5469 to 0.5520	0.5451 to 0.5553
	4	0.7028 (0.0009)	0.7029	0.7011 to 0.7046	0.6994 to 0.7063
	5	0.7089 (0.0017)	0.7089	0.7055 to 0.7121	0.7039 to 0.7141
	6	0.7066 (0.0012)	0.7066	0.7042 to 0.7089	0.7029 to 0.7103
	7	0.9067 (0.0005)	0.9067	0.9058 to 0.9077	0.9050 to 0.9081
	8	0.9301 (0.0007)	0.9301	0.9287 to 0.9315	0.9277 to 0.9321
	9	0.9196 (0.0006)	0.9196	0.9185 to 0.9208	0.9177 to 0.9217
E/O	1	1.0000 (0.0027)	1.0001	0.9946 to 1.0049	0.9911 to 1.0089
	2	1.0001 (0.0062)	1.0001	0.9884 to 1.0122	0.9794 to 1.0234
	3	1.0000 (0.0005)	1.0000	0.9991 to 1.0010	0.9985 to 1.0018
	4	0.9999 (0.0026)	0.9999	0.9948 to 1.0047	0.9906 to 1.0086
	5	1.0001 (0.0061)	0.9999	0.9885 to 1.0121	0.9809 to 1.0206
	6	1.0000 (0.0005)	1.0000	0.9991 to 1.0009	0.9983 to 1.0013
	7	1.0000 (0.0020)	1.0000	0.9962 to 1.0038	0.9924 to 1.0073
	8	1.0001 (0.0052)	1.0001	0.9901 to 1.0105	0.9841 to 1.0154
	9	1.0000 (0.0004)	1.0000	0.9993 to 1.0008	0.9987 to 1.0012
Calibration-in-the-large	1	0.0000 (0.0035)	0.0001	-0.0067 to 0.0072	-0.0109 to 0.0137
	2	-0.0001 (0.0065)	-0.0001	-0.0127 to 0.0123	-0.0244 to 0.0220
	3	-0.0001 (0.0048)	-0.0002	-0.0098 to 0.0093	-0.0179 to 0.0156
	4	0.0001 (0.0036)	0.0001	-0.0066 to 0.0074	-0.0121 to 0.0134
	5	-0.0001 (0.0067)	0.0001	-0.0132 to 0.0126	-0.0222 to 0.0210
	6	-0.0003 (0.0050)	-0.0003	-0.0099 to 0.0094	-0.0134 to 0.0177
	7	0.0000 (0.0045)	-0.0001	-0.0088 to 0.0089	-0.0169 to 0.0176
	8	-0.0002 (0.0081)	-0.0002	-0.0162 to 0.0154	-0.0236 to 0.0250
	9	-0.0003 (0.0062)	-0.0004	-0.0126 to 0.0119	-0.0194 to 0.0214
Calibration slope	1	0.9992 (0.0199)	0.9987	0.9596 to 1.0377	0.9264 to 1.0736
	2	0.9992 (0.0372)	0.9981	0.9283 to 1.0745	0.8769 to 1.1097
	3	0.9999 (0.0270)	1.0005	0.9465 to 1.0526	0.9158 to 1.1105
	4	0.9997 (0.0051)	0.9999	0.9893 to 1.0096	0.9829 to 1.0187
	5	0.9995 (0.0089)	0.9994	0.9822 to 1.0164	0.9721 to 1.0256
	6	1.0000 (0.0065)	1.0001	0.9867 to 1.0127	0.9814 to 1.0200
	7	1.0001 (0.0033)	1.0000	0.9937 to 1.0068	0.9897 to 1.0106
	8	1.0001 (0.0048)	1.0001	0.9907 to 1.0098	0.9849 to 1.0144
	9	1.0000 (0.0040)	0.9999	0.9925 to 1.0080	0.9874 to 1.0148

6.3.2 Distributions of performance statistics given heterogeneity (settings 2 to 7)

The between-study distribution of the calculated performance statistics is now summarised for settings 2 to 9, for each performance statistic separately.

Between-study distribution for true C-statistic

Heterogeneity in alpha (settings 2 to 4)

The between-study distribution for the C-statistic was approximately normally distributed for all nine scenarios in simulation settings 2 and 3 when there was little or moderate heterogeneity in α (σ_α=0.1 or σ_α=0.5 respectively). However, most scenarios exhibit little between-study variation (as seen in Table 6.7 for setting 3: σ_α=0.5). The SD of the between-study distribution increases slightly for stronger predictor effects in scenarios 4 to 6 and larger still in scenarios 7 to 9, compared to setting 1: σ_α=0 in Table 6.6, but the distributions are still very narrow as seen by the range of values (minimum to maximum). The skewness in all scenarios remains close to the ideal of 0 and kurtosis remains relatively close to the ideal of 3.

Table 6.7: Summary statistics for the distribution of the C-statistic in all scenarios when heterogeneity in α is moderate (setting 3: σ_α=0.5).

Scenario	Mean (SD)	Median	Min to max	Skewness	Kurtosis
1	0.5494 (0.0010)	0.5494	0.5459 to 0.5527	-0.0312	3.5836
2	0.5494 (0.0019)	0.5494	0.5429 to 0.5569	-0.1603	3.8797
3	0.5494 (0.0014)	0.5495	0.5442 to 0.5548	-0.0239	3.4299
4	0.7031 (0.0019)	0.7029	0.6988 to 0.7099	0.4313	3.0387
5	0.7088 (0.0022)	0.7087	0.7023 to 0.7175	0.2915	3.3969
6	0.7065 (0.0021)	0.7064	0.6997 to 0.7146	0.1509	3.2987
7	0.9071 (0.0033)	0.9068	0.8982 to 0.9177	0.3652	2.9306
8	0.9302 (0.0046)	0.9302	0.9132 to 0.9429	-0.0820	2.9636
9	0.9196 (0.0044)	0.9194	0.9075 to 0.9362	0.1799	2.9647

The between-study distribution also remained approximately normally distributed for most scenarios when there was large heterogeneity in α (setting 4: σ_α=1.0). The coefficient of skewness for the C-statistic in different scenarios was between -0.32 and 0.24 with the exceptions of scenarios 4 and 7 where the average intercept (baseline risk) was selected to give a prevalence of 0.22 if there was no heterogeneity in α (Figure 6.1). For scenarios 4 and 7, skewness of the C-statistic was estimated higher at 0.71 and 0.65 respectively which is considered moderately skewed.

The between-study distribution of the C-statistic was very narrow when there was heterogeneity in α (simulation settings 2 to 4), especially when the predictor was weak as in scenarios 1 to 3. The SDs in these scenarios are similar to those in setting 1 where there is no heterogeneity in α or β. The C-statistic is a measure of discrimination and therefore is affected by the strength of the β and does not directly depend on α. However, for scenarios with stronger predictors (and therefore larger C-statistics), the SD for the distribution of C-statistics increases as the level of heterogeneity in α increases. For example, scenarios 4 to 6 have larger C-statistics than scenarios 1 to 3 and scenarios 7 to 9 have the largest C-statistics of the nine scenarios, and therefore also have a larger increase in SD as heterogeneity in α increases. This is seen in Figure 6.1 which shows the distribution of the C-statistic for scenario 1 (weak predictor), scenario 4 (moderate predictor) and scenario 7 (strong predictor) for settings 1 to 4.

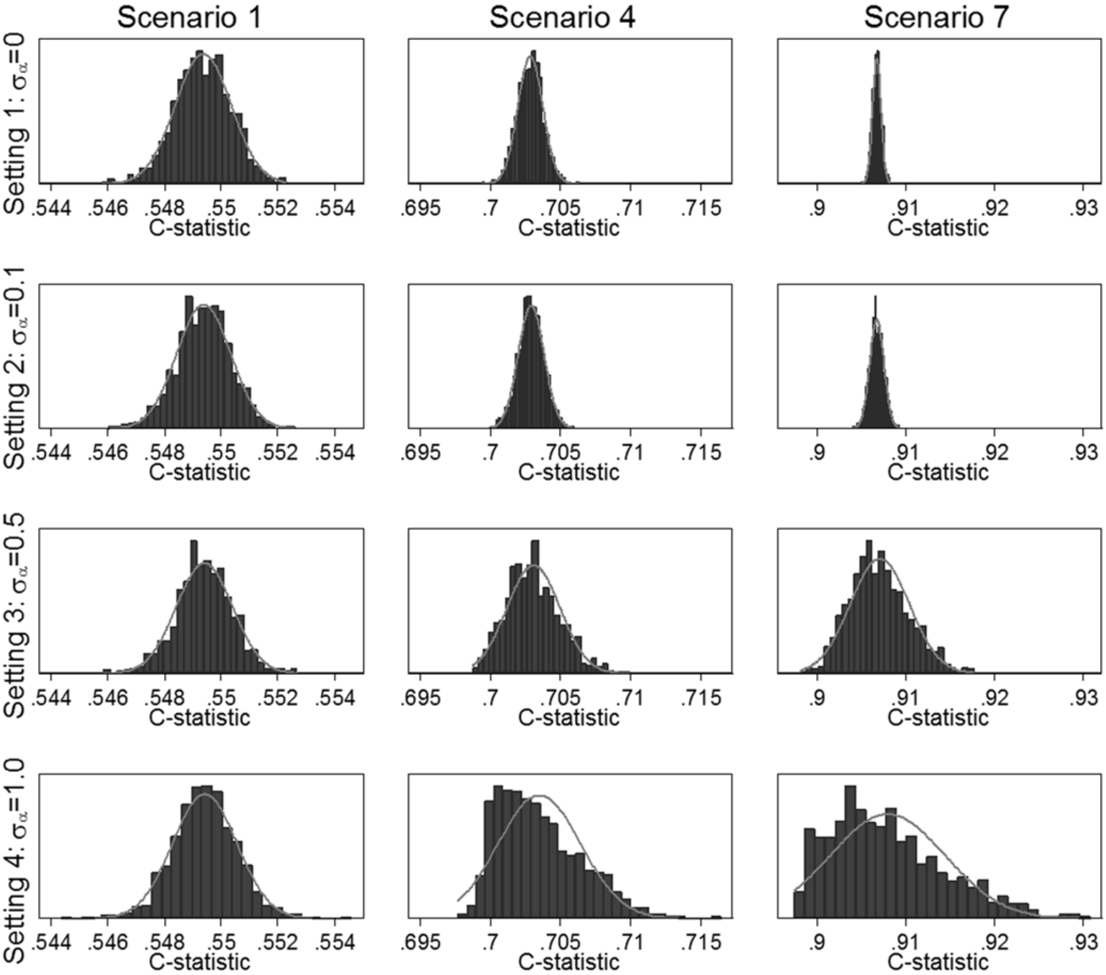


Figure 6.1: Histograms for the C-statistic in settings 1 to 4 (different values of σ_α) for scenarios 1, 4 and 7 (scenarios with average intercept from base case with weak, moderate and strong predictor effects respectively).

Heterogeneity in beta (settings 5 to 7)

As heterogeneity in β increases, the distribution of the C-statistic deviates more from normality. Firstly, when there was a small amount of heterogeneity in β (setting 5: σ_β=0.005), the distributions were approximately normal in all scenarios. When heterogeneity increased to a moderate amount (setting 6: σ_β=0.02), scenarios 7 to 9 which have a strong predictor started to skew. Lastly, when there was large heterogeneity in β (setting 7: σ_β=0.07), none of the distributions were normal (Figure 6.2). Particularly for scenarios 7 to 9, the distributions were very skewed (skewness ~ -2.3) when heterogeneity in β was large (setting 7: σ_β=0.07).

This is probably because discrimination was good (around 0.9 even when there was no heterogeneity in *β,* setting 1) and the C-statistic has a maximum value of one. For scenarios 1 to 3 with a weak predictor, the between-study C-statistic is almost uniformly distributed and includes values less than 0.5 which would indicate the model inversely discriminating between events and non-events. The normality assumption is therefore often inappropriate for the C-statistic when there is heterogeneity in the predictor effect (*β*) and the C-statistic deviates from normality more as heterogeneity in *β* increases. Hence, transformations of the C-statistic were also considered to improve the between-study distribution.

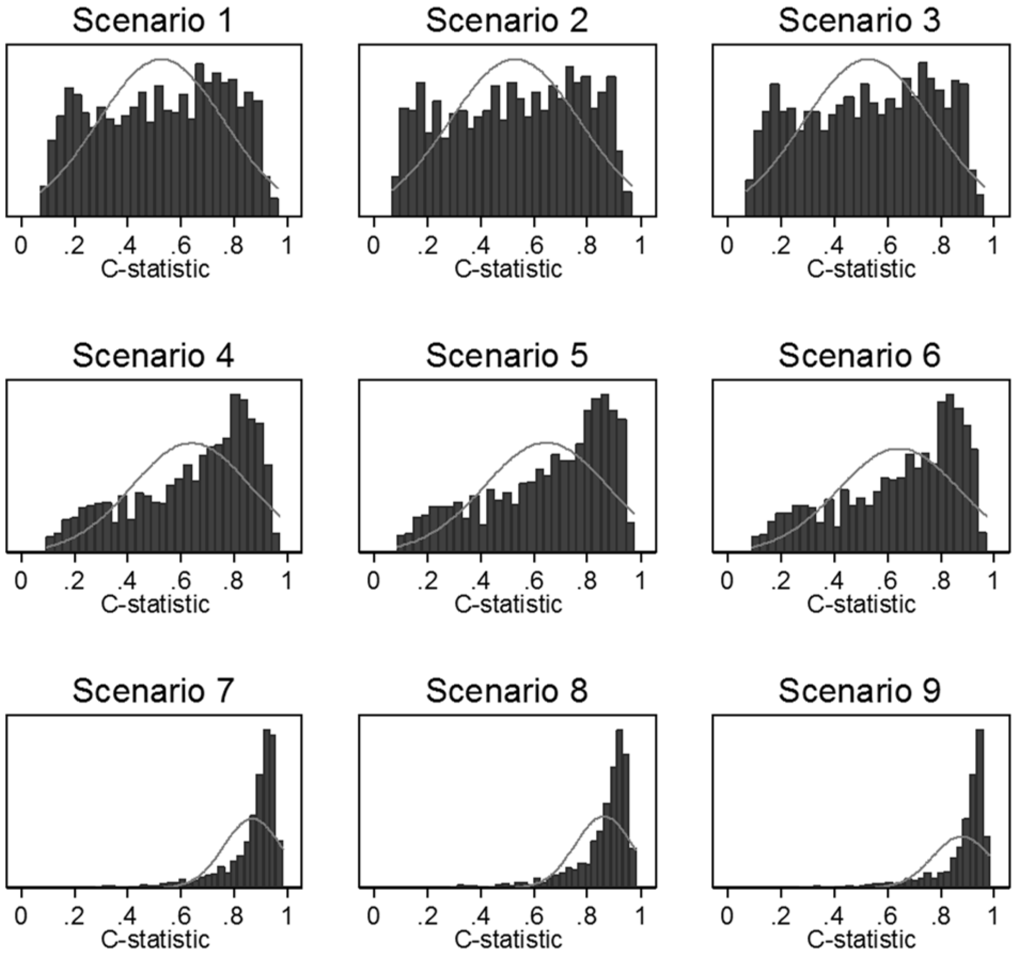


Figure 6.2: Histograms for C-statistic in all scenarios when heterogeneity in *β* is large (setting 7: σ_β=0.07).

Transformations of the C-statistic

Given the apparent non-normality of the C-statistic on its original scale in some settings, natural log, logit and arcsine transformations of the C-statistic were also evaluated to ascertain if it improved normality, especially, when heterogeneity in β was large in setting 7.

The natural log transformation did not offer any improvement in achieving normality in scenarios and settings that deviated from normality on the original C-statistic scale. For example, in setting 7 when heterogeneity in β was large, skewness was around -3.3 for scenarios 7 to 9 and kurtosis was high (~17) for log-transformed C-statistics. Histograms for setting 7: σ_β=0.07 are given in Appendix Figure E4.1).

However, the logit transformation greatly reduced the skewness in most scenarios when heterogeneity in β was large (setting 7: σ_β=0.07, see Figure 6.3 and Table 6.8). This was the simulation setting in which the C-statistic on the original scale deviated from normality the most. Scenarios 7 to 9 were still skewed (skewness of -1, Table 6.8) but far less skewed than on the original C-statistic scale (skewness of -2.4). In addition to this, the between-study distribution of the logit C-statistic was no worse in scenarios and settings that were approximately normal on the original untransformed scale.

The arcsine transformation (as suggested by Trikalinos et al.[202] for proportions) did improve some skewed distributions, however it did not perform as well as the logit transformation in the scenarios and settings considered (see Appendix Figure E4.2 and E4.3 for example comparisons of scales).

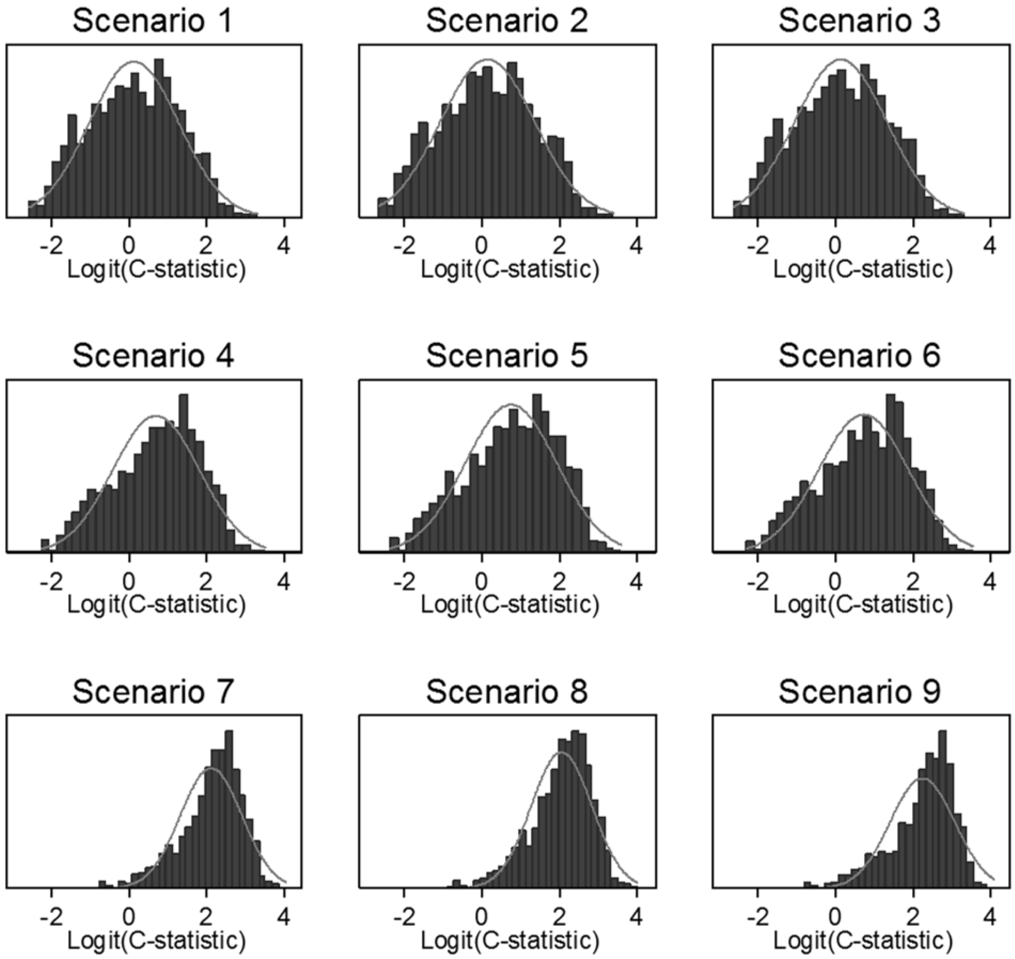


Figure 6.3: Histograms for logit(C-statistic) in all scenarios when heterogeneity in β is large (setting 7: σ_{β}=0.07).

Table 6.8: Summary statistics and distribution skewness and kurtosis for C-statistic and logit transformed C-statistic in two simulation settings that showed skewed distributions for the C-statistic on the original scale.

Simulation setting	Scenario	Mean		Median		Skewness		Kurtosis	
		C-statistic	Logit C-statistic	C-statistic	Logit C-statistic	C-statistic	Logit C-statistic	C-statistic	Logit C-statistic
Setting 4: large heterogeneity in α σ_α=1.0	1	0.5494	0.1983	0.5494	0.1982	-0.0194	-0.0173	4.0557	4.0557
	2	0.5494	0.1981	0.5494	0.1982	-0.3168	-0.3077	7.5349	7.5351
	3	0.5494	0.1984	0.5495	0.1986	0.3264	0.3318	6.5436	6.5673
	4	0.7035	0.8640	0.7030	0.8617	0.7101	0.7238	3.0712	3.1072
	5	0.7085	0.8879	0.7084	0.8878	0.1035	0.1271	3.3865	3.4050
	6	0.7062	0.8772	0.7063	0.8776	0.2023	0.2234	3.1442	3.1891
	7	0.9082	2.2943	0.9075	2.2830	0.6477	0.8231	3.0547	3.5138
	8	0.9302	2.5978	0.9303	2.5911	-0.1147	0.1787	2.7381	2.7057
	9	0.9197	2.4435	0.9193	2.4328	0.2438	0.5141	2.7482	3.1975
Setting 7: large heterogeneity in β σ_β=0.07	1	0.5261	0.1284	0.5379	0.1518	-0.1308	-0.0562	1.8153	2.2536
	2	0.5268	0.1340	0.5379	0.1521	-0.1329	-0.0591	1.8154	2.2633
	3	0.5266	0.1315	0.5379	0.1518	-0.1323	-0.0585	1.8122	2.2514
	4	0.6396	0.7005	0.6941	0.8194	-0.6462	-0.3442	2.2930	2.4276
	5	0.6445	0.7414	0.6993	0.8442	-0.6536	-0.3457	2.2961	2.4131
	6	0.6422	0.7215	0.6968	0.8321	-0.6519	-0.3491	2.2931	2.4164
	7	0.8686	2.1172	0.9047	2.2500	-2.2639	-0.8818	9.1098	3.8119
	8	0.8857	2.3479	0.9290	2.5710	-2.3922	-1.0531	9.4243	3.8061
	9	0.8784	2.2392	0.9180	2.4158	-2.3697	-1.0020	9.4760	3.8904

The logit transformed C-statistic was still skewed for scenarios 4 and 7 when there was large heterogeneity in α (setting 4: σ_α=1.0, Figure 6.4). However, across the range of scenarios and settings considered, the logit transformed C-statistic was more 'normally' distributed than the original C-statistic scale or other transformations considered. The logit scale is therefore a more appropriate scale to use for modelling the between-study distribution of C-statistics in a random-effects meta-analysis.

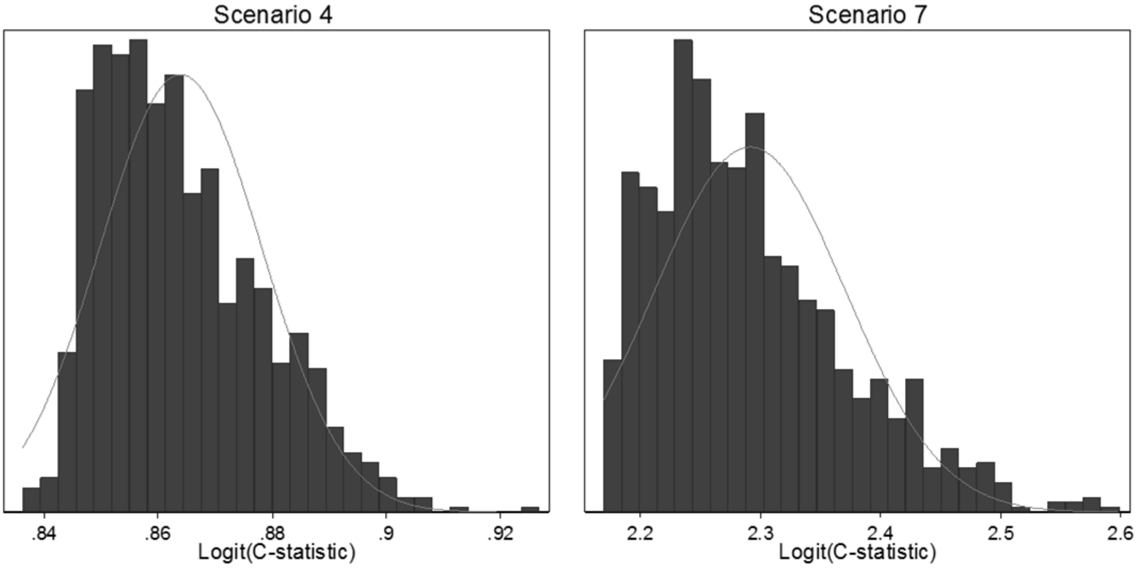


Figure 6.4: Histograms for logit(C-statistic) in scenarios 4 and 7 when heterogeneity in α is large (setting 4: σ_α=1.0).

Between-study distribution for expected/observed number of events

Heterogeneity in alpha (settings 2 to 4)

The ratio of expected and observed number of events (E/O) was centred on the ideal value of one when there was heterogeneity in α, but as σ_α increases, the SD of E/O also increased. E/O is a ratio therefore when the mean α relates to a population in which there are only a few events (as in scenarios 2, 5 and 8), a wider distribution is observed (Figure 6.5). There is more variability when the prevalence is low because small differences between the expected and observed number of events translate to larger relative differences than if the denominator was larger.

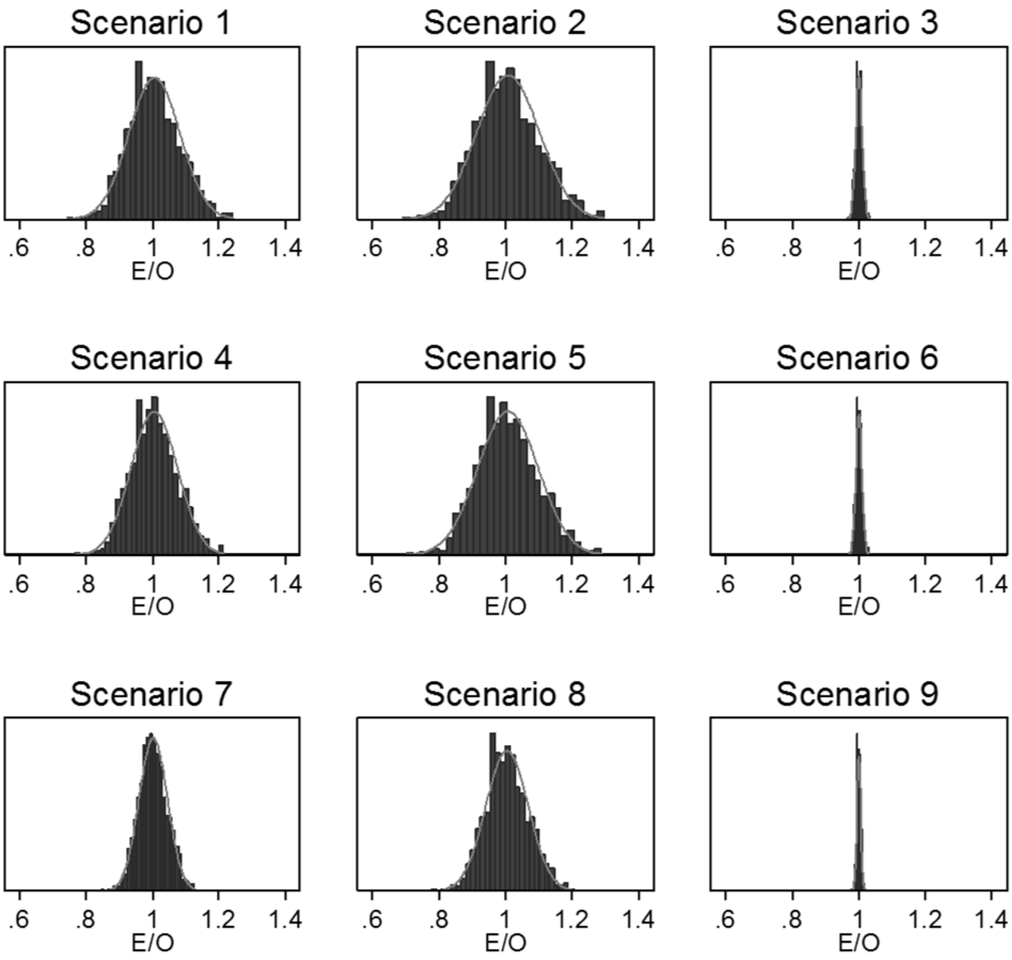


Figure 6.5: Histograms for E/O in all scenarios when there is little heterogeneity in *α* (setting 2: σ_α=0.1).

E/O was approximately normally distributed in all scenarios when there was little heterogeneity in *α* (setting 2: σ_α=0.1); however the distributions became skewed when there was moderate heterogeneity in *α* (setting 3: σ_α=0.5) and worse still when there was large heterogeneity in *α* (setting 4: σ_α=1.0). Distributions of E/O across different levels of heterogeneity in *α* (settings 2 to 4) are shown for scenario 1 in Figure 6.6, but distributions were similarly shaped for other scenarios. The estimated skewness for E/O was 1.32 and 3.00 for scenario 1 in setting 3: σ_α=0.5 and setting 4: σ_α=1.0 respectively.

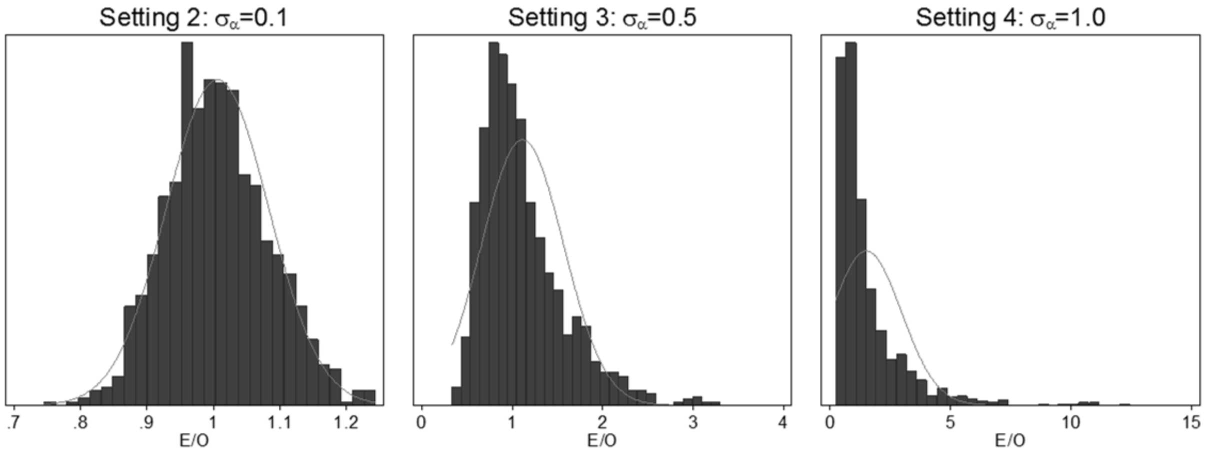


Figure 6.6: Histograms for E/O with different levels of heterogeneity in *α* in scenario 1 (settings 2 to 4).

Heterogeneity in beta (settings 5 to 7)

The distribution of E/O was skewed when there was heterogeneity in *β* (settings 5 to 7). For little heterogeneity in *β* (setting 5: σ_β=0.005), only scenarios 1 to 3 were skewed, which were the scenarios with a weak predictor. This is likely to be because heterogeneity in *β* was small relative to the *β* value in scenarios 4 to 9 with moderate or strong predictors (Figure 6.7).

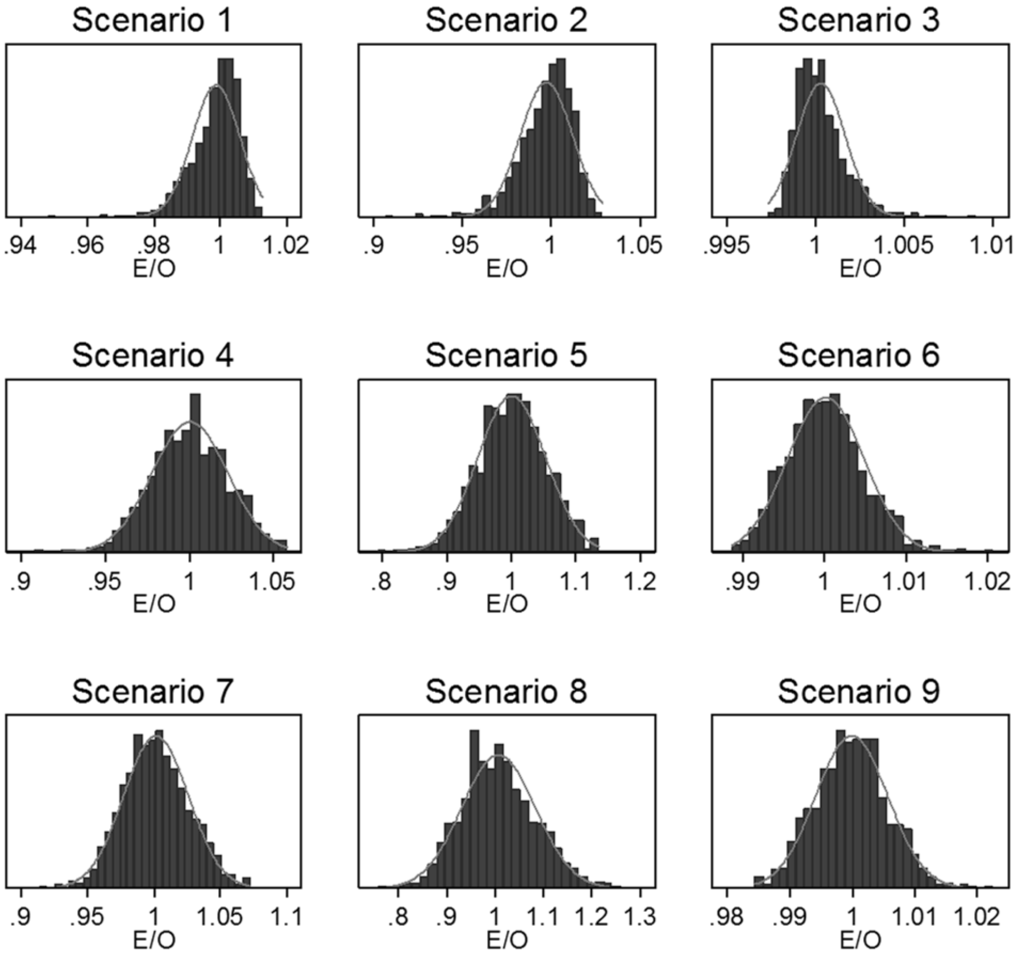


Figure 6.7: Histograms for E/O in all scenarios when there was little heterogeneity in β (setting 5: σ_β=0.005). Note: different x-axes used.

The distributions of E/O were extremely skewed for most scenarios when heterogeneity in β was moderate (setting 6: σ_β=0.02) or large (setting 7: σ_β=0.07, Figure 6.8). The distribution could be skewed in either direction but often appeared bounded close to one for scenarios 1 to 3 and bounded at different values for scenarios 4 to 6. Therefore, the peak of the distribution was at this boundary value.

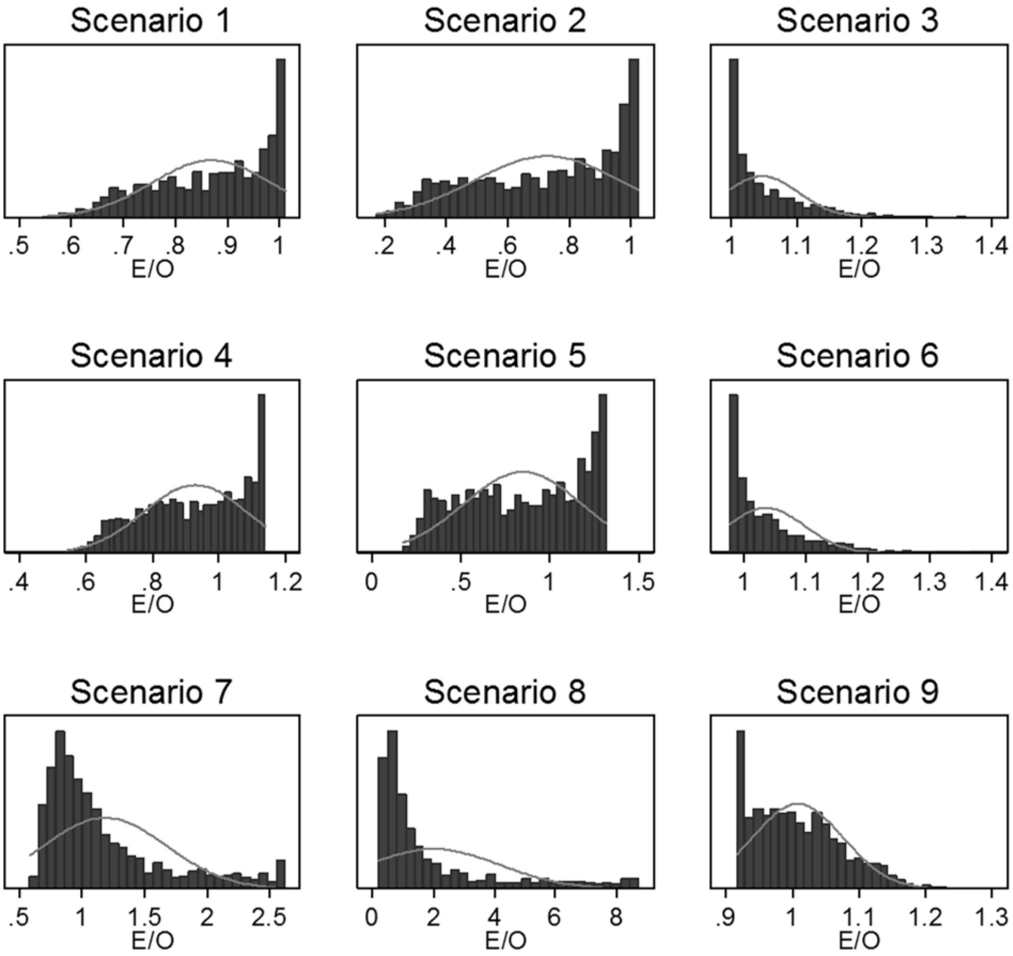


Figure 6.8: Histograms for E/O in all scenarios when heterogeneity in β was large (setting 7: σ_β=0.07). Note different x-axes used.

Transformations of E/O

A log transformation applied to E/O improved the shape of distributions when heterogeneity in α was moderate (setting 3: σ_α=0.5) or large (setting 4: σ_α=1.0), resulting in distributions that were closer to approximate normal distributions (Figure 6.9). However, the distributions remained skewed for scenarios 3, 6 and 9 where the average number of events was high. The natural log transformation resulted in distributions closer to the normal distribution compared to the square root transformation. Using the square root transformation did not improve the distributions in scenarios 3, 6 or 9 compared to the natural log transformation (see Appendix Figure E4.4 for comparison of scales).

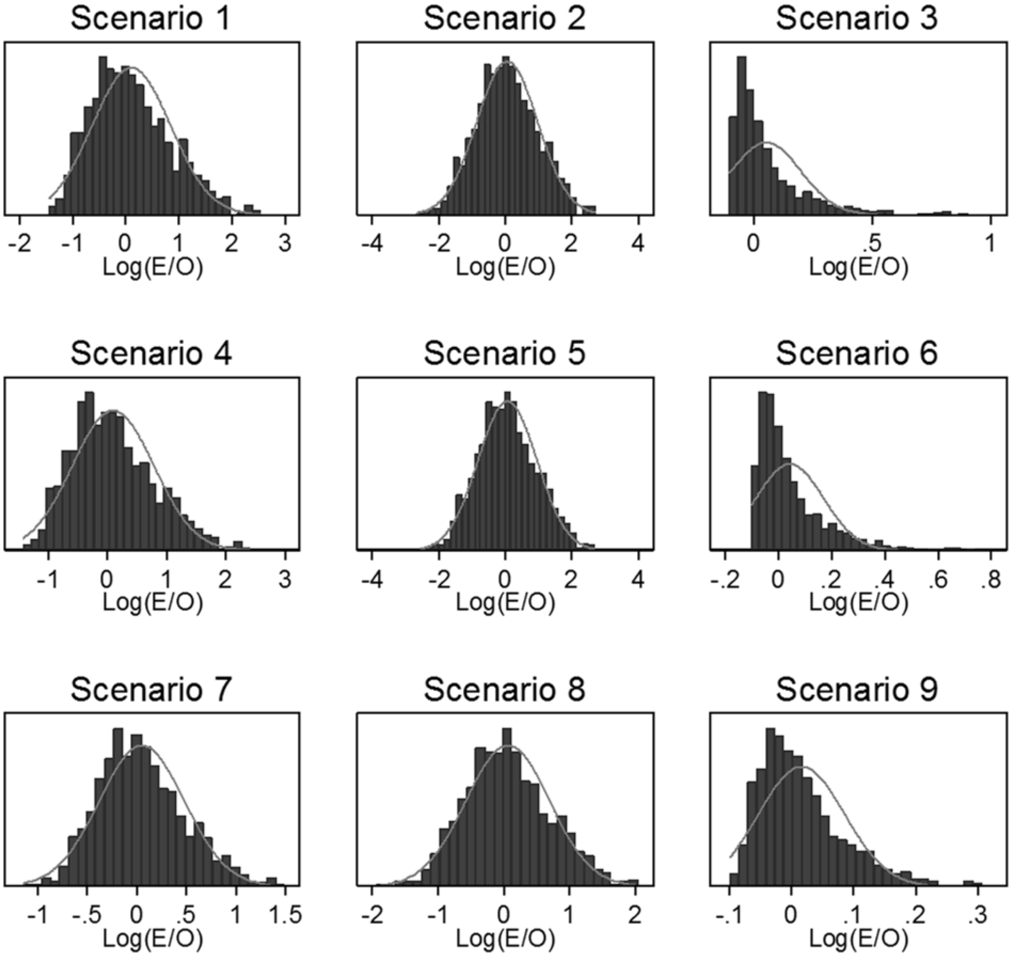


Figure 6.9: Histograms for log(E/O) in all scenarios when heterogeneity in *α* was large (setting 4: σ_α=1.0). Note different x axes used.

Using the log transformation for E/O also improved the shape of the distributions for scenarios 7 to 9 when there was moderate heterogeneity in *β* (setting 6: σ_β=0.02). The coefficient of skewness was between 0.20 and 0.67 for the log transformed E/O compared to skewness of between 0.26 and 1.31 for E/O (Table 6.9). However, the log transformation did not improve the distributions of E/O when heterogeneity in *β* was large (setting 7: σ_β=0.07), which remained very skewed in most scenarios (Figure 6.10). The square root distribution did not improve the distribution of E/O in this setting either (Appendix Figure E4.5).

Table 6.9: Summary statistics and distribution skewness and kurtosis for E/O and log transformed E/O in two simulation settings for heterogeneity in *β* (setting 6 and 7).

Simulation setting	Scenario	Mean		Median		Skewness		Kurtosis	
		E/O	Log(E/O)	E/O	Log(E/O)	E/O	Log(E/O)	E/O	Log(E/O)
Setting 6: moderate heterogeneity in β σ_{β}=0.02	1	0.9792	-0.0216	0.9908	-0.0093	-1.9947	-2.1878	8.2741	9.7516
	2	0.9570	-0.0469	0.9809	-0.0193	-2.1009	-2.6493	8.8671	13.6492
	3	1.0046	0.0046	1.0017	0.0017	3.1208	3.0277	18.7377	17.6496
	4	0.9993	-0.0040	1.0023	0.0023	-0.2750	-0.4568	2.4702	2.7713
	5	0.9955	-0.0230	1.0072	0.0072	-0.2989	-0.7601	2.4095	3.4054
	6	1.0032	0.0030	0.9993	-0.0007	1.1806	1.1100	5.0599	4.7260
	7	1.0202	0.0140	1.0008	0.0008	1.0375	0.6697	4.6344	3.5623
	8	1.0872	0.0353	1.0097	0.0097	1.3069	0.3971	5.2668	2.9890
	9	1.0001	-0.0002	0.9993	-0.0007	0.2612	0.1956	3.0482	2.9641
Setting 7: large heterogeneity in β σ_{β}=0.07	1	0.8696	-0.1490	0.8921	-0.1142	-0.5345	-0.7286	2.1408	2.5326
	2	0.7278	-0.3833	0.7740	-0.2562	-0.4290	-0.9259	1.8868	2.9171
	3	1.0473	0.0449	1.0257	0.0253	1.5889	1.4389	5.6815	4.9001
	4	0.9291	-0.0882	0.9433	-0.0584	-0.3269	-0.5632	1.9389	2.2861
	5	0.8459	-0.2649	0.8691	-0.1403	-0.1972	-0.7716	1.7144	2.5913
	6	1.0353	0.0329	1.0127	0.0126	1.5319	1.3757	5.3263	4.5887
	7	1.1852	0.0993	1.0091	0.0091	1.3655	0.7872	3.9896	2.6973
	8	2.0058	0.2151	1.0374	0.0368	1.6434	0.4816	4.6254	2.1661
	9	1.0077	0.0054	0.9973	-0.0027	0.6711	0.5319	2.9416	2.6232

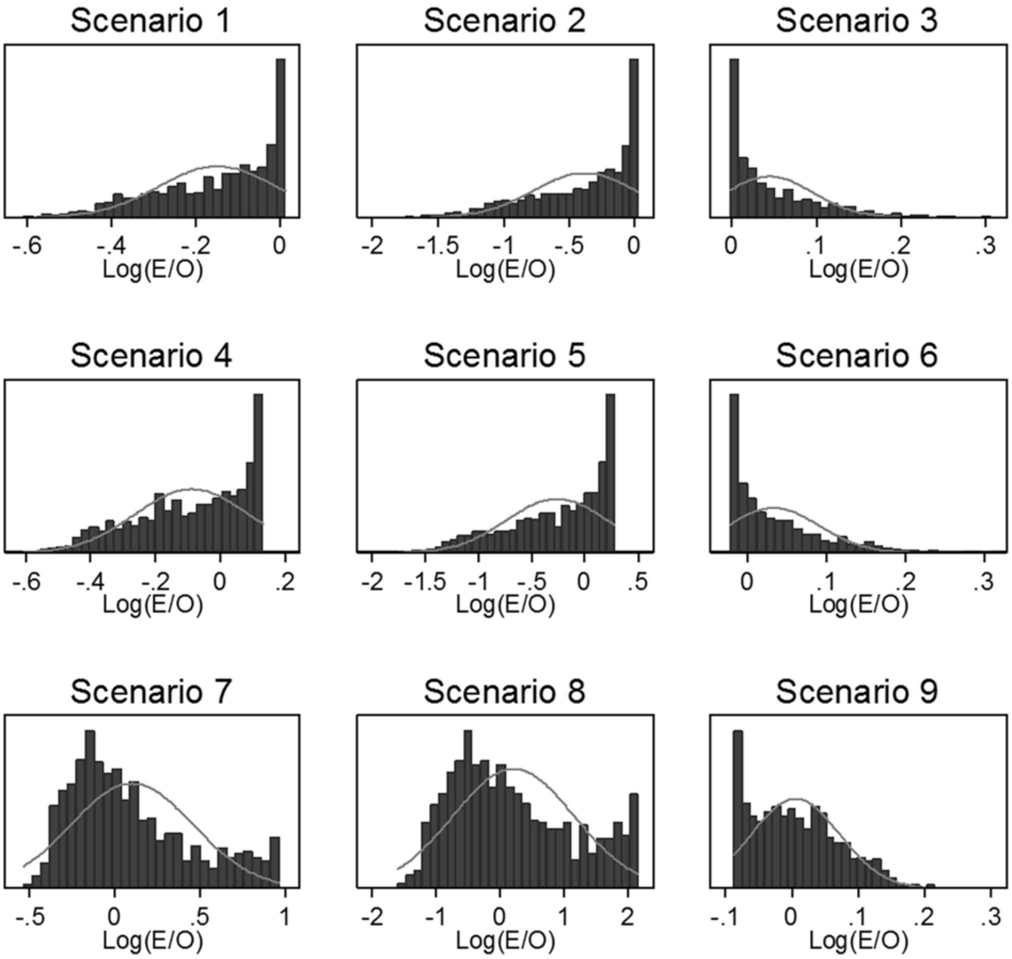


Figure 6.10: Histograms for log(E/O) in all scenarios when heterogeneity in *β* was large (setting 7: *σ*$_{\beta}$=0.07). Note different x axes used.

Between-study distribution for calibration slope

The between-study distribution of the calibration slope was approximately normal in all scenarios and in all simulation settings for heterogeneity in *α* or *β* (settings 2 to 7). The SD of the distribution was larger when the average *β* was small (weak predictor in scenarios 1, 2 and 3). Figure 6.11 shows the distributions of calibration slope in all scenarios when there was large heterogeneity in *α* (setting 4: σ_{α}=1.0). The shape of the calibration slope in Figure 6.11 is representative of all settings for heterogeneity in *α* (settings 2 to 4). Heterogeneity in *α* did not affect the distribution of calibration slope so the distributions are almost identical to setting 1: no heterogeneity for *α* or *β*, which showed only sampling error.

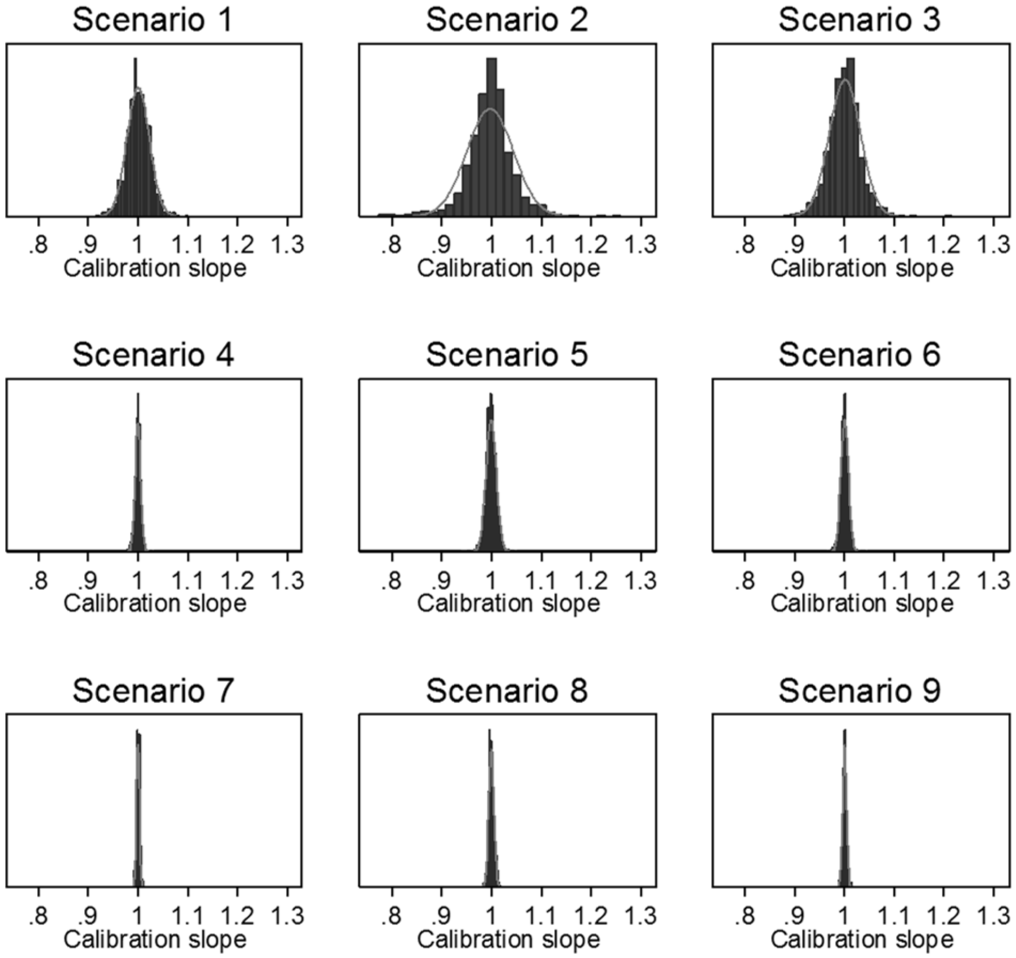


Figure 6.11: Histograms for calibration slope in all scenarios when heterogeneity in α is large (setting 4: σ_α=1.0).

As heterogeneity in β increases (settings 5 to 7), the SD of the distribution of calibration slope also increased (Table 6.10). Figure 6.12 shows that the width of the distributions of calibration slope were very large with SD=6.92 in scenarios 1 to 3 (weak predictor) when heterogeneity in β was large (setting 7: σ_β=0.07).

Table 6.10: Means and standard deviations for calibration slope for different levels of heterogeneity in *β* (settings 5 to 7).

Scenario	Mean (SD) for calibration slope		
	Setting 5: σ_β=0.005	Setting 6: σ_β=0.02	Setting 7: σ_β=0.07
1	0.9842 (0.4939)	0.9375 (1.9770)	0.7805 (6.9217)
2	0.9831 (0.4939)	0.9357 (1.9767)	0.7793 (6.9237)
3	0.9849 (0.4953)	0.9374 (1.9776)	0.7815 (6.9237)
4	0.9965 (0.1097)	0.9860 (0.4392)	0.9515 (1.5383)
5	0.9963 (0.1098)	0.9861 (0.4395)	0.9511 (1.5383)
6	0.9965 (0.1101)	0.9860 (0.4396)	0.9515 (1.5382)
7	0.9988 (0.0342)	0.9952 (0.1444)	0.9848 (0.4774)
8	0.9988 (0.0342)	0.9955 (0.1362)	0.9849 (0.4770)
9	0.9990 (0.0343)	0.9958 (0.1364)	0.9848 (0.4772)

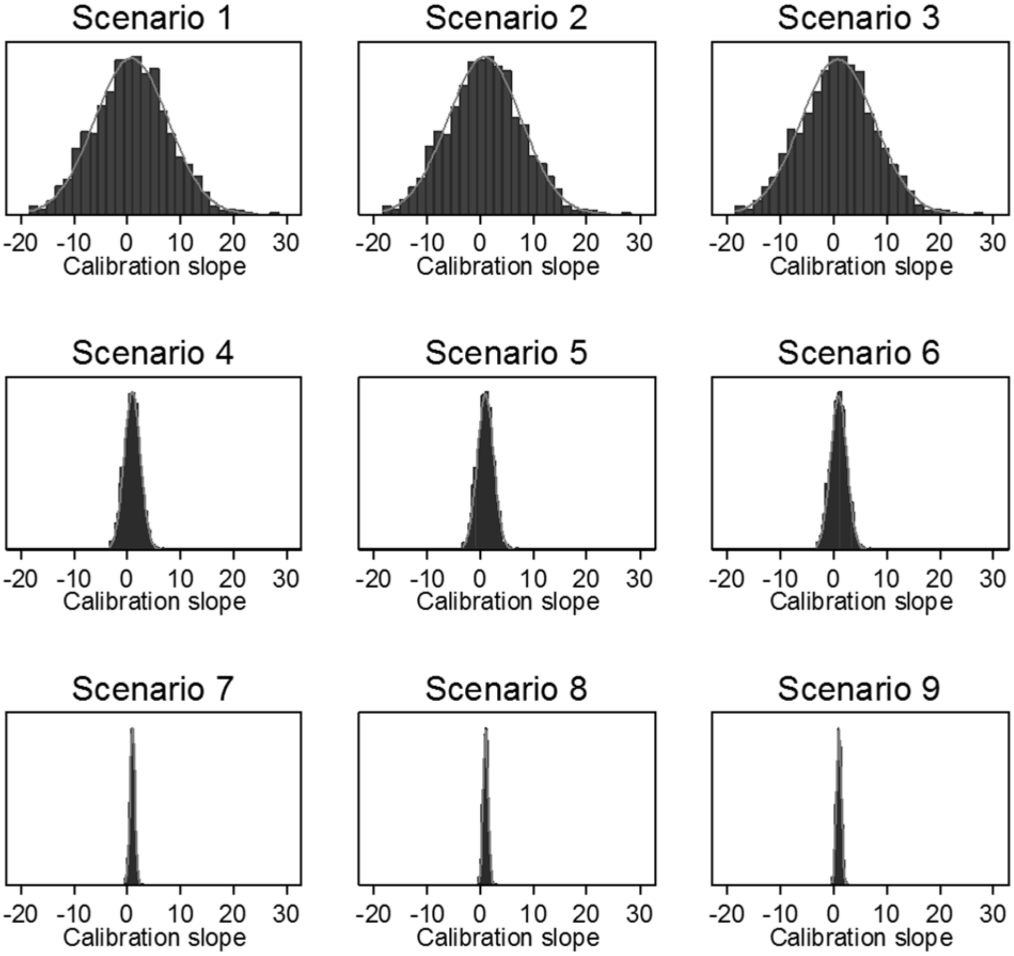


Figure 6.12: Histograms for calibration slope in all scenarios when heterogeneity in *β* is large (setting 7: σ_β=0.07).

Between-study distribution for calibration-in-the-large

Calibration-in-the-large was approximately normally distributed in all settings for heterogeneity in *α* (settings 2 to 4). The SD for the calibration-in-the-large distribution is approximately the same in different scenarios but increases as heterogeneity in *α* increases (Figure 6.13).

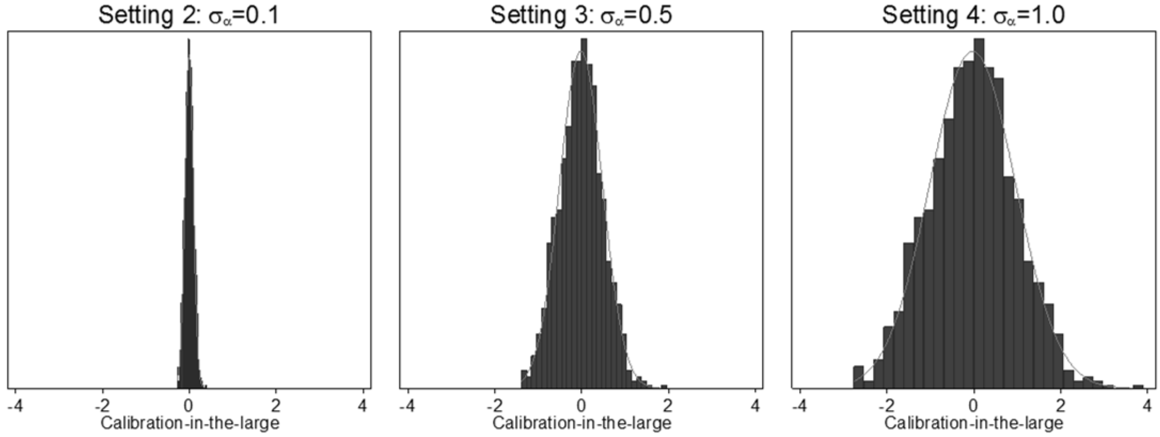


Figure 6.13: Histograms for calibration-in-the-large across different levels of heterogeneity in *α* (settings 2 to 4) for scenario 1.

The distribution of calibration-in-the-large was skewed (in either direction) when there was heterogeneity in *β*, but only when the heterogeneity was large relative to the value of *β*. Therefore, distributions for scenarios 1 to 3 (weak predictor) were skewed when there was little heterogeneity in *β* (setting 5: σ_β=0.005), distributions for scenarios 1 to 6 were skewed when there was moderate heterogeneity in *β* (setting 6: σ_β=0.02) and distributions for all scenarios were skewed when there was large heterogeneity in *β* (setting 7: σ_β=0.07). The distributions of calibration-in-the-large when there is moderate heterogeneity in *β* (setting 6: σ_β=0.02) are shown in Figure 6.14. The distributions appear to be bounded at different values for scenarios 1 to 6.

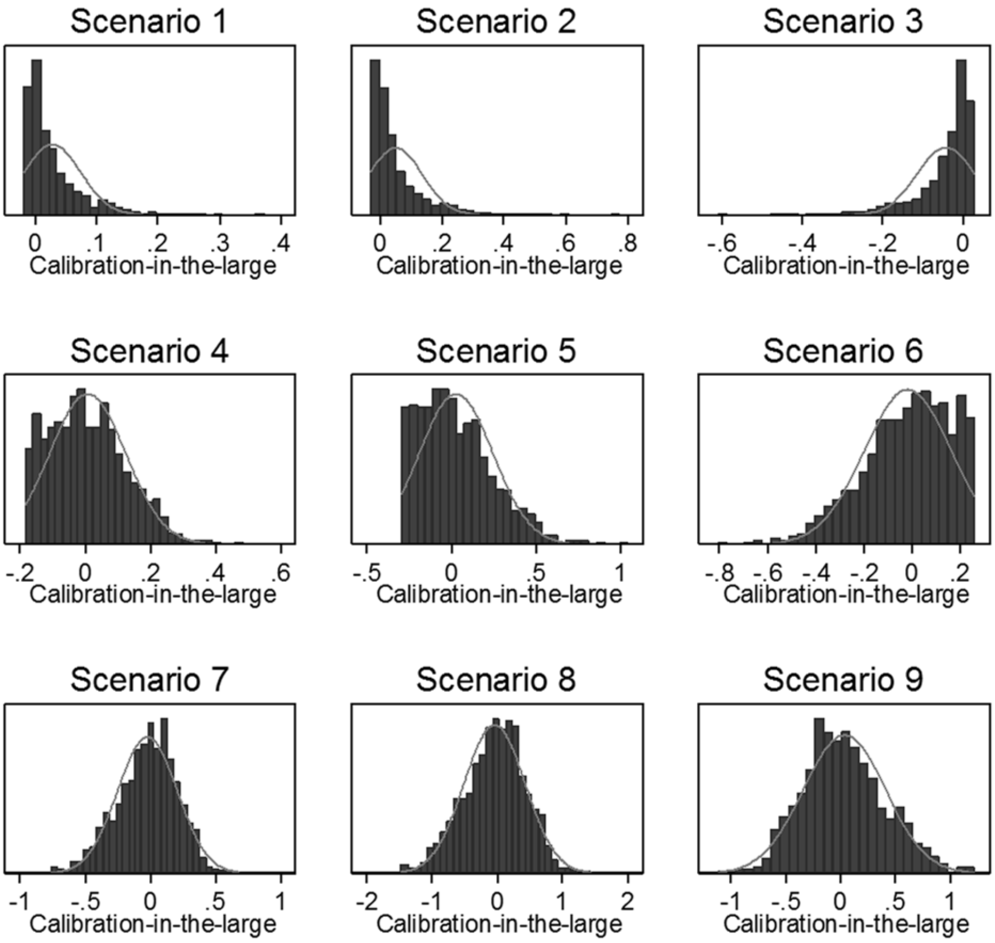


Figure 6.14: Histograms for calibration-in-the-large in all scenarios when heterogeneity in *β* is moderate (setting 6: σ_β=0.02). Note different x axes used.

6.3.3 Extensions

Extension 1: Limiting the age range to between 18 and 100 years

Limiting the range of values for age affected the performance statistics in different ways depending on the scenario. When the predictor was weak (scenarios 1 to 3), the distribution means and SDs were very similar to the original simulation setting 1 (no heterogeneity in *α* or *β*). The largest differences between limiting the age range and not limiting the age range, were seen in distributions when the predictor was strong (scenarios 7 to 9) where the width of the distributions increased or decreased for the performance statistics but distributions still appeared approximately normal. Figure 6.15 shows the histograms for all four performance statistics with restricted and unrestricted age ranges for scenario 7 as an example.

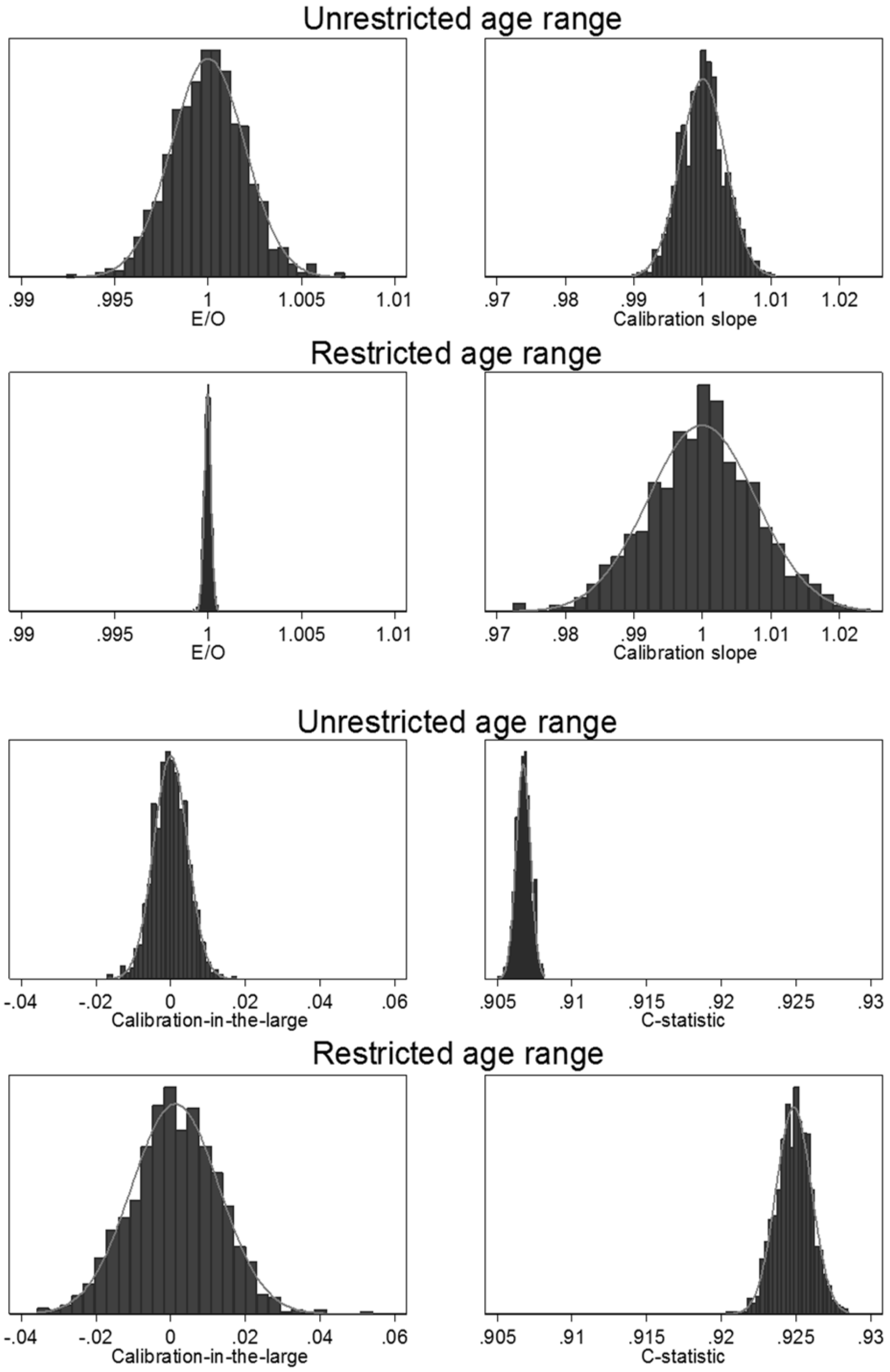


Figure 6.15: Histograms for performance statistics in scenario 7, with data generated using the original age distribution N(60, 17.6^2) and age restricted to between 18 and 100 years.

Extension 2: Varying the distribution of age values across studies

Examples of study-specific age distributions with varying mean and SD are shown in Figure 6.16. Varying the mean and SD for the study-specific age distributions did not have any effect on measures of calibration (E/O, calibration slope or calibration-in-the-large) compared to distributions of performance statistics with restricted age range (between 18 and 100 years discussed in the previous section). For these calibration statistics, the distributions were very similar to the distributions of the statistics when the mean and SD of age were fixed for all studies (Figure 6.17).

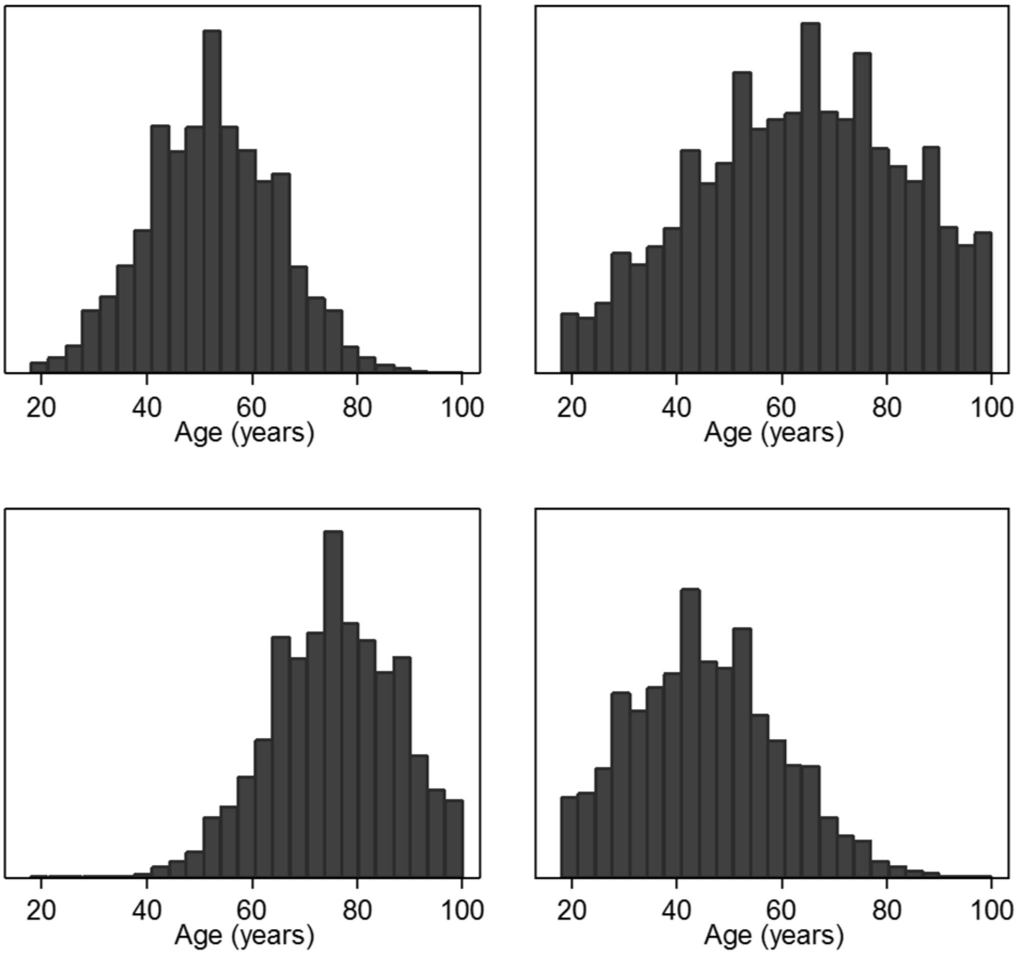


Figure 6.16: Examples of study-specific age distributions when the mean and SD of age were allowed to vary across studies but were still restricted between 18 and 100 years.

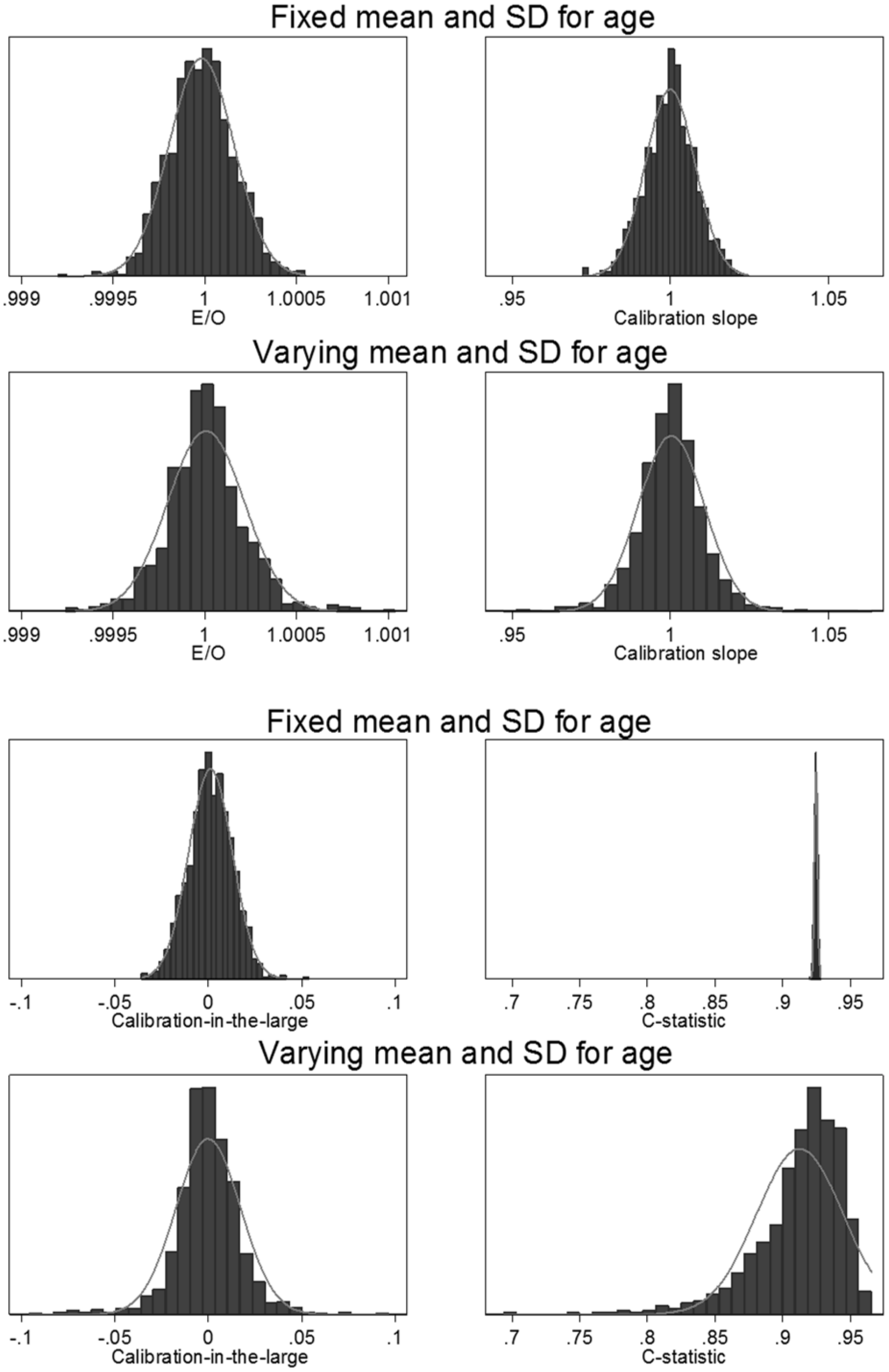


Figure 6.17: Histograms for performance statistics comparing fixed mean and SD for age with random-effects on the mean and SD for age in scenario 7.

Unlike measures of calibration, the width of the between-study C-statistic distribution increased when the age distribution was allowed to vary across studies. This is plausible as the C-statistic directly relates to case mix variability. This was observed in all scenarios and distributions were more skewed for strong predictors in scenarios 7 and 8. The C-statistic distribution was only slightly skewed for scenarios with a weak or moderate predictor (scenarios 1 to 6) and the logit transformation improved the shape of the distribution in scenarios 7 and 8 (skewness and kurtosis in Table 6.11). Note that scenario 9 was defined to have a high number of events and a strong predictor, therefore with varying age distributions, computation problems were encountered and performance statistics could not be calculated for all studies. This scenario was excluded as it is likely that for some distributions of age in studies, all patients would have the event.

Table 6.11: Summary statistics and distribution skewness and kurtosis for original and logit transformed C-statistic when age distributions vary across studies.

Scenario	Mean		Median		Skewness		Kurtosis	
	C-statistic	Logit C-statistic	C-statistic	Logit C-statistic	C-statistic	Logit C-statistic	C-statistic	Logit C-statistic
1	0.5443	0.1776	0.5450	0.1804	-0.6439	-0.6359	3.6332	3.6144
2	0.5444	0.1783	0.5456	0.1830	-0.7048	-0.6974	3.4317	3.4167
3	0.5444	0.1780	0.5450	0.1805	-0.5158	-0.5088	3.1124	3.1001
4	0.6858	0.7837	0.6898	0.7994	-0.8403	-0.7066	3.8160	3.5098
5	0.6820	0.7659	0.6867	0.7848	-0.8249	-0.6895	3.8448	3.5190
6	0.6879	0.7937	0.6916	0.8076	-0.8647	-0.7125	3.9810	3.6176
7	0.9129	2.4064	0.9198	2.4398	-1.6267	-0.5105	7.5334	3.4738
8	0.8991	2.2329	0.9062	2.2676	-1.7110	-0.7289	7.5627	3.7947

Extension 3: Including an additional predictor and interaction

When the additional predictor included in the data generating model was categorical and an interaction was included between age and the additional predictor, the performance of the 'true' prediction model (defined in Table 6.4, excludes additional predictor and interaction term) deteriorated as the strength of the interaction increased. However, the width of the between-study distribution of performance statistics was not affected, only the average performance. This was the case for all four performance statistics.

When the additional predictor was continuous, for each of the calibration slope, calibration-in-the-large and the C-statistic, the width of their between-study distribution increased slightly as the interaction effect increased (difference between the maximum and minimum values given in Table 6.12). The opposite effect was seen for E/O, where the difference between maximum and minimum values decreased slightly as the interaction effect increased. However, all the between-study distributions still appeared approximately normal and the variances remained very small, mostly representing sampling error in these distributions. Figure 6.18 shows distributions of all performance statistics in settings for missing 1 to 3 for scenario 4.

Table 6.12: Range of values for performance statistics in scenarios 4 to 6 when an additional predictor and interaction effect were used to generate the data but not included in the prediction model.

		Difference between min and max values		
Performance statistic	Scenario	Small interaction effect	Moderate interaction effect	Large interaction effect
E/O	4	0.0027	0.0021	0.0016
	5	0.0066	0.0025	0.0021
	6	0.0005	0.0003	0.0002
Calibration slope	4	0.0362	0.0561	0.0851
	5	0.0263	0.0391	0.0512
	6	0.2005	0.3135	0.5154
Calibration-in-the-large	4	0.0235	0.0400	0.0553
	5	0.0215	0.0225	0.0275
	6	0.1525	0.2081	0.3312
C-statistic	4	0.0065	0.0078	0.0078
	5	0.0041	0.0045	0.0046
	6	0.0367	0.0441	0.0464

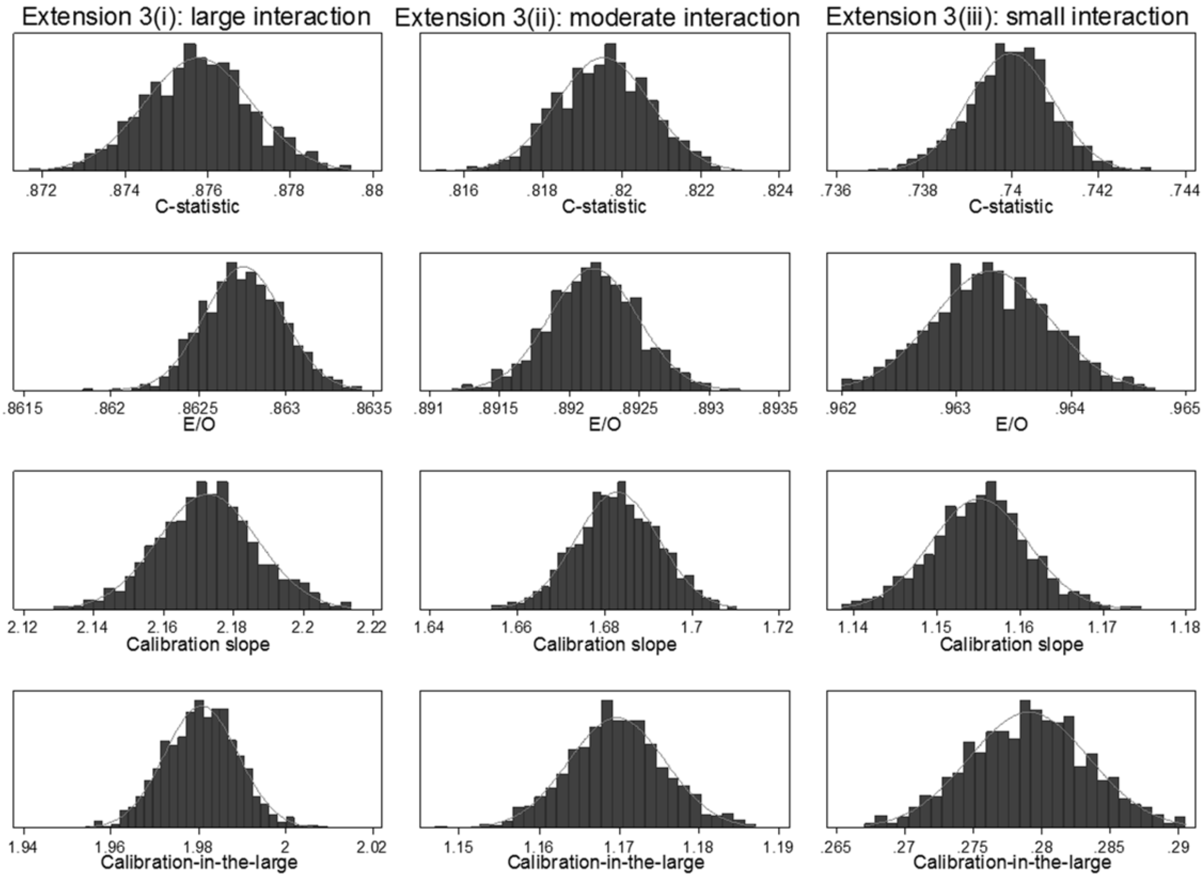


Figure 6.18: Distributions of all performance statistics for different missing (continuous) predictor settings (extension 3(i) to 3(iii)) for scenario 4. Note different x axes used.

6.4 Discussion

Between-study normality of the C-statistic and other performance statistics was assumed in the meta-analysis methods proposed in Chapter 5. This assumption of normality is especially important when deriving prediction intervals for the potential performance of the assumed prediction model in new populations or settings similar to those included in the meta-analysis. Therefore, it was important to verify whether this normality assumption is valid and, if not, when it is likely to break down.

The aim of this study was therefore to assess if the normality of true performance statistics is a reasonable assumption and if not, whether a simple transformation of the performance statistic provides a more normally distributed scale for meta-analysis. Through simulation, the distribution of true performance statistics was observed in different settings in which heterogeneity was introduced either in the baseline risk (intercept), the predictor effect (beta), the distribution of predictor values, or by including an additional predictor and interaction effect (that are not included in the assumed prediction model). Key findings from the defined simulation settings are given in Box 6.2.

Box 6.2: Key findings from simulation study looking at true between-study distributions of performance statistics.

C-statistic:

- The C-statistic distribution was most skewed when there was heterogeneity in the predictor effect (β) and when the distribution of the predictor values varied across studies (mean and SD for distribution of predictor were also sampled from normal distributions for each study), therefore varied across studies. However, the logit transformation greatly improved the shape of the C-statistic in these situations. Distributions were still slightly skewed for scenarios with a strong predictor (scenarios 7 to 9) when heterogeneity in β was large, and thus large heterogeneity in predictor effects is undesirable.
- Heterogeneity in the baseline risk (α) had very little effect on the C-statistic in that the distribution remained very narrow indicating mostly sampling error (as expected due it measuring discrimination which relates to the predictor effect). However when the heterogeneity in α was large and the predictor effect (β) was moderate to strong (scenarios 4 to 9), the distribution was slightly skewed. Neither log, logit nor arcsine transformations improved this.

Box 6.2 continued…

Expected/observed number of events:

- Heterogeneity in *α* increased the width of the E/O distribution but it remained approximately normal when heterogeneity in *α* was small. As heterogeneity in *α* increased, the distribution of E/O increasingly got more skewed.
- E/O was skewed when heterogeneity in *β* was moderate to large.
- Using the log transformation improved the E/O distribution (towards normality) when there was heterogeneity in *α*, except when the baseline risk was high (scenarios 3, 6 and 9). The log transformation also improved the shape of the E/O distribution when heterogeneity in *β* was moderate but not when heterogeneity in *β* was large.
- The width of the distribution was wider when the baseline risk was low (scenarios 2, 5 and 8) as would be expected. E/O is a ratio with a small denominator when the baseline risk is low and could be misleading in such situations.

Calibration slope:

- Calibration slope was approximately normal in all settings considered. Heterogeneity in *α* or *β* only affected the mean value and SD but not the shape of the distribution.

Calibration-in-the-large:

- Calibration-in-the-large was approximately normal when there was heterogeneity in *α*, although the width of the distribution increased with increasing heterogeneity in *α*.
- Calibration-in-the-large was skewed when heterogeneity in *β* was large relative to the size of *β*, again emphasizing that large heterogeneity in predictor effects is undesirable.

Heterogeneity in β, especially when large, resulted in skewed distributions for all performance statistics except the calibration slope. Large heterogeneity in β was an extreme setting, especially for scenarios with weak or moderate predictor effects such as in scenarios 1 to 6. This level of heterogeneity in β is unlikely to occur in reality as a predictor is unlikely to be considered useful in a model if the effect could vary so much that it is not predictive of outcome in some studies (β=0) or could potentially be predictive in the opposite direction in some studies (predictive of non-events, β<0 when β>0 in other studies). Rather than a predictor effect that varies that much across studies, it is more likely that there are other unknown or unmeasured predictors that would explain differences in studies or populations but are not appropriately being accounted for in the prediction model. What this highlights is the need to have reliable data in which to estimate the predictor effects and also to ensure that the prediction model includes as many relevant predictors as possible, so that included predictors in the model (or at least their combination) gives little heterogeneity in their (combined) effect.

An article published by van Klaveren et al. considered meta-analysis of C-statistics on the probability scale and log-odds (logit) scale and concluded that the probability scale was most appropriate for their data based on checking residuals.[62] However, such residuals are an amalgamation of within-study sampling error and between-study error. The simulation study in this chapter used study sample sizes of 500000 in order to reduce within-study error to a tiny amount, in order to reduce this issue considerably. The simulations then found that the logit transformed C-statistic was more normally distributed than the C-statistic on the original scale in many settings, even when there was large heterogeneity in the predictor effect. The true prediction models considered in this chapter and the van Klaveren et al. article differ, as van Klaveren et al. used a Cox proportional hazards model, whereas this chapter considered logistic models. However, the recommendation differs as this chapter recommends that the logit scale is preferable for meta-analysis of the C-statistic.

Another simulation study by Austin and Steyerberg investigated the relationship between the C-statistic and a continuous explanatory variable. The authors showed that the C-statistic was dependent on both the log odds ratio and variance of the explanatory variable.[203] The findings in this chapter support this as when the age distribution was fixed, the C-statistic increased as the predictor effect increased (relating to the log odds ratio). The distribution of the C-statistic was also wider when the age distribution varied across studies. Therefore it is likely that when the age distribution was narrower, the C-statistic was lower than when the age distribution was wider.

Key recommendations for the scale on which to pool performance statistics in a meta-analysis are given in Box 6.3.

Box 6.3: Recommendations for pooling performance statistics in a meta-analysis.

The following scales should be used for modelling the between-study distribution when pooling performance statistics in a random-effects meta-analysis and subsequently deriving prediction intervals.

- Use logit transformed C-statistics
- Use log transformed ratio of expected/observed (E/O) number of events.
- Use original scale for calibration slope
- Use original scale for calibration-in-the-large

But caution should still be taken in assuming normality on these scales in situations where the predictor effect (β) varies a lot across studies. This may indicate that important predictors have not been included in the prediction model, and the normality assumption between-studies is likely to be less reliable (and therefore calculation of 95% prediction intervals following random effects meta-analysis may be unreliable under a normal assumption).

6.4.1 Findings in relation to previous chapter

In Chapter 5, the log transformation was used to pool E/O. This simulation study supports the use of the log transformation as log(E/O) was more normally distributed than E/O when there was heterogeneity in α or β and the log transformation performed better than the square root transformation. However, there were a few simulation settings in which the distribution of log(E/O) remained skewed, such as when the baseline risk was high and there was heterogeneity in α or when heterogeneity in β was large. These are extreme settings and therefore this is unlikely to be a problem in reality. Therefore using the log transformed E/O is considered the most suitable scale for meta-analysis.

In Chapter 5, C-statistics were pooled on the original scale whereas the simulation study showed the logit transformation to be more suitable in settings where there is heterogeneity in β or when the predictor distribution varies from study to study. It would be useful to check whether using the logit scale for pooling C-statistics in Chapter 5 would have any impact on the overall results and prediction intervals. Unfortunately, this was not possible to do as the performance statistics were provided by other authors and standard errors and correlations were not immediately available on the logit scale. Further research will look to update the meta-analysis of C-statistics from Chapter 5 accordingly to examine if and how conclusions (such as prediction intervals and choice of implementations strategies) change if the logit scale is used.

6.4.2 Limitations and further research

This chapter considered a logistic regression prediction model with a single predictor included and heterogeneity was introduced by assuming the study-specific intercept or predictor effect came from a normal distribution. This chapter highlights some important findings about which scale is suitable for pooling performance statistics using meta-analysis methods, however further research is required to further check the normality assumption for meta-analysis. For example, the calibration slope looks approximately normally distributed in all scenarios and under all settings for heterogeneity in either α or β. This may be because of the way in which heterogeneity was introduced as α or β was drawn from a normal distribution. Therefore further simulations in which the distributions of study-specific intercepts and predictor effects are sampled from a different distribution would be required to check if performance statistics remain normally distributed or if the shape of the distribution is related to the way in which heterogeneity was introduced.

In this chapter, a binary outcome was selected to keep the model simple for the simulation study, therefore only a logistic prediction model has been considered. It would be useful to extend this to a survival model to check if the findings are robust against a different model being used to develop the prediction model. It would therefore also be possible to check the distributions of additional performance statistics such as the D-statistic used in Chapter 5.

6.5 Conclusions

This simulation study has investigated the between-study distributions for performance statistics in a variety of settings in which different levels of heterogeneity in baseline risk or predictor effects have been considered. The normality assumption of the between-study distributions is made when pooling performance statistics in a meta-analysis and is important when predicting how well a model is likely to perform in a new population or setting.

Where between-study distributions were not approximately normally distributed, transformations were considered and recommendations for the scale on which to pool various performance statistics have been made. This chapter therefore adds important findings to the meta-analysis literature regarding clinical prediction models, and builds on the work from earlier chapters. The next chapter brings together the key findings from all the previous chapters, for a final, broader discussion on their implications and contributions.

CHAPTER 7: DISCUSSION

7.1 Overview of thesis

Prognosis is an important area of research as it aims to understand, explain and predict the risk of a future outcome in patients with a particular disease or condition. In particular, this thesis has focused on prognostic modelling which combines multiple prognostic factors to predict risk for patients and can be used to help inform clinicians when advising their patients on the likely course of their disease or condition.[8] Knowing the predicted risk of the outcome (or conversely the probability of survival) for an individual can also help both the clinician and patient decide on an appropriate treatment plan, in order to optimise treatment benefit and reduce unnecessary treatment.[8] Thus the core aim of prognosis research is to improve patient health outcomes, yet unfortunately many prognostic models do not make their way into clinical practice.[3,204] Several papers over recent years have highlighted the need to improve the methodology and reporting of prognostic model studies,[1,18,81] and the work presented in this thesis has contributed towards that goal.

The overall aims of this thesis were to apply, develop and evaluate novel statistical methods for prognosis research. In particular, early chapters (Chapters 2 to 4) aimed to apply and evaluate the use of flexible parametric survival models as an alternative to Cox models when developing prognostic models using survival data. The last two chapters of this thesis (Chapters 5 and 6) focused on validation of prognostic models, by developing meta-analysis methods that extend the internal-external cross-validation approach by synthesising performance statistics across multiple ‘external’ validation studies. A short summary of the chapters is given below.

7.1.1 Summary of thesis chapters

The chapters contained a mixture of clinical application and methodology development. Chapter 2 applied novel flexible parametric survival models to registry data from osteoarthritis patients that received a hip replacement. The clinical aim was to ascertain whether there were differences in mortality between patients receiving different procedure types. Previous literature in this field had concentrated mainly on revision rates but mortality should also be a crucial patient outcome of interest, and the work raised interesting findings (see below). This application also helped identify some of the statistical advantages of using flexible parametric models rather than Cox proportional hazards models for prognostic modelling, especially in regard to individualised, absolute risk prediction. Given the advantages observed in Chapter 2, a literature review of published prognostic models was conducted in Chapter 3, where the main aims were to establish (i) if and how the baseline hazard was being modelled; (ii) how absolute risk prediction was being presented; and (iii) how the developed models would enable absolute risk predictions for new patients. This review found that the Cox model was used to develop all models in the included articles, and thus there is a need to promote the use of novel methods that actually model the baseline hazard, such as the Royston-Parmar flexible parametric model. Chapter 4 therefore used the Royston-Parmar approach to develop a prognostic model for advanced pancreatic cancer using data from two randomised clinical trials. In doing so, several challenges were identified in the development of prognostic models, including general issues and issues relating to the use of clinical trials data specifically.

The last two chapters then focused on methods relating to the validation of prognostic models and in particular the setting in which several studies are available for development and validation of a prognostic model. Performance statistics were provided by other authors on request, following their use of internal-external cross-validation in two clinical areas, DVT and breast cancer.[149,176] In Chapter 5, random-effects meta-analysis was proposed to provide a summary measure of each performance statistic across the multiple studies and also to provide an estimate of the between-study variance of each performance statistic. Implementation strategies suggested by Debray et al.[149] for different intercepts in the validation study were compared in terms of average performance and heterogeneity in performance. Multivariate meta-analysis was also implemented in which two or more performance statistics are pooled in the same meta-analysis, utilising estimates of within-study correlations between performance statistics. Emphasis was placed on using the results of the meta-analyses to predict model performance in a new but similar setting. This was done by calculating prediction intervals or, in the multivariate setting, joint prediction ellipses. Chapter 5 also proposed how the results of a multivariate meta-analysis (matrices of pooled estimates and variance-covariance matrices) can be used to calculate the predicted probability that a model will have adequate performance when used in practice, according to some pre-defined criteria. This allows model implementation strategies to be ranked, to help ascertain the best approach to take.

The focus of Chapter 5 was on predicting performance of a model in a new study or setting, yet a fundamental assumption in predicting the performance in new studies is that the ‘true’ performance statistics are normally distributed across studies. Chapter 6 therefore aimed to verify if the normality assumption was reasonable for the true between-study distributions of performance statistics in a variety of settings, and examined the best scales to achieve this.

7.1.2 Publications from this thesis

The work in this thesis has led to a publication in the BMJ in 2012 on mortality rates in patients that received hip replacements (Chapter 2). Another article detailing the methods proposed in Chapter 5 (multivariate meta-analysis for pooling performance statistics following the IECV approach) has also been accepted for publication in the Journal of Clinical Epidemiology in 2015. Further articles arising from Chapters 3, 4 and 6 will be drafted and submitted over the coming year.

7.1.3 Areas of contribution to the field of prognosis research

This thesis has contributed to both applied and methodology research. Applied research has been carried out in the following clinical research areas:

- **Total hip replacements or resurfacing**

  Mortality rates and revision rates were compared between patients that had cemented and uncemented total hip replacements. This study found small but statistically significant differences in absolute patient survival and implant survival between the cemented and uncemented procedure groups after adjusting for other factors. Patient survival was slightly better for uncemented procedures and implant survival was slightly better for cemented procedures. Comparing both these procedures to Birmingham hip resurfacing in men aged under 55 years, showed that absolute survival probabilities over time were slightly higher in patients that had a BHR. However, differences between procedures were very small in terms of differences in absolute survival probabilities and residual confounding is likely to be present. Therefore further research was recommended to examine these mortality findings further, which has prompted further work by others.[103]

- **Advanced pancreatic cancer**

  A prognostic model was developed for patients with advanced pancreatic cancer. Overall model performance was only assessed internally and showed good performance up to 6 months but calibration of the model got progressively worse at later time points. Discrimination was also only moderate, and thus further work needs to include additional predictors in an updated model. Such work should incorporate the predictors identified in this chapter, and utilise the flexible parametric modelling approach that was shown to be useful over and above the Cox model. If external data becomes available, the performance of the developed model should be examined, possibly considering short term prognosis (up to 6 months) which may still be clinically useful as prognosis is generally quite poor in this group of patients.

- **Deep vein thrombosis and breast cancer**

  Validation performance was summarised across multiple studies for two developed prediction models, a model for diagnosing DVT and a model for mortality risk prediction in patients with breast cancer. The meta-analysis findings revealed that both models performed best in the validation studies when the intercept/baseline hazard was re-estimated in the population in which the model will be applied. In other words, recalibration of the intercept/baseline hazard was recommended so that the model had more consistent performance in populations of interest. However, the DVT model was not suitable for use, firstly because it was developed to illustrate methods rather than being the 'best' clinical model and, discrimination of the model was not very good. Discrimination could potentially be improved by including additional predictors in the model. The breast cancer model shows promise for use, because discrimination was slightly better and calibration was good on average for all implementations strategies, but had the narrowest prediction interval (therefore most consistent calibration) when the baseline hazard was estimated in the population for which it was intended for use. Data from sources such as electronic hospital records may be available to estimate the intercept for a particular population in which the model is intended to be used (as in strategy 1), but this is unlikely to always be the case. Debray et al. explored different ways in which a relevant intercept could be selected.[149] For example, using the intercept from a study with a similar prevalence (strategy 3) or by comparing baseline characteristics to select an intercept from the study included in the IECV approach with the most similar population characteristics. It is important to compare the different implementation strategies, as the model may be implemented in different ways depending on the data available for the relevant population.

There are also three main areas of contribution to prognosis research methodology from this thesis. These are:

1) Application of flexible parametric survival models for developing prognostic models.
2) Contributions towards developing prognostic models, especially with regard to the use of trials data.
3) Novel methods for external validation of prognostic models.

The key findings and recommendations were summarised in the discussion section of each chapter. However, the key methodology areas are now each discussed once more.

7.2 Application of flexible parametric survival models for prognostic model development

The literature review in Chapter 3 showed that 100% of the 31 articles included in the review used Cox proportional hazards models to develop their prediction model. A review of prognostic models in cancer also showed that Cox models were the most popular choice for survival data with 94% of studies using Cox models.[83] Flexible parametric models were first published in 2001 by Royston and Parmar and use restricted cubic spline functions to flexibly model the baseline cumulative hazard function.[32] They offer an alternative to the popular Cox proportional hazards models and standard parametric survival models such as the Weibull or exponential models which are often not flexible enough to model hazard functions in ‘real’ data.

The following sections summarise the findings of this thesis in terms of why Royston-Parmar flexible parametric models should be considered when developing a prognostic model for time-to-event data. These sections are very much related and therefore there is overlap between them.

7.2.1 The advantages of modelling the baseline hazard

Through application of flexible parametric models in hip replacement data and advanced pancreatic cancer, Chapters 2 and 4 highlighted several statistical advantages of modelling the baseline hazard. When the baseline cumulative hazard function is flexibly modelled using restricted cubic splines, the estimated hazard ratios from the model are almost identical to the hazard ratios that would be obtained from fitting a Cox model. Also, modelling the baseline hazard usually only requires a few degrees of freedom (generally between 2 and 5 d.f. are recommended depending on the complexity of the function and the amount of data available).[34] The main advantages of modelling the baseline hazard for prognosis are given in Box 7.1. In summary, modelling the baseline hazard is beneficial when developing prognostic models as it helps to understand the hazard profile of patients (by plotting the hazard function over time). It also facilitates absolute risk predictions at *any* time point for groups of patients as well as for individuals.

Box 7.1: Advantages of modelling the baseline hazard using flexible parametric models.

- The baseline hazard function can be plotted over time.
    - This helps to understand the course of the disease or condition over time. For example, when the hazard of the outcome is highest and how it changes over time.
    - This helps to inform decisions such as combining treatment groups (if necessary). This can be done by comparing the baseline hazard functions of each category to check how reasonable combining them may be.
- After fitting the prognostic model, population-averaged survival functions can be estimated and plotted.
    - This is done by averaging across all individuals using their individual predictor values in the model and possibly fixing the value of one predictor (e.g. treatment group). Survival functions are then averaged across all individual and these survival functions are therefore ‘adjusted’ for the other predictors (in contrast to Kaplan-Meier curves, which are unadjusted).
    - The difference between two survival functions (and 95% confidence intervals) can be calculated for selected time points. Therefore absolute survival differences can be reported in addition to relative measures such as hazard ratios.
    - This helps to identify if a predictor effect is clinically meaningful rather than using statistical significance alone.
- Survival functions (survival probabilities over time) can be predicted for individual patients by using their individual predictor values in the model.
- Time-dependent effects can be modelled by including an interaction between the time-dependent predictor and the baseline hazard function.
    - This essentially gives different baseline hazard functions for different values of the time-dependent predictor.
    - The baseline hazard function can be plotted for different values of the time-dependent predictor, which can help to understand how they differ over time.

7.2.2 Absolute risk prediction over time

Hazard ratios should be interpreted in relation to the baseline hazard as they could otherwise be misleading. For example, a large hazard ratio might reveal a small absolute effect when the baseline hazard is low. On the other hand, a hazard ratio close to one might reflect a large absolute effect if the baseline hazard is high. This was highlighted in Chapter 2 where the hazard ratio for cemented procedures relative to BHR in males under 55 years old was 3.86, but the baseline hazard rate was very low so the absolute difference in average survival probabilities was very small at 0.018 at 6 years.

Reporting absolute risk (probability of experiencing the event) for patients is important for prognostic studies.[18] Survival probabilities are more useful in understanding the likely prognosis for a disease and in making treatment decisions; hazard ratios are less meaningful in this context as they are relative measures. The literature review in Chapter 3 found that some authors used an estimate of baseline survival at a single time point to predict the survival probability at that time (Section 3.3.6). Alternatively, patients were categorised into risk groups based on a risk score (linear predictor). Survival probabilities for the risk groups could then be estimated using Kaplan-Meier estimates. Ideally, survival probabilities should not be limited to a single time point and should be calculated for individuals rather than categorising patients into risk groups.

Modelling the baseline hazard function using flexible parametric models allows absolute risk to be easily predicted and plotted for individual patients and/or risk groups over time, therefore not restricted to a single or a few time points (illustrated in Section 4.7). The survival functions from flexible parametric models are also smooth functions unlike the step-functions estimated using the non-parametric Kaplan-Meier method.

When Kaplan-Meier estimates are used to calculate survival probabilities for risk groups, only the patients within that risk group are used to estimate the survival probability for that risk group. This can result in large steps in survival probability when the number of patients in each risk group is small. Another benefit of using flexible parametric models is that all the data are used to estimate parameters in the prognostic model, and these model estimates are used to predict survival curves, therefore all of the data are being used to estimate each survival curve.

7.2.3 Individual predictions

A prognostic model is intended for predicting the probability of an outcome in individual patients. Therefore classifying patients into risk groups may not be the best approach. A patient's prediction could lie close to the border between two risk groups. If this is the case, using the average survival probability of the risk group may severely over- or under-estimate their actual survival probability. This was shown in Chapter 4 for two patients with very different survival functions that would both fall within the same risk group (Section 4.7).

Therefore, it is recommended that where possible, predictions should be made on an individual basis rather than using the average survival probability for a risk group. Of course this assumes that the model performs well in terms of discriminating between patients that have the event and those that do not, and that the model calibrates well for the setting in which it will be applied. The TRIPOD guidelines suggest that the full prediction model should be specified to allow individual prediction, including baseline survival at a given time point.[59] Using flexible parametric methods and providing the baseline survival over time would allow the user to select which time point is most relevant to them rather than restricting it to a time point selected by the author.

From a statistical point of view, it is also important to consider the uncertainty in any predictions made, which could be addressed by providing the range in which the survival probability is most likely to lie rather than giving the point estimate alone.

7.2.4 Statistical versus clinical viewpoint

This thesis has primarily focused on the statistical aspects of developing prognostic models but it is also important to consider the clinical viewpoint as this is where the models are intended for use. Clinicians may be sceptical about incorporating prognostic models into their decision making and management of patients, and so the translation of prognostic models into clinical practice is a pivotal area to now discuss.

From a statistical viewpoint, a prognostic model that explicitly models the baseline hazard and therefore can be used to predict survival probability over time is better than the Cox model which does not explicitly model the baseline hazard and therefore requires other means of estimating the baseline cumulative hazard or baseline survival probability at a chosen time point. Perhaps from a clinical perspective, the ability to predict over time may not be necessary and a simple scoring algorithm to provide risk at a single time point is easier to use. However, providing a model that can predict at any time point allows the user to select the most relevant time point for their setting.

From a clinical perspective, it may be desirable to have a model in which the associations (predictor effects) are clinically understandable and subsequent predictions are easily calculated. Prognostic information may not all be available immediately and may rather come in waves as more tests are run.[205] It may not be clear how predictions from prognostic models can be used to help manage patients or perhaps clinicians are looking for a clinical decision rule rather than the prognostic model itself. In other words, they may want the model with defined cut-points for risk in order to categorise patients into high/low risk groups. Patients that are high or low risk for the outcome may then be managed differently in terms of treatment plans. Statistically, this simplification of the model results in loss of information and, as discussed earlier, patients with a predicted risk close to the cut-point may not be dissimilar to patients in another risk group. Therefore, even if risk groups are required clinically, it is better to know the actual predicted risk for an individual, so that more exact individual decisions can be made. The use of cut-points is also context specific, and should come after presenting the full model. This would avoid problems with misclassification due to miscalibration in some settings.

7.3 Using clinical trials data and other challenges in developing prognostic models

The literature review in Chapter 3, like many other reviews detailing development and reporting of prediction models,[11,83,85,144,145] highlighted issues in the methods used to develop the models as well as poor reporting of statistical methods in some studies. This included topics such as checking linearity of continuous variables, dichotomising continuous variables, variable selection methods and handling of missing data. Several articles have been published on the need to improve methods and reporting,[1,11,18,82] and the recently published TRIPOD statement provides clear guidelines for reporting prediction model studies that should hopefully help improve the quality of reporting (and methods used) in future prediction studies.[42,59]

In chapter 4, clinical trials data were used to develop a prognostic model for advanced pancreatic cancer. Several challenges were identified, some specific to using clinical trials data and other more general issues. Two clinical trials were included in model development but a trial variable was not included in the model as it would not be useful to future patients.

Also, as the two trials randomised patients to different doses of an experimental treatment, marimastat, this would not be very useful to future patients. Therefore the decision was made to combine treatments according to whether they received the standard treatment gemcitabine or not as this is likely to be more relevant, especially as marimastat was not shown to improve survival of patients in the original trials.[154,155] The shape of the baseline hazard also supported this decision to combine treatment groups in this way. There has been debate about whether prognostic models should be developed using only the control arm of a trial.[12,206] However, Moons et al. suggest combining treatment groups if the treatment is ineffective (relative risk=1), or if treatment is effective, using the data from all treatment groups and including the treatment variable as a predictor in the model.[8]

In summary, many of these challenges are well known and have been published in a number of articles. However, developing a prognostic model can still be difficult and decisions need to be made on how to handle these difficulties. Although one example has been considered in depth in this thesis (Chapter 4), the experience may be very different using different clinical trials data. Also, although any one of the challenges may have a simple solution, dealing with them becomes much more complex when they occur in combination. For example in Chapter 4, multiple imputation was used to handle missing predictor data. This made variable selection more complex, while also trying to simultaneously model continuous variables appropriately, now using multiple imputed datasets. There was also a time-dependent effect that added further complexity to the prognostic model. Box 7.1 lists some considerations when developing prognostic models. Many of these are general issues when fitting any model and have been reported by other authors, but it is worth mentioning them as they are crucial when developing prognostic models. This list is by no means exhaustive as different challenges are faced in each dataset.

Box 7.2: Recommendations for improving the development of a prognostic model, as identified from the literature review (Chapter 3) and development of a prognostic model using trials data (Chapter 4).

- Handling missing data
    - If a variable contains a lot of missing values, consider whether it would be useful to include it in a model if the variable is not routinely recorded.
    - If a dataset contains missing values for several variables, consider whether the data are missing at random and if so, use multiple imputation with the number of imputed datasets equal to the percentage of observations with missing values for any variables considered in the modelling process.
- Modelling continuous variables
    - Consider non-linear functions, e.g. simple transformations or fractional polynomials, but avoid categorising continuous variables where possible.
- Baseline hazard function
    - If the aim is for prediction, the baseline hazard function should be modelled using a model such as the flexible parametric Royston-Parmar model.
    - Use AIC and BIC to guide choice of d.f. (usually between 2-5 d.f.) and plot baseline hazards using different d.f. against non-parametric estimate (using a smoother) to compare functions and select appropriate d.f.
- Variable selection process
    - Include known prognostic variables regardless of statistical significance (including trial stratification factors if data come from clinical trials).
    - Backward elimination has been recommended as a better approach to forward selection.
    - Consider a variable selection method that allows for non-linear functions of continuous variables as part of the modelling process, such as MFP.
    - MFPMI can be used for variable selection using multiply imputed data.

Box 7.2 continued…

  - Trial specific: do not include a trial variable as not useful for prediction in future patients but check that trials are similar in terms of inclusion/exclusion criteria and baseline hazard function.
  - Trial specific: carefully consider whether a treatment variable is necessary in the model and how it should be included. Use the baseline hazard function to guide this if necessary.
- Proportional hazards assumption
  - Check proportional hazards assumption for all variables considered for the multivariable model.
  - If not proportional, are the other predictor effects consistent if separate models are fitted for different values of the non-proportional variable? If so, consider a time-dependent effect which models an interaction between the baseline hazard function and the time-dependent variable. If the other predictor effects are not consistent, consider fitting separate models if the effective sample size is reasonable within each group, defined by different values for the time-dependent variable.

7.4 Validating a prognostic model

With so many models being developed, it may be difficult for clinicians to decide which models are worth implementing.[205] PROGRESS also encourage researchers to share IPD in order to develop and validate better prognostic models.[3] The authors also suggest a shift in focus to validating and updating existing models rather than constantly developing new models.[3] Therefore, in addition to the previously mentioned model development considerations, it is important that a prognostic model predicts well in data external to that in which the model was developed. Ideally, this should be done in several external settings.

If IPD is to be shared and prognostic models are developed and validated using multiple studies, internal-external cross-validation can be used to evaluate model performance across multiple studies. The findings from Chapter 5 and 6 are summarised and discussed below.

7.4.1 Using multiple studies to develop and validate a prognostic model

When IPD from multiple studies is available, one option is to allocate some studies to a development set and reserve the others for external validation. However, it is more efficient to use all available data for model development, therefore increasing the power to identify important predictors, possible interactions and non-linear functions. The issue then is that there is no external data in which to evaluate model performance. The aim of external validation is to ensure that the model performs well in other related populations or settings, which is referred to as generalisability of the model. Internal-external cross-validation was proposed by Royston et al.[69] for validating a prognostic model when multiple studies are available, which provides a useful way to validate a prognostic model by fitting the developed model multiple times, each time excluding one study and reserving it for validation. The *β*-coefficients for the predictors in the model are re-estimated each time a study is excluded so that the validation study is external to model derivation, thereby assessing generalisability of the model performance across the different ‘external’ studies.[69] Debray et al.[149] extended this work by developing a framework that considers different implementation strategies (for the intercept) when validating the model in the excluded study.

Although not considered in this thesis, the internal-external cross-validation approach could also be adapted to incorporate a model development strategy phase, for when the set of included predictors and their functional form is not pre-specified (for example, backwards selection of predictors and possible non-linear functions might be used). The use of internal-external cross-validation in this context would help evaluate the reproducibility of a model as derived from a particular development strategy. This may help identify the best strategy for use (e.g. shrinkage of predictor effects), by revealing (through the meta-analysis methods proposed) which strategy leads to consistently the best performance upon external validation.

7.4.2 Meta-analysis to summarise validation performance of a model across studies

Univariate random-effects meta-analysis was proposed to summarise validation performance statistics such as the C-statistic and calibration slope across studies individually. Model performance was evaluated considering different implementation strategies for the intercept or baseline hazard of the model when applied to the validation study, as proposed in the framework by Debray et al.[149] In Chapter 5, model performance was summarised for two prediction models (DVT diagnosis and breast cancer prognosis). Random-effects meta-analysis highlighted that the average value for a performance statistic is an incomplete summary measure and that estimating heterogeneity across the studies is important when evaluating model performance, as this indicates how much variability there is in performance in different settings. An ideal model would have good average performance and no heterogeneity across studies, therefore performing consistently well across studies. However, in reality a model may perform better in some settings than in others. Prediction intervals incorporate an estimate of the between-study variance therefore accounting for heterogeneity in the performance statistic across studies, and should therefore be considered when evaluating how well a model is expected to perform (for a particular performance statistic) if it were to be applied in a new but similar setting.

Comparing performance statistics for the different implementation strategies showed that it is best to re-estimate the intercept or baseline hazard in the setting in which the model will be used. Average model performance, considering the calibration slope for example, was not necessarily closest to the ideal value of one (perfect calibration) for this implementation strategy but calibration performance was far more consistent across studies, and thus far more likely to be acceptable upon application. A model could severely over-predict in some settings and under-predict in others, but still have good average performance. Such a model would not predict particularly well in many of the individual settings and this was seen for the DVT model when evaluating calibration-in-the large if the average intercept was used in the validation study (Figure 5.2), which had an extremely wide 95% prediction interval ranging from -1.24 to 1.24. Conversely, if a model has no heterogeneity (and thus a very narrow prediction interval), this is only adequate if the average performance is itself acceptable. Therefore, when selecting the best implementation strategy one needs to potentially compromise between the average performance and the amount of between-study heterogeneity. This can be summarised by predicted probabilities of a good performance (see below).

7.4.3 Multivariate meta-analysis for predicting performance of a model in a new setting

In addition to pooling performance statistics using univariate meta-analysis, multivariate meta-analysis was proposed to summarise two or more performance statistics using the additional information of within-study correlations between the performance statistics (obtained by bootstrapping). The utilisation of correlation can lead to slightly narrower confidence intervals and prediction intervals but more importantly, better reflects how performance of a model should consider measures of both discrimination and calibration performance. A recent literature review of studies evaluating prediction models found that discrimination and calibration were not reported in 27% and 67% of included studies respectively,[54] highlighting that calibration is often not reported. Therefore in Chapter 5, bivariate random-effects meta-analysis was used to jointly summarise the C-statistic and calibration slope of each prediction model. Based on the results of the bivariate meta-analysis, a 95% prediction ellipse can be calculated and plotted to give a region in which the model performance in a new but similar setting is expected to lie for both calibration and discrimination performance.

Altman et al. have suggested that it may be helpful to pre-specify criteria for acceptable model performance in terms of calibration and discrimination performance.[20] Using this approach, Chapter 5 also illustrated how the predicted probability of a model performing to specified criteria can be calculated, assuming the results of the bivariate meta-analysis are correct. This is done by sampling from a bivariate t-distribution (using the results of the meta-analysis as input for the matrix of mean values and the variance-covariance matrix) for the measures of calibration and discrimination, and calculating the proportion in which the criteria for both performance statistics are satisfied. In addition to the probability of 'good' overall performance this approach provides, it also allows models and/or implementation strategies to be ranked in terms of preference if the overall performance is acceptable. Of course, deciding which performance statistics to consider and what constitutes acceptable performance is unclear.[20,207]

Caution should be taken in predicting model performance when the between-study correlation is poorly estimated (usually -1 or +1) as this will have an impact on the predicted probability. In Chapter 5, between-study correlations were poorly estimated for the DVT example when the intercept was estimated in the validation study (implementation strategy 1) resulting in an estimated correlation of +1 which usually indicates that it has reached the end of its parameter space without converging.[193] This can happen when the within-study variance is very large relative to the between-study variance.[193] A sensitivity analysis in Chapter 5 assumed a between-study correlation of +0.5 rather than +1.0 when calculating predicted probabilities for the DVT model using the average intercept (strategy 2), meant a change in predicted probabilities of acceptable joint performance from 2.4% to 9.6%. Pooling several performance statistics together, for example in a trivariate meta-analysis, may improve estimation of the between-study correlation as additional information. When this was done for the DVT model, between-study correlations did converge but were still estimated very close to +1. This warrants further investigation as to whether this is likely to be genuine and therefore ensure any joint predictions using the meta-analysis results are valid.

7.4.4 Assumption of normality for true between-study performance

A key assumption in a random-effects meta-analysis is the assumption of normality. The meta-analysis model firstly assumes that the study-specific (within-study) performance statistic is normally distributed and secondly that the true (between-study) performance statistic is normally distributed. The first normality assumption is reasonable using the Central Limit Theorem as justification if sample sizes are relatively large. However, normality of the between-study performance statistics had not previously been tested reliably. Therefore, through simulation, Chapter 6 aimed to establish whether the true performance statistics were normally distributed in a variety of scenarios, or if transformations of the performance statistics considered should be used. In the simulation settings and scenarios considered in Chapter 6, the between-study distributions for calibration-in-the-large and the calibration slope were reasonably normally distributed. The expected/observed ratio (E/O) was skewed as expected and the log transformation was a better scale for normality of the between-study distribution. The C-statistic was skewed in many simulation settings and a key finding of Chapter 6 is the recommendation of using the logit transformation for pooling C-statistics, providing a different recommendation to that proposed by van Klaveren et al.[62]

The simulation study also highlighted the problem with heterogeneous predictor effects. When heterogeneity in the predictor effect was large, many of the performance statistics were skewed and transformations did not produce approximately normal between-study distributions. In reality, the level of heterogeneity considered in some scenarios was extreme but caution should be taken if heterogeneity in predictor effects is suspected. Differences in the predictor effects could arise due to differences in populations, such as truly different patient populations, different definitions of either the outcome or predictors, or different distributions of predictors that are missed (not modelled).[12]

To conclude this section, key findings and recommendations from Chapters 5 and 6 on internal-external cross-validation are summarised in Box 7.3 below.

Box 7.3: Summary of key finding and recommendations for internal-external cross-validation of a prognostic model using multiple studies.

- Random-effects meta-analysis is recommended for summarising model performance statistics across multiple validation studies.
    - This provides a measure of the between-study variance in addition to the average for the performance statistic.
    - Ideal performance should have good average performance and little or no heterogeneity in performance across studies.
    - 95% prediction intervals can be derived for the expected model performance (of any performance statistic) in a new but similar study or setting.

Box 7.3 continued…

- Bivariate and multivariate random-effects meta-analysis can be used to jointly summarise measures of calibration and discrimination (or other relevant performance statistics).
    - Utilises within-study correlations between performance statistics which can result in slightly narrower marginal confidence intervals and prediction intervals.
    - 95% prediction ellipses can be plotted that show the region of expected combinations of values of both calibration and discrimination performance of the model in a new but similar study or setting.
    - Predicted probabilities of the model performing to specified criteria can be estimated by sampling from a bivariate t-distribution, using the results of the bivariate meta-analysis to specify the mean and variance-covariance matrices for the bivariate t-distribution.
    - Predicted probabilities can be used to rank models or implementation strategies and select the one with the highest probability of 'good' joint performance.
- Performance statistics should be pooled on a scale that can reasonably be assumed to be approximately normally distributed for the between-study distribution. The following scales or transformations are recommended, based on the findings of the simulation study in Chapter 6:
    - Original scale for calibration slope or calibration-in-the-large
    - Log transformation for expected/observed ratio
    - Logit transformation for the C-statistic

7.5 Further research

Several areas of further work or updating previous work have been identified for several of the chapters and are discussed below.

The literature review in Chapter 3 could be extended to assess if in the years since the review was done (2012), more researchers are modelling the baseline hazard or summarise how they are attempting to report absolute risk for survival models. The review was limited to six general medical journals and perhaps this is not representative of the models published in more specialist journals. However, a review of cancer prognostic models, which searched Pubmed for articles (published in 2005) and included high impact (specialist) cancer journals, also found that Cox models were fitted in the majority of included studies.[83] One way to check if the baseline hazard is being modelled more in the time since the literature review in Chapter 3 was conducted, might be through citation searching, although new or different methods may be missed. There are a large number of prediction or prognostic model development studies being published, therefore it is necessary to narrow down the search in some way.

Chapter 4 presented a prognostic model for advanced pancreatic cancer. This extended previous work by Stocken et al.,[156,157] where issues such as how best to combine treatment groups, handling missing data and appropriately modelling non-linear functions for variables such as CA19-9 were carefully considered. The model has been internally validated but the next step would be to externally validate the model in independent data to assess how generalizable it is. Before doing so, shrinkage factors should be applied to the model parameters to adjust for over-optimism.

Chapters 4 and 5 considered more than one study when fitting Royston-Parmar models. This involved making assumptions about the shape of the baseline hazard function, generally that the shape of the baseline hazard function was the same in the studies, and proportional to each other. An area for further work could be on how to ensure that baseline hazard functions are comparable and can be combined, how similar they should be and what the potential effects of combining studies with different underlying shapes has on predictions. In particular, whether recalibration that includes a new baseline hazard shape is preferable.

Chapter 6 assessed the normality assumption for four performance statistics (C-statistic, calibration slope, E/O and calibration-in-the-large) based on a ‘true’ logistic prediction model being validated across many studies through simulation. This could be extended to consider further performance statistics such as the D-statistic and consider a survival model as the prediction model.

Further simulation studies are needed to evaluate the statistical properties of the meta-analytic approach proposed for IECV in Chapter 5. A simulation study could help evaluate whether the approach produces unbiased estimates of predictive performance for performance statistics (e.g. C-statistic), appropriate coverage and prediction intervals. If the approach performs well, further work into developing a program would be advantageous in allowing other researchers to easily apply the proposed methods. The program would aim to automate the IECV approach, by excluding one study at a time, re-fitting the specified model, calculating appropriate performance statistics and summarising model performance using random-effects meta-analysis. It could also be extended to include multivariate meta-analysis which would require bootstrapping for within-study correlations and allow joint predictions. Further simulation studies could evaluate the use of multivariate meta-analysis to compare competing existing models, and the use of multivariate meta-analysis to evaluate the added benefit of an additional predictor.

The simulation study in Chapter 6 only considered the scale for the between-study distribution of performance statistics; further work could evaluate whether the same scale should be used for the within-study distribution.

7.6 What is the future for prognosis research?

While this thesis has explored specific areas relating to the development and validation of prognostic models, there are still many challenges that are left unaddressed. Over the last few years, many articles have been published in an attempt to highlight current weaknesses and areas that require methodology research.[1-4] It is also essential to improve the transparency and reporting of prognostic and prediction studies, with Peat et al. encouraging transparency of prognosis research to avoid bias and improve conduct by proposing prognostic studies should be registered and protocols published, similar to RCTs.[82] Guidance documents for reporting of studies have been published for many types of studies, such as STROBE for observational studies,[208] STARD for diagnostic accuracy studies,[209] and REMARK for tumour marker studies.[210] More recently, the TRIPOD Statement was published for reporting of multivariable diagnostic and prognostic prediction models,[42,59] as many aspects of developing or validating prediction models and prognostic models has been shown to be poor in literature reviews conducted over recent years.[54,83,85,145] Use of guidelines can be slow on the uptake, but does slowly improve the quality of reporting studies as found for STARD.[211] This is likely to be the same for TRIPOD and requires journals to encourage authors to use the published guidelines.

Many methodology articles including TRIPOD,[42] Steyerberg et al.[212] and Moons et al.[58] discuss development and validation of prognostic models and give clear advice on certain issues, for example, how to handle missing data, model continuous variables, select variables for inclusion etc.; however, there will still be challenges in developing prognostic models that are specific to the data being used and this can easily be complicated when multiple challenges arise. Flexible parametric models have received little attention in the area of prognosis research, yet offer advantages such as those identified in this thesis. Perhaps discussion of flexible parametric models in articles appealing for better conduct and reporting by recognised methodologists may encourage researchers to consider alternatives to the Cox model.

PROGRESS encourage a shift in focus to externally validating and updating prognostic models rather than developing a new model each time more data are collected in a particular clinical area.[3] If the research community responds to the plea for data sharing (of IPD) and transparency, better models can be developed across a wider range of settings, hopefully resulting in models that are more generalizable to different settings. It is also important to consider how applicable a model is to different settings (generalisability), whether the model can be improved, for example recalibrated or including additional predictors, or whether different models should be applied in different settings. This remains a challenge but potentially by sharing data in a collaborative effort, these questions could be evaluated more reliably if many studies are available for external validation. Developing a ‘good’ prediction model should be an iterative process as models can be improved with additional data and new promising prognostic factors. If done in this way, it may help clinicians keep track of prognostic research in their clinical field and encourage use of prognostic models in clinical practice.

Croft et al. discuss the idea of a shift in framework from diagnosis to prognosis for managing patients and evaluating clinical research.[213] Potential reasons for this are that diagnosis is not always useful if it does not lead to better patient outcomes; also, generally the idea underpinning diagnosis and treatment is patient prognosis. Another reason is that for many conditions, diagnosis may not really be as simple as having the disease or not and it could be thought of on more of a continuous scale without the need for a dichotomy. Such diagnostic information could rather be incorporated in prognostic models for predicting patient outcome.[213]

Another area of progression for prognosis research may be to incorporate longitudinal data into the survival model using joint modelling,[214,215] thereby incorporating the trajectory of biomarkers which change over time and potentially improve predictions for patient prognosis.

In terms of validation of prognostic models, it is still unclear what constitutes 'good' performance. However, researchers such as Vickers and Cronin state that calibration and discrimination do not really tell us if a model is any good and instead suggest a move towards decision analytics.[207] Assuming treatment decisions are based on classification of patients to high or low risk groups (using a decision threshold), decision curve analysis can be used to calculate net benefit and harm of the model over a range of thresholds. This may warrant further consideration but may not be applicable in all settings as prognostic models may be used to aid decision making by providing a predicted probability rather than assuming it will always result in a decision rule. However, such methods warrant further investigation and may have use in comparing models or assessing the incremental value of additional predictors. Consideration of meta-analysis methods in this regard may be important.

7.7 Conclusions

Prognosis research and in particular prognostic modelling is still a challenging area that requires more methodology research to improve the models being developed and validated. The aim is to provide useful models that will be implemented in clinical practice, and ultimately, improve patient outcomes and their wellbeing. Though many issues remain, this thesis has contributed toward improvements in the prognostic modelling field through application and methodological development. In particular, the use of flexible parametric modelling and meta-analysis methods will hopefully improve the development, evaluation and uptake of robust prognostic models in the coming years.

APPENDICES

Appendix A

Appendix A1: Log-log plots for the proportional hazards assumption

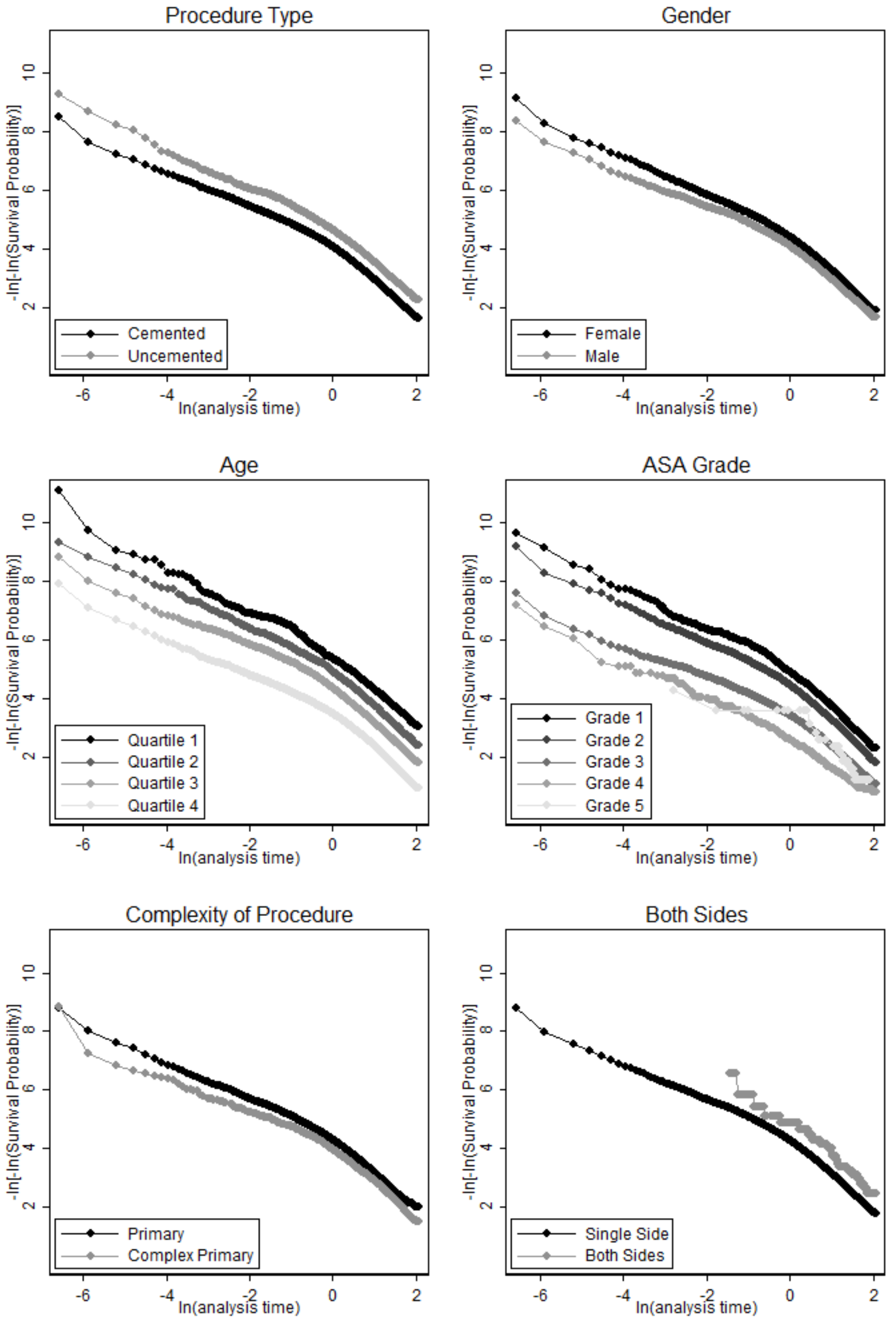


Figure A1.1: Log-log plots for mortality as the outcome.

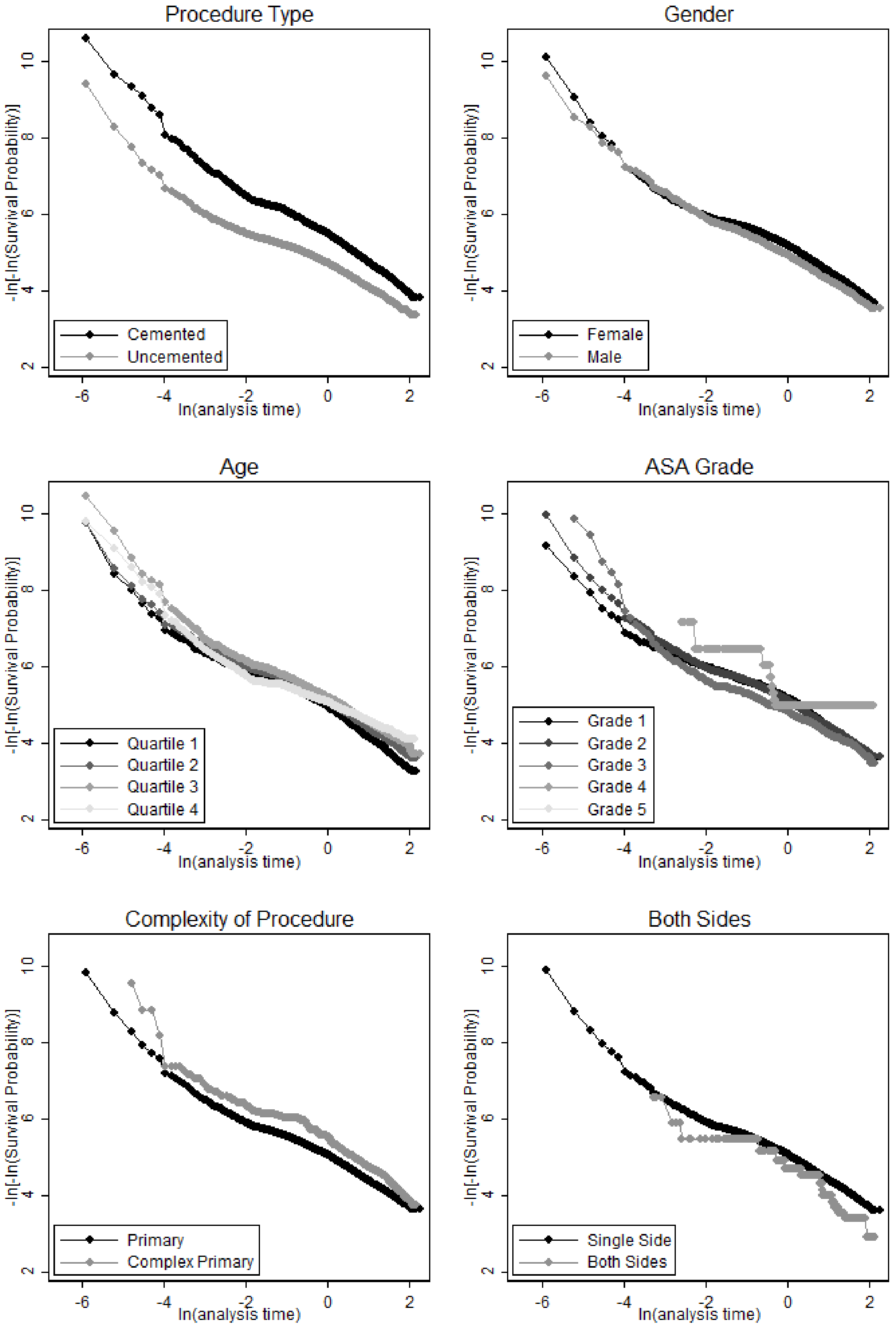


Figure A1.2: Log-log plots for revision as the outcome.

Appendix A2: Kaplan-Meier (unadjusted) survival curves

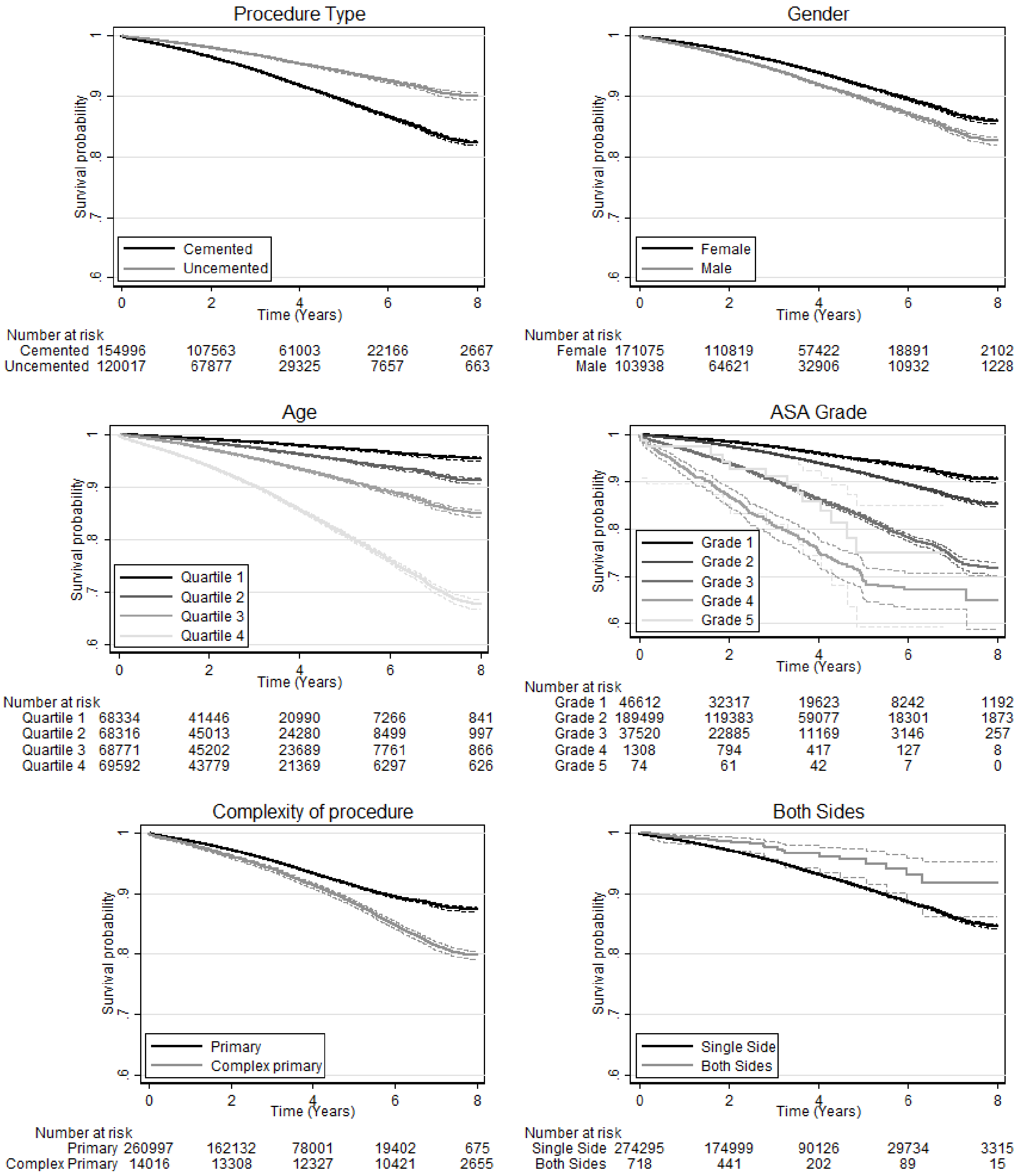


Figure A2.1: Kaplan-Meier curves for mortality as the outcome.

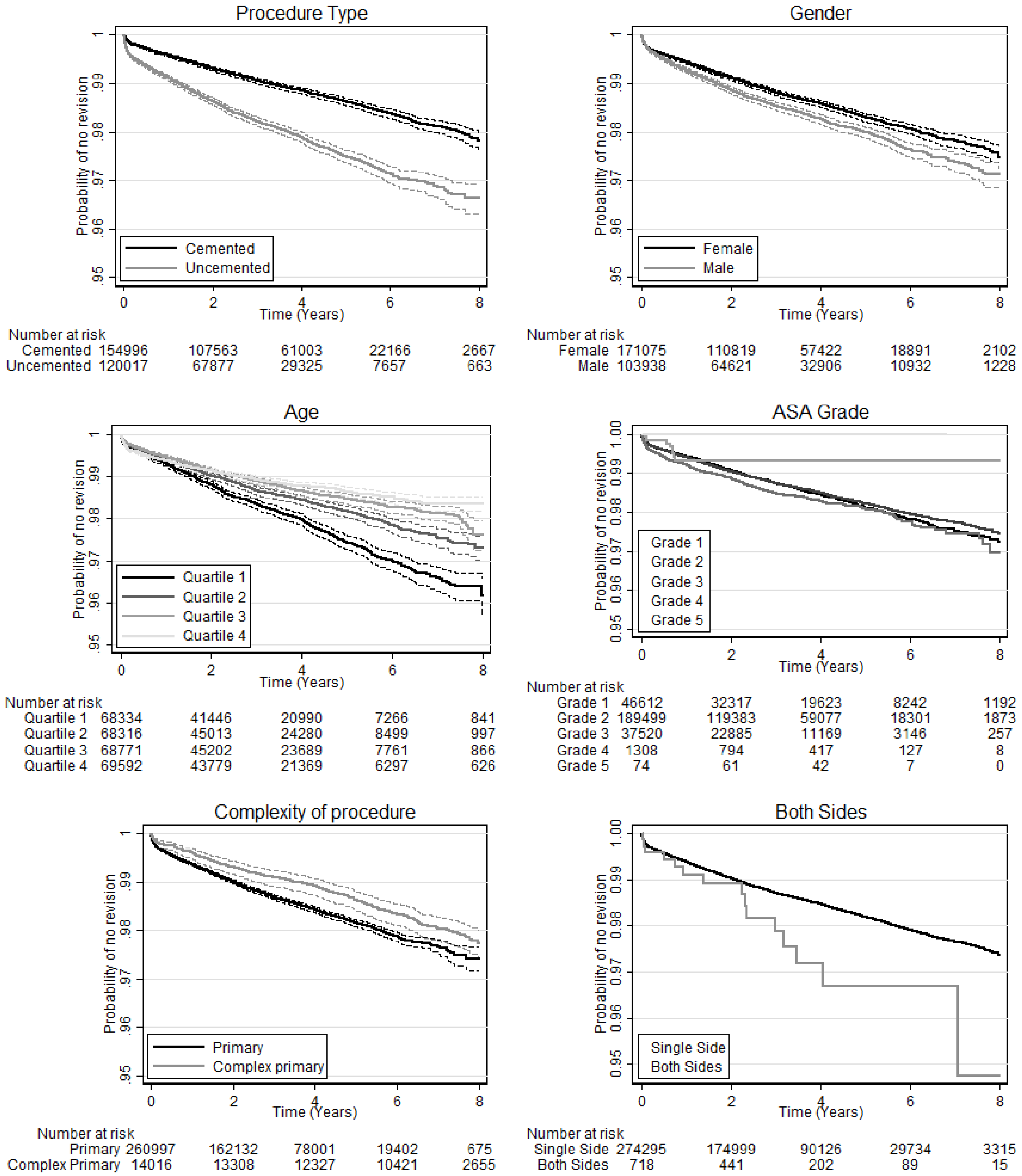


Figure A2.2: Kaplan-Meier curves for revision as the outcome.

Appendix A3: Sensitivity to approach

Excluding patients with missing information for surgical approach resulted in a loss of 15247 observations. Table A3.1 shows that the hazard ratios change slightly compared to the original model and approach is highly significant in the model (Wald test p<0.001). However, there is still a significant difference in mortality between cemented and uncemented procedures and although the predicted population-averaged survival probabilities for the procedure types change, the difference in mean predicted survival is still very similar even at 8 years of follow-up with a difference in mean survival probabilities 0.0142 compared to 0.0131 in the primary analysis of mortality (Figure A3.1).

Table A3.1: Sensitivity to approach analysis results including model estimates from original model and model including approach.

	Hazard ratio (95% confidence interval)	
Variable	Original model	Model inc. approach
Uncemented procedure	1	1
Cemented procedure	1.111 (1.069 to 1.155)	1.144 (1.098 to 1.191)
Age (years)	1.090 (1.088 to 1.092)	1.089 (1.087 to 1.092)
ASA grade 1	1	1
ASA grade 2	1.192 (1.133 to 1.255)	1.174 (1.106 to 1.246)
ASA grade 3	2.152 (2.033 to 2.279)	2.112 (1.979 to 2.255)
ASA grade 4	3.518 (3.092 to 4.004)	3.577 (3.120 to 4.101)
ASA grade 5	2.938 (1.702 to 5.071)	2.781 (1.610 to 4.804)
Female	1	1
Male	1.537 (1.488 to 1.587)	1.557 (1.503 to 1.613)
Non-complex	1	1
Complex	1.395 (1.332 to 1.462)	1.466 (1.285 to 1.672)
Anterior approach	-	1
Antero-lateral approach	-	1.220 (0.928 to 1.603)
Hardinge approach	-	0.810 (0.615 to 1.066)
Lateral (inc. harding) approach	-	1.238 (0.944 to 1.623)
Other approach	-	0.935 (0.696 to 1.257)
Posterior approach	-	1.006 (0.767 to 1.318)
Trochanteric osteotomy approach	-	0.630 (0.373 to 1.064)

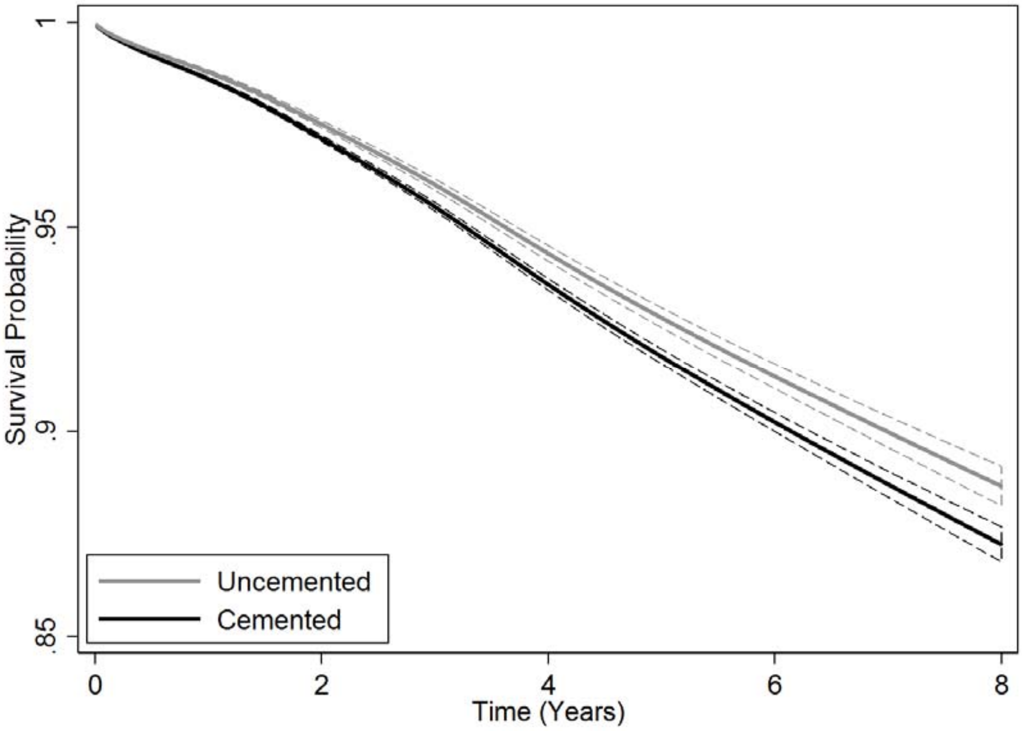


Figure A3.1: Adjusted survival curve for procedure type and mortality as the outcome, from the model including approach (95% CIs given by dashed lines).

Appendix A4: Inclusion of Birmingham Hip Resurfacing: male only analyses

The following analyses used the males from the original dataset with the addition of males that received the Birmingham Hip Resurfacing (BHR) procedure. There is again an imbalance between the three groups as can be seen from the baseline characteristics in Table A4.1. The BHR males are younger with a mean age of 55.9 years and have a higher proportion of males in ASA grade 1 (47.9%) and ASA grade 2 (48.9%), indicating younger and healthier patients receiving this procedure.

Table A4.1: Summary of baseline characteristics, outcome and follow-up by procedure type in males only.

		BHR (n=8352)	Cemented (n=53409)	Uncemented (n=50529)
Baseline characteristics				
Age, years	Mean (SD)	55.92 (8.58)	72.26 (8.61)	66.22 (9.94)
	Median	56.58	72.89	66.67
	IQR	50.35 – 61.93	67.15 – 78.08	60.27 – 73.14
	Range	19.03 – 84.99	18.1 – 101.69	17.13 – 98.76
ASA grade, n (%)	1	3999 (47.88)	8246 (15.44)	10761 (21.30)
	2	4087 (48.93)	35838 (67.10)	33594 (66.48)
	3	256 (3.07)	8947 (16.75)	5918 (11.71)
	4	8 (0.10)	360 (0.67)	246 (0.49)
	5	2 (0.02)	18 (0.03)	10 (0.02)
Complexity	Non-complex	8262 (98.92)	48743 (91.26)	50176 (99.30)
	Complex	90 (1.08)	4666 (8.74)	353 (0.70)
Both sides, n (%)	No	8297 (99.34)	53319 (99.83)	50304 (99.55)
	Yes	55 (0.66)	90 (0.17)	225 (0.45)
Follow-up				
Endpoint, n (%)	Death	93 (1.11)	4821 (9.03)	1872 (3.70)
	Revision	159 (1.90)	645 (1.21)	830 (1.64)
	Unrevised	8100 (96.98)	47943 (89.77)	47827 (94.65)
Length of follow-up, person-years	Total	27961	183101	134702

Mortality

The length of follow-up is shorter for BHRs than for cemented or uncemented procedures. Therefore, to avoid extrapolating results for BHRs, follow-up length in the analyses was restricted to 6 years. This means that any patient that had not had an event by 6 years was censored at this time.

After adjustment for age, ASA grade and complexity, there is a significant difference between BHR and both cemented and uncemented procedures (both p<0.001) in males. The mortality rate is significantly higher for cemented and uncemented procedures when compared to BHR with hazard ratios of 1.67 (95% CI: 1.35 to 2.06) and 1.47 (95% CI: 1.19 to 1.82) respectively (Table A4.2).

Table A4.2: Multivariable model estimates including BHR for mortality as the outcome in males only.

Variable	Hazard ratio	95% Confidence interval	P-value
BHR procedure	1.000	-	-
Cemented procedure	1.667	1.349 to 2.061	<0.001
Uncemented procedure	1.473	1.192 to 1.821	<0.001
Age (years)	1.089	1.086 to 1.093	<0.001
ASA grade 1	1.000	-	-
ASA grade 2	1.163	1.076 to 1.257	<0.001
ASA grade 3	2.169	1.990 to 2.364	<0.001
ASA grade 4	3.687	3.064 to 4.437	<0.001
ASA grade 5	0.546	0.077 to 3.880	0.546
Non-complex	1.000	-	-
Complex	1.274	1.179 to 1.377	<0.001

The population-averaged survival curves are shown in Figure A4.1. After adjusting for confounding, the mean predicted probability of survival for BHR in males is lower than the unadjusted survival curves but remains higher than cemented and uncemented with a probability of 0.921 (95% CI: 0.906 to 0.936) at 6 years, compared to 0.877 (95% CI: 0.873 to 0.881) and 0.888 (95% CI: 0.883 to 0.893) for cemented and uncemented respectively.

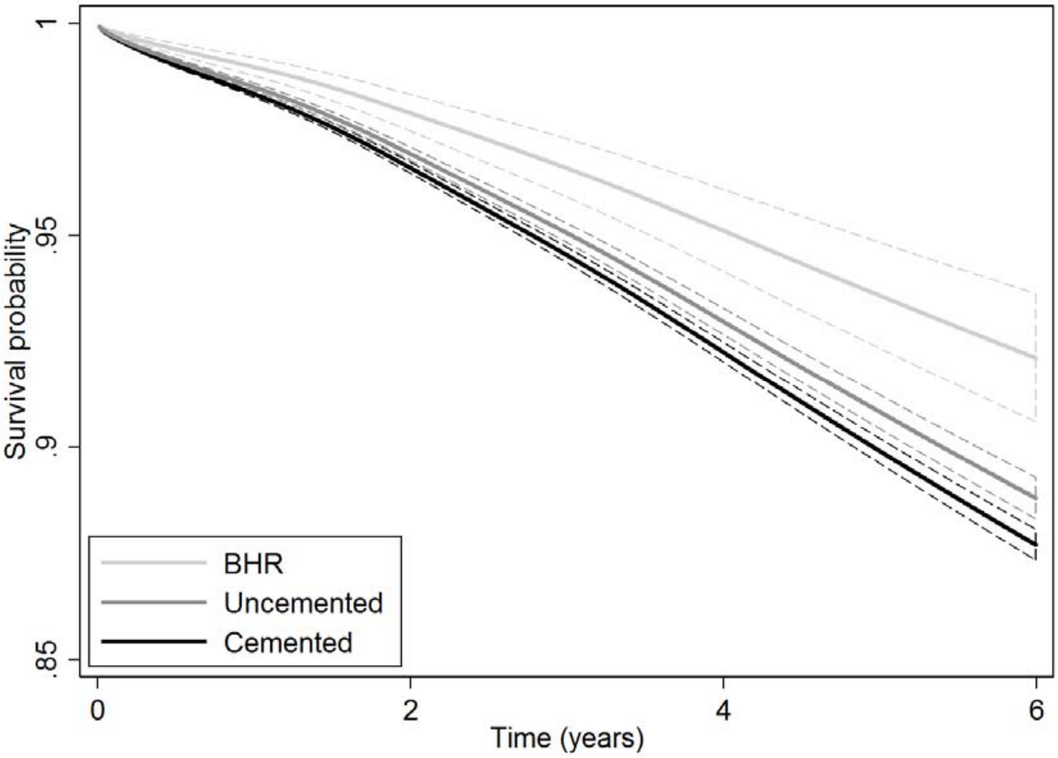


Figure A4.1: Population-averaged survival curves for procedure type including BHR in males for mortality as the outcome (95% CIs given by dashed lines).

Revision

After adjustment for age and ASA grade, the hazard of revision is 35% lower (95% CI: 22% to 46%) for the cemented group compared to BHR. No significant difference is seen between BHR and uncemented procedures with a hazard ratio of 1.04 (95% CI: 0.87 to 1.25), as shown in Table A4.3 below. Age was included in the model, given its importance in previous models and borderline significant p-value of 0.129.

Table A4.3: Multivariable model estimates including BHR in males for revision as the outcome.

Variable	Hazard ratio	95% Confidence interval	P-value
BHR procedure	1.000	-	-
Cemented procedure	0.645	0.530 to 0.784	<0.001
Uncemented procedure	1.044	0.872 to 1.249	0.639
Age (years)	0.996	0.990 to 1.001	0.129
ASA grade 1	1.000	-	-
ASA grade 2	1.109	0.977 to 1.257	0.109
ASA grade 3	1.255	1.051 to 1.498	0.012
ASA grade 4	0.551	0.205 to 1.480	0.237
ASA grade 5*	-	-	-

*Not estimable

The adjusted survival curves in Figure A4.2 show very little difference in the probability of no revision between uncemented and BHR procedures at any time point and the greatest probability of no revision is observed for the cemented group. However, the absolute mean probability of no revision is high for all three procedure groups. At 6 years of follow-up, the mean probability of no revision is 0.981 (95% CI: 0.980 to 0.983) for cemented, 0.970 (95% CI: 0.968 to 0.972) for uncemented and 0.971 (95% CI: 0.966 to 0.976) for BHRs.

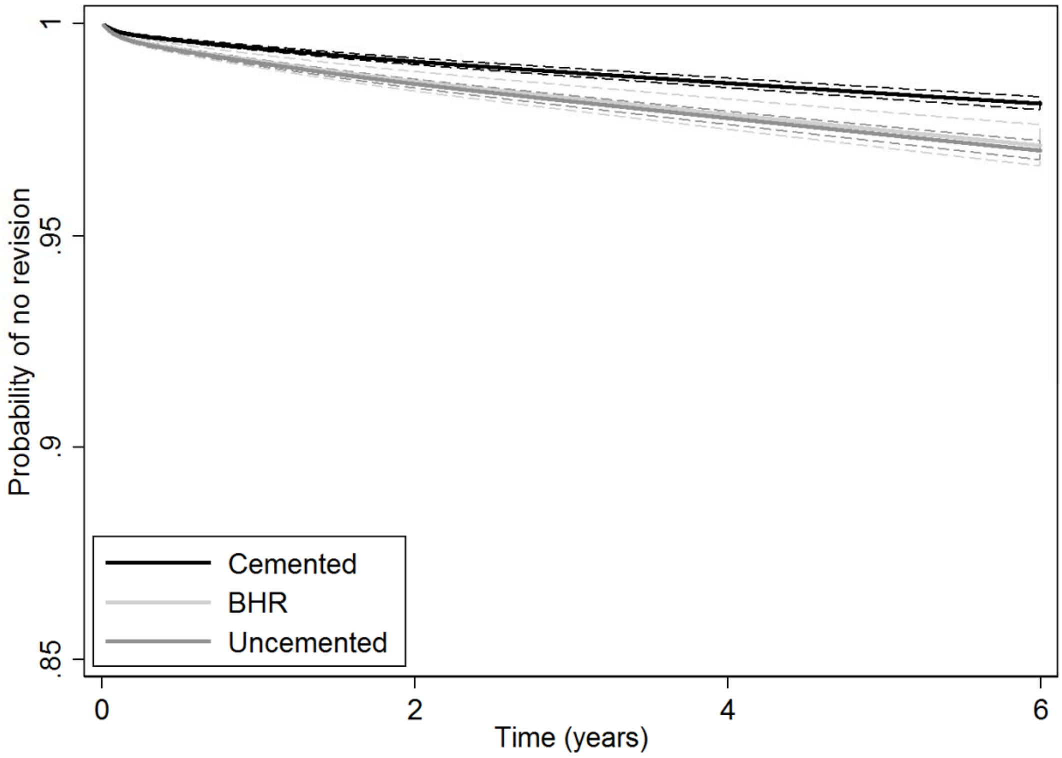
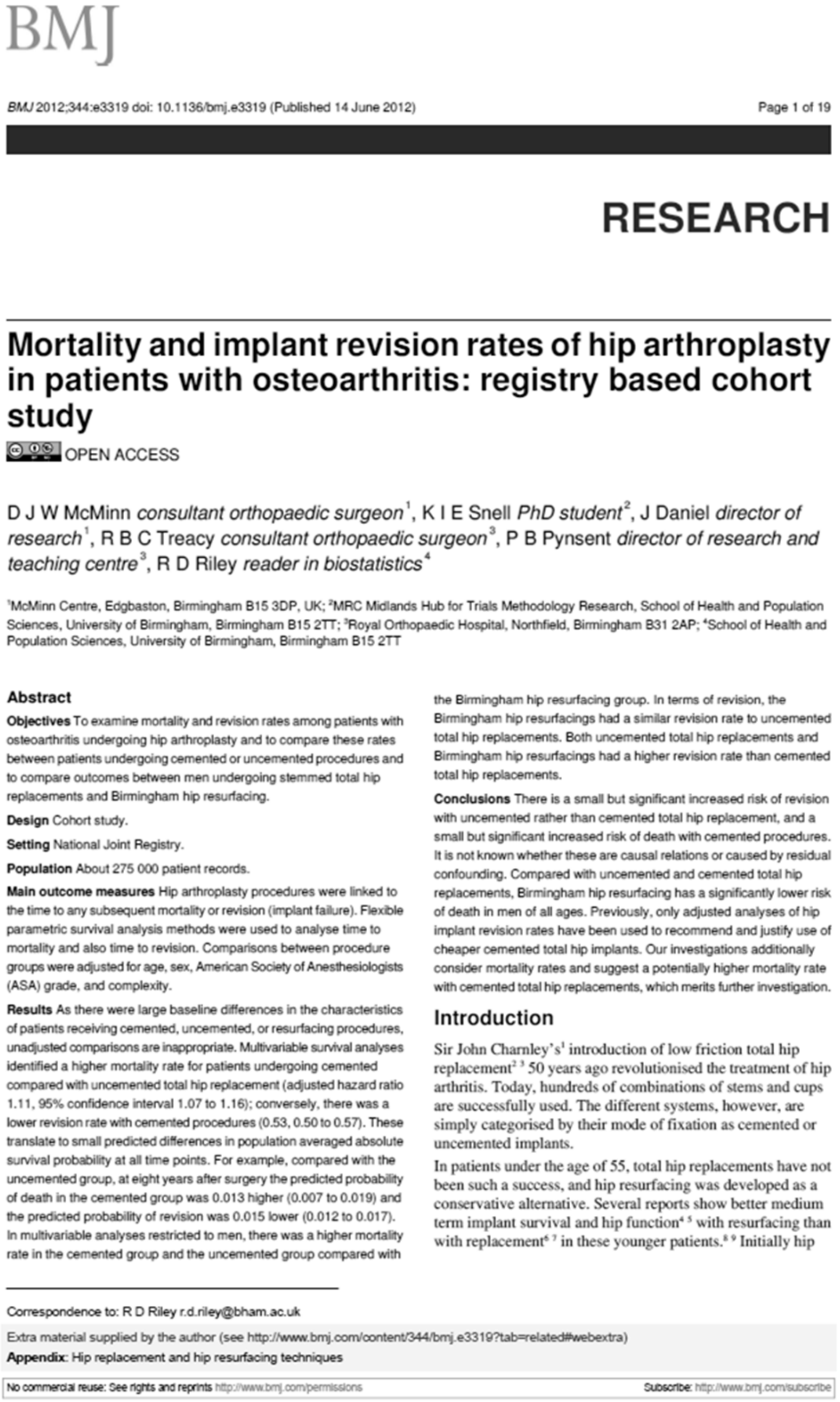


Figure A4.2: Adjusted survival curve for procedure type including BHR in males where the outcome is revision.

Appendix A5: Article published in the BMJ



BMJ

BMJ 2012;344:e3319 doi: 10.1136/bmj.e3319 (Published 14 June 2012) Page 1 of 19

RESEARCH

Mortality and implant revision rates of hip arthroplasty in patients with osteoarthritis: registry based cohort study

D J W McMinn *consultant orthopaedic surgeon*[1], K I E Snell *PhD student*[2], J Daniel *director of research*[1], R B C Treacy *consultant orthopaedic surgeon*[3], P B Pynsent *director of research and teaching centre*[3], R D Riley *reader in biostatistics*[4]

[1]McMinn Centre, Edgbaston, Birmingham B15 3DP, UK; [2]MRC Midlands Hub for Trials Methodology Research, School of Health and Population Sciences, University of Birmingham, Birmingham B15 2TT; [3]Royal Orthopaedic Hospital, Northfield, Birmingham B31 2AP; [4]School of Health and Population Sciences, University of Birmingham, Birmingham B15 2TT

Abstract

Objectives To examine mortality and revision rates among patients with osteoarthritis undergoing hip arthroplasty and to compare these rates between patients undergoing cemented or uncemented procedures and to compare outcomes between men undergoing stemmed total hip replacements and Birmingham hip resurfacing.

Design Cohort study.

Setting National Joint Registry.

Population About 275 000 patient records.

Main outcome measures Hip arthroplasty procedures were linked to the time to any subsequent mortality or revision (implant failure). Flexible parametric survival analysis methods were used to analyse time to mortality and also time to revision. Comparisons between procedure groups were adjusted for age, sex, American Society of Anesthesiologists (ASA) grade, and complexity.

Results As there were large baseline differences in the characteristics of patients receiving cemented, uncemented, or resurfacing procedures, unadjusted comparisons are inappropriate. Multivariable survival analyses identified a higher mortality rate for patients undergoing cemented compared with uncemented total hip replacement (adjusted hazard ratio 1.11, 95% confidence interval 1.07 to 1.16); conversely, there was a lower revision rate with cemented procedures (0.53, 0.50 to 0.57). These translate to small predicted differences in population averaged absolute survival probability at all time points. For example, compared with the uncemented group, at eight years after surgery the predicted probability of death in the cemented group was 0.013 higher (0.007 to 0.019) and the predicted probability of revision was 0.015 lower (0.012 to 0.017). In multivariable analyses restricted to men, there was a higher mortality rate in the cemented group and the uncemented group compared with the Birmingham hip resurfacing group. In terms of revision, the Birmingham hip resurfacings had a similar revision rate to uncemented total hip replacements. Both uncemented total hip replacements and Birmingham hip resurfacings had a higher revision rate than cemented total hip replacements.

Conclusions There is a small but significant increased risk of revision with uncemented rather than cemented total hip replacement, and a small but significant increased risk of death with cemented procedures. It is not known whether these are causal relations or caused by residual confounding. Compared with uncemented and cemented total hip replacements, Birmingham hip resurfacing has a significantly lower risk of death in men of all ages. Previously, only adjusted analyses of hip implant revision rates have been used to recommend and justify use of cheaper cemented total hip implants. Our investigations additionally consider mortality rates and suggest a potentially higher mortality rate with cemented total hip replacements, which merits further investigation.

Introduction

Sir John Charnley's's[1] introduction of low friction total hip replacement[2 3] 50 years ago revolutionised the treatment of hip arthritis. Today, hundreds of combinations of stems and cups are successfully used. The different systems, however, are simply categorised by their mode of fixation as cemented or uncemented implants.

In patients under the age of 55, total hip replacements have not been such a success, and hip resurfacing was developed as a conservative alternative. Several reports show better medium term implant survival and hip function[4 5] with resurfacing than with replacement[6 7] in these younger patients.[8 9] Initially hip

Correspondence to: R D Riley r.d.riley@bham.ac.uk

Extra material supplied by the author (see http://www.bmj.com/content/344/bmj.e3319?tab=related#webextra)

Appendix: Hip replacement and hip resurfacing techniques

Appendix B

Appendix B1: Literature review protocol

General Information

- Author Names
- Journal of publication
- Year of publication
- What clinical / disease area was the model developed for?
- What is the starting point for follow-up?
- What is the outcome or event of interest?
- What is the primary aim of the article?
- What is the study design (e.g. cohort, case-control, trial)?
- Was the data collected to answer the study question (prospective), or was existing data used to answer the study question (retrospective use of existing data)? If existing data were used, what was the aim of the original study that the data were collected for?
- How large is the study (sample size)? Full dataset size and size of data used for modelling (if reduced, e.g. subset analysis, due to missing data etc.).
- How many events were there?
- How many candidate predictors were there?
- What is the length of follow-up (e.g. min, max, median follow-up, depending on what is reported)?

Model Development

- What modelling techniques were used to obtain the prognostic model: parametric, semi-parametric or non-parametric?

- Were time-dependent covariates considered in the survival analysis?
- Were continuous variables modelled linearly, non-linearly or categorised? If categorised, how were the cut-points decided? If modelled linearly, were the variables transformed? If modelled non-linearly, how were they modelled?
- What process was used to select variables for inclusion in the model? (E.g. automatic selection process, selection based on univariable results)
- If multiple studies (datasets) were available, how have they been used? Have they been combined to fit the prognostic model or were some used for model development and others used for validation?
- Was missing data a problem and if it was, how was missing data handled?

Baseline Hazard

- Has the baseline hazard been modelled explicitly? If so, what method was used to model the baseline (cumulative) hazard function?

Reporting of Results

- How have the model results been reported? Detail the following:
  i. Are the full fitted model parameter estimates reported? If not, what is given from the fitted model?
  ii. Do they report the baseline hazard function?
  iii. Are the 'beta' estimates reported on their original scale (i.e. log hazard ratios), or reported transformed (e.g. hazard ratios), or both?
  iv. How is absolute risk summarised (e.g. is probability of survival at time t given for all patients (Kaplan-Meier curve), is it given for individuals with particular covariate values, or perhaps a risk scoring system is presented, etc.)?
  v. How is the model intended to be used in clinical practice?

Validation

- Was the model validated before publication? If so:

  i. What validation was performed i.e. internal, external or temporal? Include brief description on how this was done.

  ii. What validation techniques were used and what validation statistics were presented?

  iii. Was the baseline hazard function (alpha term) compared in the validation data to the development data?

  iv. If no validation or internal validation was performed, is there any discussion on how the estimated baseline hazard can be used toward different patient populations or how the authors suggest dealing with changes in baseline hazard?

  v. Were absolute risk probabilities computed in the validation data and compared to the predicted ones? If so, how was this done?

Appendix B2: Literature review exclusions

Table B2.1: Articles excluded from current literature review with reason for exclusion.

First Author, year	Reason for exclusion
Not modelling a rate over time (33 articles excluded)	
Smits, 2007	Logistic regression
van der Steeg, 2007	Logistic regression
Cornelis, 2009	Logistic regression
Billings, 2006	Logistic regression
Aylin, 2007	Logistic regression
Frank, 2008	Logistic regression
Perel, 2008	Logistic regression
Lee, 2006	Logistic regression
Robbins, 2007	Logistic regression
Wijeysundera, 2007	Logistic regression
Green, 2008	Logistic regression
Burroughs, 2006	Logistic regression
Johnston, 2007	Logistic regression
Carlin, 2008	Logistic regression
Puhan, 2009	Logistic regression
Lyssenko, 2008	Logistic regression
Peacock, 2008	Logistic regression
Lopman, 2006	Logistic regression
Maitland, 2006	Logistic regression
Hughes, 2007	Logistic regression
Steyerberg, 2008	Logistic regression
Whitely, 2009	Logistic regression
Tyson, 2008	Logistic regression
Meigs, 2008	Logistic regression & GEEs
Potti, 2006	Binary classification-tree analysis & logistic regression
Dehghan, 2008	Linear regression
Lyness, 2006	Linear mixed-effects models, proportional-odds regression
Freemantle, 2009	Poisson mixed models (for live births not time to event)
Reis, 2009	Naive Bayesian classifiers (for event at *any* time)
Tamborlane, 2008	ANCOVA
Chaves, 2006	Time series analysis & forecasting
Drake, 2006	Stochastic epidemic models & forecasting
Kelen, 2006	Disposition classification system
Validation only study (5 articles excluded)	
Parmigiani, 2007	No model development
Collins, 2009	No model development
Nigrovic, 2007	No model development
Morrison, 2006	No model development
Lapidus, 2009	No model development
Time-to-event data not right censored	
Parikh, 2008	Interval censoring
Kahn, 2009	Interval censoring

References of articles excluded from literature review in Chapter 3:

Aylin P, Bottle A, Majeed A. Use of administrative data or clinical databases as predictors of risk of death in hospital: comparison of models. *British Medical Journal* 2007; 334(7602):1044-1047.

Billings J, Dixon J, Mijanovich T, Wennberg D. Case finding for patients at risk of readmission to hospital: development of algorithm to identify high risk patients. *British Medical Journal* 2006; 333(7563):327-330.

Burroughs AK, Sabin CA, Rolles K, Delvart V, Karam V, Buckels J et al. 3-month and 12-month mortality after first liver transplant in adults in Europe: predictive models for outcome. *Lancet* 2006; 367(9506):225-232.

Carlin JB, Darmstadt GL, Hamer DH, Weber MW, Chowdhury A, Saha S et al. Clinical signs that predict severe illness in children under age 2 months: a multicentre study. *Lancet* 2008; 371(9607):135-142.

Chaves LF, Pascual M. Climate cycles and forecasts of cutaneous leishmaniasis, a nonstationary vector-borne disease. *Plos Medicine* 2006; 3(8):1320-1328.

Collins GS, Altman DG. An independent external validation and evaluation of QRISK cardiovascular risk prediction: a prospective open cohort study. *British Medical Journal* 2009; 339.

Cornelis MC, Qi L, Zhang CL, Kraft P, Manson J, Cai TX et al. Joint Effects of Common Genetic Variants on the Risk for Type 2 Diabetes in U. S. Men and Women of European Ancestry. *Annals of Internal Medicine* 2009; 150(8):541-W98.

Dehghan A, Kottgen A, Yang Q, Hwang SJ, Kao WHL, Rivadeneira F et al. Association of three genetic loci with uric acid concentration and risk of gout: a genome-wide association study. *Lancet* 2008; 372(9654):1953-1961.

Drake JM. Limits to forecasting precision for outbreaks of directly transmitted diseases. *Plos Medicine* 2006; 3(1):57-62.

Frank PI, Morris JA, Hazell ML, Linehan MF, Frank TL. Long term prognosis in preschool children with wheeze: longitudinal postal questionnaire study 1993-2004. *British Medical Journal* 2008; 336(7658):1423-+.

Freemantle N, Wood J, Griffin C, Gill P, Calvert MJ, Shankar A et al. What factors predict differences in infant and perinatal mortality in primary care trusts in England? A prognostic model. *British Medical Journal* 2009; 339.

Green BB, Cook AJ, Ralston JD, Fishman PA, Catz SL, Carlson J et al. Effectiveness of home blood pressure monitoring, Web communication, and pharmacist care on hypertension control - A randomized controlled trial. *Jama-Journal of the American Medical Association* 2008; 299(24):2857-2867.

Hughes AE, Orr N, Patterson C, Esfandiary H, Hogg R, McConnell V et al. Neovascular age-related macular degeneration risk based on CFH, LOC387715/HTRA1, and smoking. *Plos Medicine* 2007; 4(12):1993-2000.

Johnston SC, Rothwell PM, Nguyen-Huynh MN, Giles MF, Elkins JS, Bernstein AL et al. Validation and refinement of scores to predict very early stroke risk after transient ischaemic attack. *Lancet* 2007; 369(9558):283-292.

Kahn HS, Cheng YJ, Thompson TJ, Imperatore G, Gregg EW. Two Risk-Scoring Systems for Predicting Incident Diabetes Mellitus in US Adults Age 45 to 64 Years. *Annals of Internal Medicine* 2009; 150(11):741-W134.

Kelen GD, Kraus CK, McCarthy ML, Bass E, Hsu EB, Li GH et al. Inpatient disposition classification for the creation of hospital surge capacity: a multiphase study. *Lancet* 2006; 368(9551):1984-1990.

Lapidus N, Luquero FJ, Gaboulaud V, Shepherd S, Grais RF. Prognostic Accuracy of WHO Growth Standards to Predict Mortality in a Large-Scale Nutritional Program in Niger. *Plos Medicine* 2009; 6(3).

Lee SJ, Lindquist K, Segal MR, Covinsky KE. Development and validation of a prognostic index for 4-year mortality in older adults. *Jama-Journal of the American Medical Association* 2006; 295(7):801-808.

Lopman BA, Barnabas RV, Boerma JT, Chawira G, Gaitskell K, Harrop T et al. Creating and validating an algorithm to measure AIDS mortality in the adult population using verbal autopsy. *Plos Medicine* 2006; 3(8):1273-1281.

Lyness JM, Heo M, Datto CJ, Ten Have TR, Katz IR, Drayer R et al. Outcomes of minor and subsyndromal depression among elderly patients in primary care settings. *Annals of Internal Medicine* 2006; 144(7):496-504.

Lyssenko V, Jonsson A, Almgren P, Pulizzi N, Isomaa B, Tuomi T et al. Clinical Risk Factors, DNA Variants, and the Development of Type 2 Diabetes. *New England Journal of Medicine* 2008; 359(21):2220-2232.

Maitland K, Berkley JA, Shebbe M, Peshu N, English M, Newton CRJC. Children with severe malnutrition: Can those at highest risk of death be identified with the WHO protocol? *Plos Medicine* 2006; 3(12):2431-2439.

Meigs JB, Shrader P, Sullivan LM, McAteer JB, Fox CS, Dupuis J et al. Genotype Score in Addition to Common Risk Factors for Prediction of Type 2 Diabetes. *New England Journal of Medicine* 2008; 359(21):2208-2219.

Morrison LJ, Visentin LM, Kiss A, Theriault R, Eby D, Vermeulen M et al. Validation of a rule for termination of resuscitation in out-of-hospital cardiac arrest. *New England Journal of Medicine* 2006; 355(5):478-487.

Nigrovic LE, Kuppermann N, Macias CG, Cannavino CR, Moro-Sutherland DM, Schremmer RD et al. Clinical prediction rule for identifying children with cerebrospinal fluid pleocytosis at very low risk of bacterial meningitis. *Jama-Journal of the American Medical Association* 2007; 297(1):52-60.

Parikh NI, Pencina MJ, Wang TJ, Benjamin EJ, Lanier KJ, Levy D et al. A risk score for predicting near-term incidence of hypertension: The Framingham Heart Study. *Annals of Internal Medicine* 2008; 148(2):102-110.

Parmigiani G, Chen S, Iversen ES, Friebel TM, Finkelstein DM, Anton-Culver H et al. Validity of models for predicting BRCA1 and BRCA2 mutations. *Annals of Internal Medicine* 2007; 147(7):441-450.

Peacock WF, De Marco T, Fonarow GC, Diercks D, Wynne J, Apple FS et al. Cardiac troponin and outcome in acute heart failure. *New England Journal of Medicine* 2008; 358(20):2117-2126.

Perel P, Arango M, Clayton T, Edwards P, Komolafe E, Pocock S et al. Predicting outcome after traumatic brain injury: practical prognostic models based on large cohort of international patients. *British Medical Journal* 2008; 336(7641):425-429.

Potti A, Mukherjee S, Petersen R, Dressman HK, Bild A, Koontz J et al. A genomic strategy to refine prognosis in early-stage non-small-cell lung cancer. *New England Journal of Medicine* 2006; 355(6):570-580.

Potti A. A genomic strategy to refine prognosis in early-stage non-small-cell lung cancer (Retraction of vol 355, pg 570, 2006). *New England Journal of Medicine* 2007; 356(2):201-202.

Puhan MA, Garcia-Aymerich J, Frey M, ter Riet G, Anto JM, Agusti AG et al. Expansion of the prognostic assessment of patients with chronic obstructive pulmonary disease: the updated BODE index and the ADO index. *Lancet* 2009; 374(9691):704-711.

Reis BY, Kohane IS, Mandl KD. Longitudinal histories as predictors of future diagnoses of domestic abuse: modelling study. *British Medical Journal* 2009; 339.

Robbins J, Aragaki AK, Kooperberg C, Watts N, Wactawski-Wende J, Jackson RD et al. Factors associated with 5-year risk of hip fracture in postmenopausal women. *Jama-Journal of the American Medical Association* 2007; 298(20):2389-2398.

Smits M, Dippel DWJ, Steyerberg EW, de Haan GG, Dekker HM, Vos PE et al. Predicting intracranial traumatic findings on computed tomography in patients with minor head injury: The CHIP prediction rule. *Annals of Internal Medicine* 2007; 146(6):397-405.

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):1251-1261.

Tamborlane WV, Beck RW, Bode BW, Buckingham B, Chase HP, Clemons R et al. Continuous glucose monitoring and intensive treatment of type 1 diabetes. *New England Journal of Medicine* 2008; 359(14):1464-1U65.

Tyson JE, Parikh NA, Langer J, Green C, Higgins RD. Intensive care for extreme prematurity - Moving beyond gestational age. *New England Journal of Medicine* 2008; 358(16):1672-1681.

van der Steeg WA, Boekholdt SM, Stein EA, El-Harchaoui K, Stroes ESG, Sandhu MS et al. Role of the apolipoprotein B-apolipoprotein A-I ratio in cardiovascular risk assessment: A case-control analysis in EPIC-Norfolk. *Annals of Internal Medicine* 2007; 146(9):640-648.

Whiteley W, Jackson C, Lewis S, Lowe G, Rumley A, Sandercock P et al. Inflammatory Markers and Poor Outcome after Stroke: A Prospective Cohort Study and Systematic Review of Interleukin-6. *Plos Medicine* 2009; 6(9).

Wijeysundera DN, Karkouti K, Dupuis JY, Rao V, Chan CT, Granton JT et al. Derivation and validation of a simplified predictive index for renal replacement therapy after cardiac surgery. *Jama-Journal of the American Medical Association* 2007; 297(16):1801-1809.

Appendix B3: Kalbfleisch & Prentice estimator

The Kalbfleisch & Prentice estimator is an extension of the Kaplan-Meier estimator and will give the same survival function estimates as Kaplan-Meier when there are no covariates in the model. However, when covariates are included in the model, this estimator can be used to estimate the baseline survival function (or adjusted survival curves) as a step-function.

The maximum likelihood estimate of the baseline survival function is

$$\widehat{S}_0(t)=\prod_{i|t_i\leq t}\widehat{\alpha}_i$$

where t_i represents each unique failure time. The $\alpha_i = 1\text{-}\lambda_i$ where λ_i is the hazard at time t_i.[22] By differentiating the log of the partial likelihood function for the Cox model and rearranging for α_i, the following equation can be solved for the maximum likelihood estimate of α_i if there are no ties at time t_i.

$$\widehat{\alpha}_i=\left[1-\frac{\exp(X_{(i)}\widehat{\beta})}{\sum_{j\in R(t_i)}\exp(X_j\widehat{\beta})}\right]^{\exp(-X_{(i)}\widehat{\beta})}$$

where $X_{(i)}$ denotes the covariate values for the subject that failed at time t_i and $R(t_i)$ is the set of all individuals that have not had the event or been censored just prior to time $t_{(i)}$. If there is more than one failure at time t_i (i.e. tied observations), then this would have to be solved iteratively.

Appendix C

Appendix C1: Previously published results

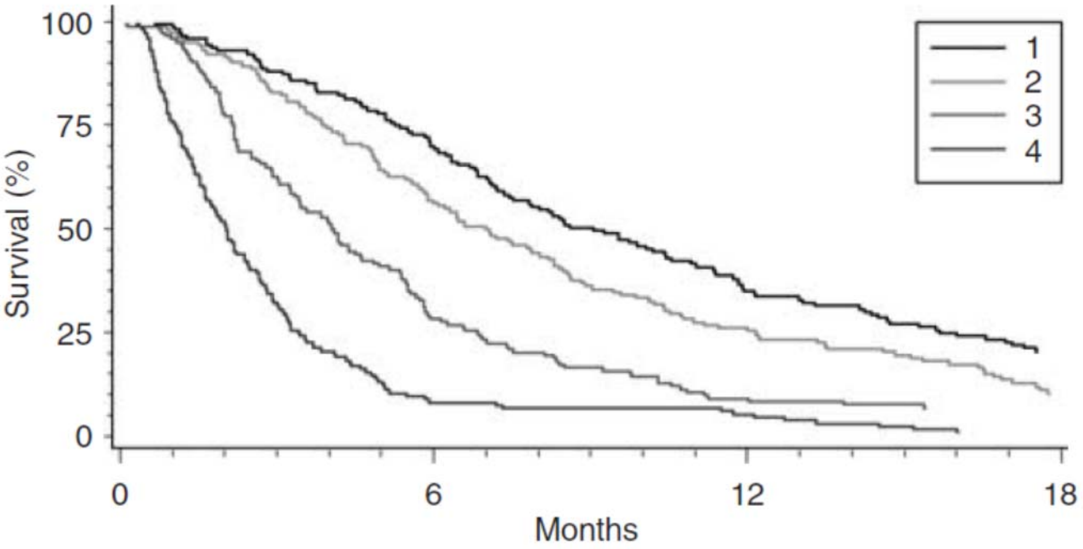


Figure C1.1: Kaplan-Meier survival curves for risk groups 1-4 from Stocken model for advanced pancreatic cancer. Reprinted by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: British Journal of Cancer, Stocken et al.[156] Copyright 2008.

Appendix C2: Proportional hazards assumption

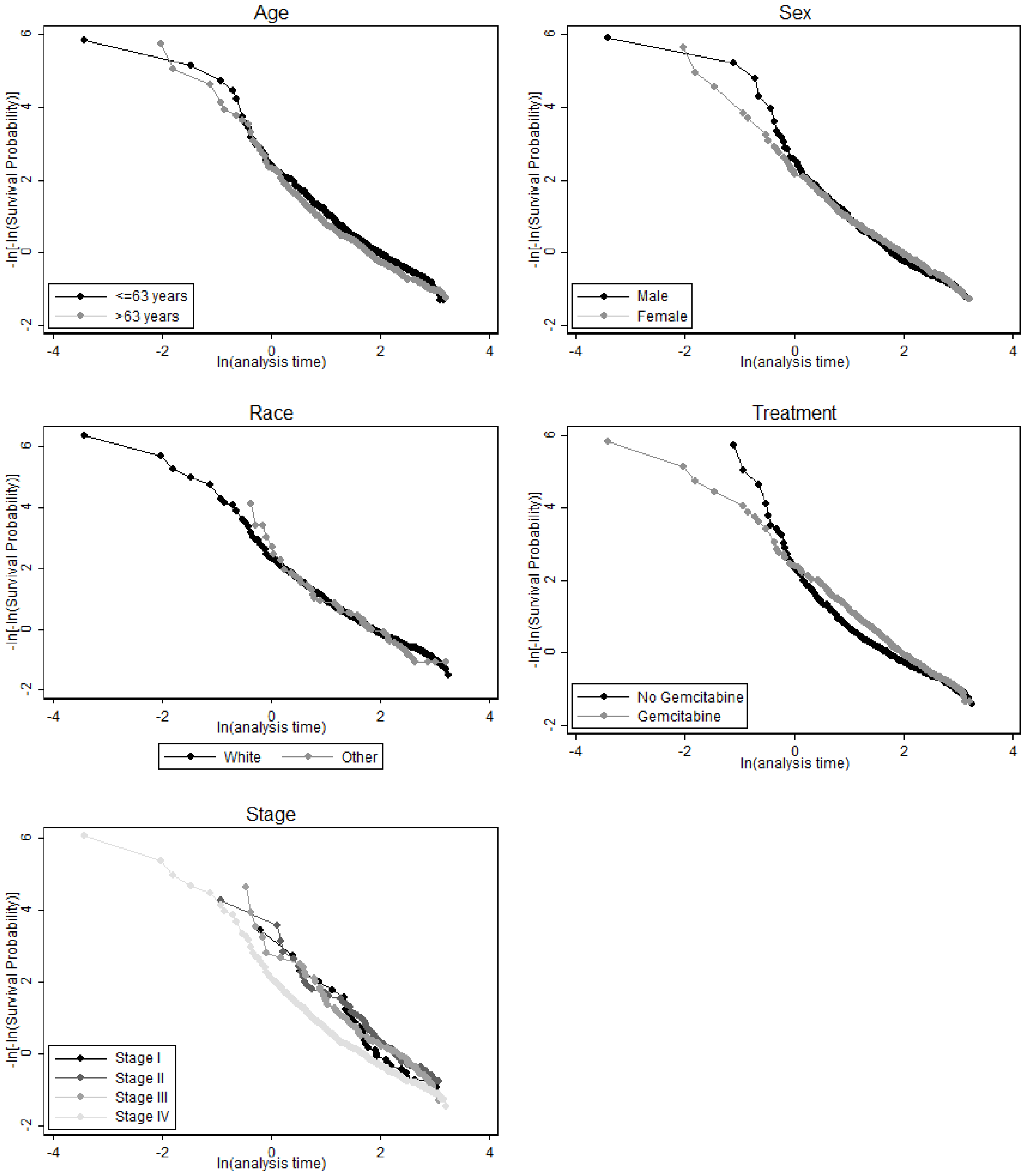


Figure C2.1: Log-log plots for demographic and cancer variables with death as the outcome.

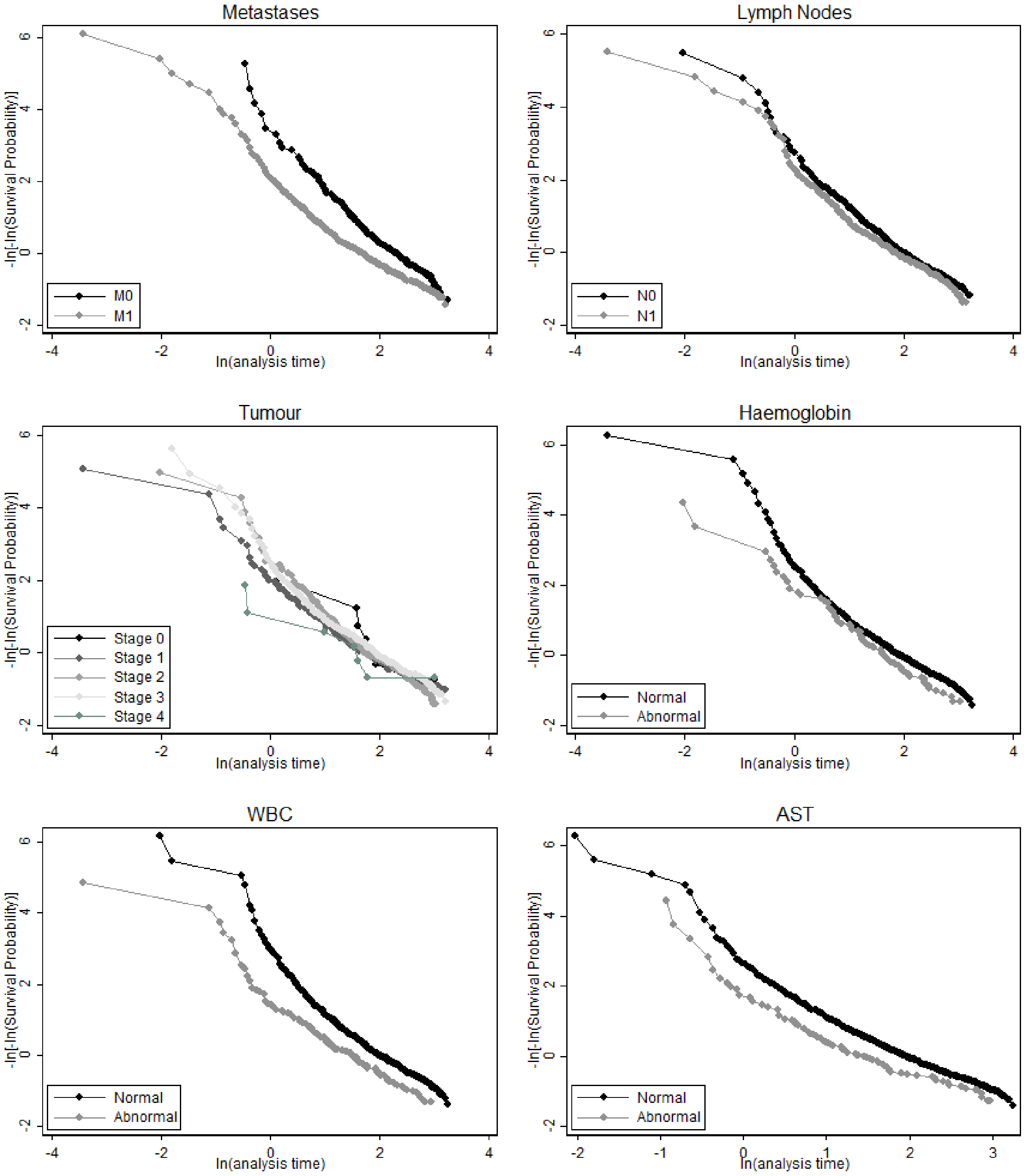


Figure C2.2: Log-log plots for cancer variables and laboratory variables with death as the outcome.

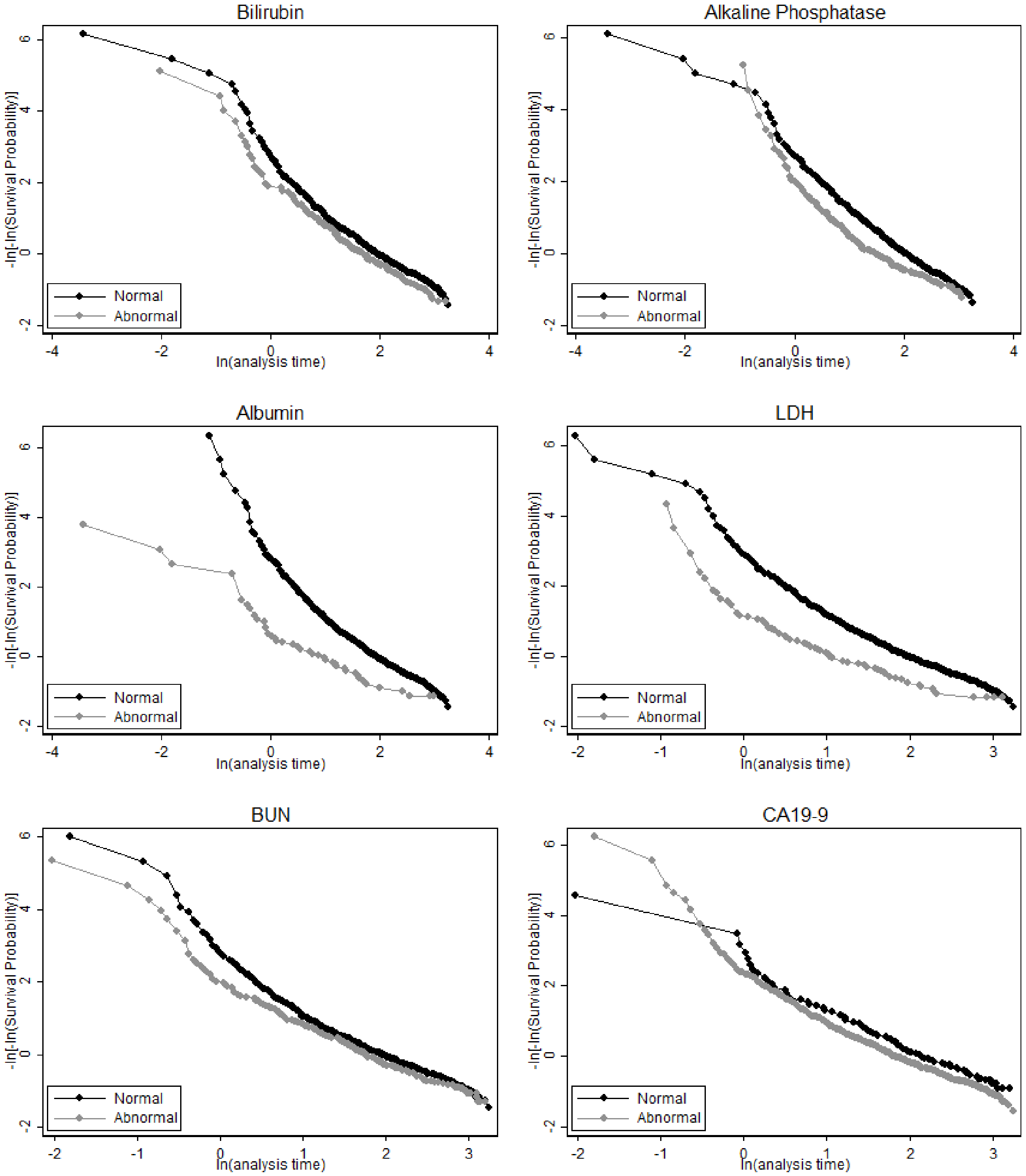


Figure C2.3: Log-log plots for laboratory variables with death as the outcome.

Appendix D

Appendix D1: R-code used to calculate joint probabilities for performance statistics

```
# load library
library("mnormt")

# set seed
set.seed(86454)

########################################################################
#   Function using the number of samples (N) as input                  #
#   mu = mean vector, Sigma = var-cov matrix,                          #
#   df = degrees of freedom and validation criteria=a,b & c            #
########################################################################

#Function for joint probability of satisfying validation criteria for DVT model
f_jointprob1<-function(N,mu,Sigma,df,a,b,c){

  #Draw random samples from bivariate t-distribution
  Nsamples<-rmt(N,mu,Sigma,df)

  #Define first column from the random samples
  col1<-Nsamples[,1]

  #Define second column from the random samples
  col2<-Nsamples[,2]

  #Set count to zero so that we can count the number of samples for which
   the validation criteria are satisfied
  count<-0

  #For each row from 1 to N:
  for(i in 1:N){

    #Check the entries of both columns satisfy the validation criteria
    if(col1[i]>=a  & col1[i]<=b & col2[i]>=c){

      #If criteria are satisfied, add 1 to count
      count<-count+1
    }
  }

  #Return proportion of samples meeting criteria as output
  return=count/N
}

#Function for joint probability of satisfying validation criteria for Breast
Cancer model
f_jointprob2<-function(N,mu,Sigma,df,d,e,f){

  #Draw random samples from bivariate t-distribution
  Nsamples2<-rmt(N,mu,Sigma,df)
  #Define first column from the random samples
  col1<-Nsamples2[,1]

  #Define second column from the random samples
  col2<-Nsamples2[,2]

  #Set count to zero so that we can count the number of samples for which
   the validation criteria are satisfied
  count<-0

  #For each row from 1 to N:
  for(i in 1:N){

    #Check the entries of both columns satisfy the validation criteria
    if(col1[i]>=d  & col2[i]<=e & col2[i]>=f){

      #If criteria are satisfied, add 1 to count
      count<-count+1
    }
  }

  #Return proportion of samples meeting criteria as output
  return=count/N
}

########################################################################
#   Input for samples - DVT model, Strategy 1                          #
########################################################################
N<-500000
mu <- c(0.9750971,0.6866886)
Sigma <- matrix(c(0.02911928,0.00325565, 0.00325565,0.00037329), 2, 2)
df <- 10
a<-0.9
b<-1.1
c<-0.7

###############
#   Output    #
###############
proportion_met1<-f_jointprob1(N,mu,Sigma,df,a,b,c)

########################################################################
#   Input for samples - DVT model, Strategy 2                          #
########################################################################
N<-500000
mu2 <- c(0.9751832,0.6863293)
Sigma2 <- matrix(c(0.03012634,0.00337928, 0.00337928,0.00038859), 2, 2)
df <- 10
a<-0.9
b<-1.1
c<-0.7

###############
#   Output    #
###############
proportion_met2<-f_jointprob1(N,mu2,Sigma2,df,a,b,c)
```

```
########################################################################
#   Input for samples - DVT model, Strategy 3                          #
########################################################################
N<-500000
mu3 <- c(0.9716777,0.6859794)
Sigma3 <- matrix(c(0.0295297,0.00331324, 0.00331324,0.00038108), 2, 2)
df <- 10
a<-0.9
b<-1.1
c<-0.7

###############
#   Output    #
###############
proportion_met3<-f_jointprob1(N,mu3,Sigma3,df,a,b,c)

########################################################################
#   Input for samples - Breast Cancer model, C-stat & D-stat           #
########################################################################
N<-500000
mu4 <- c(0.71044666,0.3268395)
Sigma4 <- matrix(c(0.00050658,-0.00027979,-0.00027979,0.02601139), 2, 2)
df <- 6
d<-0.7
e<-99999
f<-0.3

###############
#   Output    #
###############
proportion_met4<-f_jointprob2(N,mu4,Sigma4,df,d,e,f)

#########################################################################
#   Input for samples - Breast Cancer model, C-stat & calib (new)       #
#########################################################################
N<-500000
mu5 <- c(0.711926,1.001583)
Sigma5 <- matrix(c(0.00054073,0.00070951,0.00070951,0.00109693), 2, 2)
df <- 6
d<-0.7
e<-1.1
f<-0.9

###############
#   Output    #
###############
proportion_met5<-f_jointprob2(N,mu5,Sigma5,df,d,e,f)
```

```
########################################################################
#   Input for samples - Breast Cancer model, C-stat & calib (average)  #
########################################################################
N<-500000
mu6 <- c(0.7104769,0.9940746)
Sigma6 <- matrix(c(0.00056394,0.0001089,0.0001089,0.05578766), 2, 2)
df <- 6
d<-0.7
e<-1.1
f<-0.9

###############
#   Output    #
###############
proportion_met6<-f_jointprob2(N,mu6,Sigma6,df,d,e,f)

########################################################################
#   Input for samples - Breast Cancer model, C-stat & calib (closest)  #
########################################################################
N<-500000
mu7 <- c(0.7094562,0.9573295 )
Sigma7 <- matrix(c(0.00063688,-0.00226936,-0.00226936,0.1100915), 2, 2)
df <- 6
d<-0.7
e<-1.1
f<-0.9

###############
#   Output    #
###############
proportion_met7<-f_jointprob2(N,mu7,Sigma7,df,d,e,f)

########################################################################
#   Input for samples - Breast Cancer model, D-stat & calib (new)      #
########################################################################
N<-500000
mu8 <- c(0.3199665,0.9985328)
Sigma8 <- matrix(c(0.02378233,-0.00139718,-0.00139718,0.00041476), 2, 2)
df <- 6
d<-0.3
e<-1.1
f<-0.9

###############
#   Output    #
###############
proportion_met8<-f_jointprob2(N,mu8,Sigma8,df,d,e,f)
```

```
########################################################################
#   Input for samples - Breast Cancer model, D-stat & calib (average)  #
########################################################################
N<-500000
mu9 <- c(0.3345175,0.9945885)
Sigma9 <- matrix(c(0.02826252,0.03439829,0.03439829,0.05663072), 2, 2)
df <- 6
d<-0.3
e<-1.1
f<-0.9

###############
#   Output    #
###############
proportion_met9<-f_jointprob2(N,mu9,Sigma9,df,d,e,f)

########################################################################
#   Input for samples - Breast Cancer model, D-stat & calib (closest)  #
########################################################################
N<-500000
mu10 <- c(0.3369767,0.9619958)
Sigma10 <- matrix(c(0.02864098,0.03283924,0.03283924,0.10922132), 2, 2)
df <- 6
d<-0.3
e<-1.1
f<-0.9

###############
#   Output    #
###############
proportion_met10<-f_jointprob2(N,mu10,Sigma10,df,d,e,f)
```

Appendix D2: Sensitivity analysis for breast cancer data

Table D2.1: Bivariate random-effects meta-analysis results of calibration and discrimination performance for the breast cancer model including and excluding Sweden, for implementation strategy 2 (average baseline hazard).

Sensitivity analysis	Validation statistic	Pooled estimate (95% CI)	95% prediction interval	I^2 %	Estimate of τ	Joint probability of ‘good’ performance in a new population*
IECV including all countries	Calibration slope	0.994 (0.836 to 1.153)	0.416 to 1.572	98	0.221	0.21
	C-statistic	0.710 (0.688 to 0.733)	0.660 to 0.769	52	0.021	
IECV excluding Sweden	Calibration slope	0.999 (0.883 to 1.114)	0.594 to 1.404	95	0.146	0.32
	C-statistic	0.712 (0.688 to 0.735)	0.650 to 0.773	52	0.021	

* Good performance defined by a C-statistic≥0.7 and a calibration slope between 0.9 and 1.1.

Appendix D3: Trivariate random-effects meta-analysis

Table D3.1: Trivariate random-effects meta-analysis results for the DVT model performance statistics, using different implementation strategies for the intercept.

	Performance statistic	Pooled estimate (SE)	95% confidence interval for pooled estimate	Marginal 95% prediction interval for pooled estimate	I^2 % (approx. 95% CI for I^2)	τ estimate (approx. 95% CI for τ)
Strategy 1: Intercept estimated in validation study	Calibration-in-the-large	-0.130 (0.028)	-0.185 to -0.075	-0.195 to -0.065	1	0.008 (0.000 to 0.054)
	Calibration slope	0.975 (0.062)	0.854 to 1.097	0.597 to 1.353	57	0.158 (0.059 to 0.258)
	Log(expected/observed)	0.086 (0.019)	0.047 to 0.124	0.041 to 0.128	0	0.0009 (0.000 to 0.034)
	C-statistic	0.687 (0.009)	0.670 to 0.704	0.645 to 0.729	34	0.017 (0.004 to 0.031)
Strategy 2: Average intercept taken from derived random-intercept model	Calibration-in-the-large	-0.004 (0.158)	-0.313 to 0.305	-1.240 to 1.232	97	0.532 (0.301 to 0.763)
	Calibration slope	0.980 (0.065)	0.853 to 1.107	0.585 to 1.357	59	0.165 (0.063 to 0.268)
	Log(expected/observed)	0.022 (0.116)	-0.206 to 0.250	-0.887 to 0.931	97	0.390 (0.220 to 0.560)
	C-statistic	0.687 (0.009)	0.669 to 0.705	0.640 to 0.734	37	0.019 (0.004 to 0.033)
Strategy 3: Intercept from a study included in derivation set with a similar prevalence	Calibration-in-the-large	0.047 (0.085)	-0.120 to 0.214	-0.584 to 0.678	89	0.270 (0.136 to 0.404)
	Calibration slope	0.976 (0.064)	0.851 to 1.102	0.578 to 1.375	59	0.167 (0.061 to 0.272)
	Log(expected/observed)	-0.029 (0.062)	-0.150 to 0.093	-0.485 to 0.427	89	0.195 (0.094 to 0.295)
	C-statistic	0.687 (0.009)	0.669 to 0.705	0.640 to 0.734	38	0.019 (0.004 to 0.034)

* A trivariate meta-analysis was fitted to calibration-in-the-large, calibration slope and C-statistic, and then again for log(Expected/Observed), calibration slope, and C-statistic. Perfect negative correlation between calibration-in-the-large and expected/observed within studies prevents all four statistics being analysed together (due to collinearity). Results were practically the same for calibration slope and C-statistic, regardless of the trivariate model fitted.

Appendix E

Appendix E1: Sample size and number of studies

The first stage was to agree upon the number of studies and the number of individuals within studies to be used in all simulations. The histograms of C-statistics calculated using different sample sizes (rows) and numbers of studies (columns) are shown below in Figure E1.1. The size of the samples has a greater impact on the variability of the estimated C-statistic than the number of studies used. Based on the observed graphs in Figure E1.1, 500000 individuals was selected as appropriate. The distribution is narrower compared to 100000 individuals, and increasing the sample size to 1000000 individuals only results in a slightly narrower distribution but would increase computation time considerably.

The number of studies was chosen to be 1000. Based on the histograms, 500 would probably be adequate, however the main aim of this chapter is to evaluate the shape of the true distributions for performance statistics, therefore 1000 studies was considered more appropriate to better show the true distributional shape.

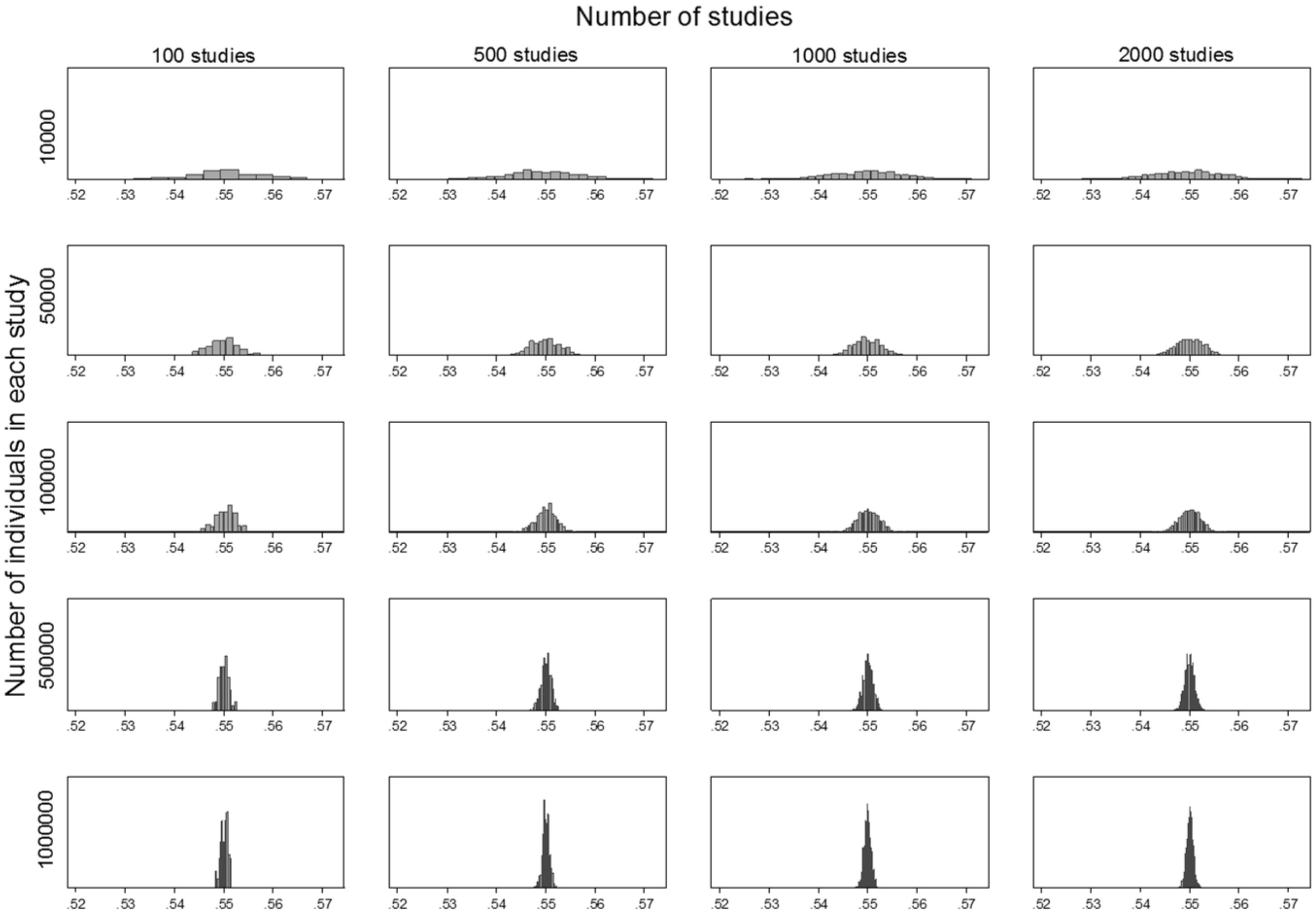


Figure E1.1: Distribution of C-statistic using different number of studies and different number of individuals within each study. Note: columns show different number of studies and rows show different number of individuals within each study.

Appendix E2: Example Stata code for simulation study

An example of a Stata do-file used to generate data and calculate the performance statistics. Below is the do-file for scenario 1 (base case) with large heterogeneity in alpha.

```
*********************************************************************************
*                   Scenario 1: Large heterogeneity in alpha                    *
*********************************************************************************
*Set working directory
cd "U:\PhD\Simulation study\Sim results\Alpha heterogeneity"

*Set up postfile to save performance statistics after each study has been generated
tempname perf
tempfile perf_stats
postfile `perf' pop alpha beta prevalence exp obs eo cstat cslope citl using
`perf_stats', replace

*Set local macros for no. of studies, no. of individuals in each study, predictor
values, alpha & beta
local i=0
local p=1000
local size=500000
local beta_mu=0.010
local beta_sd=0
local alpha_mu=-1.274
local alpha_sd=1
local x1_mu=0
local x1_sd=17.6

*Loop for each study
qui forvalues pop=1/`p' {
       noi _dots `pop' 0
       local ++i
       local alpha=rnormal(`alpha_mu',`alpha_sd')
       local beta=rnormal(`beta_mu',`beta_sd')

       *Generate data for each population
       clear
       set obs `size'
       gen alpha=`alpha'
       gen beta=`beta'
       gen x1=round(rnormal(`x1_mu',`x1_sd'))
       gen lp=`alpha'+`beta'*x1
       gen p=exp(lp)/(1+exp(lp))
       gen outcome=uniform()<=p

       *Prevalence
       summ outcome
       local prevalence=r(mean)

       *Fit true model (fixed alpha and beta)
       logistic outcome x1, coef iterate(0) from(`beta_mu' `alpha_mu', copy)
       *Calculate performance statistics for each population
       *Expected/observed
       predict xb, xb
       predict prob, pr
       summ prob
       local exp=r(mean)
       summ outcome
       local obs=r(mean)
       local eo=`exp'/`obs'

       *C-statistic
       lroc, nograph
       local cstat=r(area)

       *Calibration slope
       logistic outcome xb
       local cslope=_b[xb]

       *Calibration-in-the-large
       logistic outcome, offset(xb)
       local citl=_b[_cons]

       *Save performance statistics for each study to postfile
       post `perf' (`pop') (`alpha') (`beta') (`prevalence') (`exp') (`obs') (`eo')
(`cstat') (`cslope') (`citl')
}

postclose `perf'
use `perf_stats', clear
save "results 2-1-3 scenario 1 alpha het 1_0", replace

*Look at distributions of performance statistics including transformations
use "results 2-1-3 scenario 1 alpha het 1_0", clear
foreach var in eo cstat cslope {
       gen log_`var'=log(`var')
       gen logit_`var'=logit(`var')
}
gen asin_cstat=asin(cstat)
gen sqrt_eo=sqrt(eo)

*Set up postfile for summary statistics
tempname summary
tempfile summ_stats
postfile `summary' str15 var n mean sd median lci uci min max skewness kurtosis ///
       using `summ_stats', replace
foreach var in eo log_eo sqrt_eo cstat log_cstat logit_cstat asin_cstat ///
       cslope log_cslope citl {
       local varname="`var'"
       *Histogram and normal probability plots
       hist `var', name(hist2_1_3_`var'_1_0, replace) normal lcolor(ebg)
fcolor(navy)
       graph save "hist2-1-3_`var'_1_0", replace
       pnorm `var', name(pnorm2_1_3_`var'_1_0, replace) color(blue)
       graph save "pnorm2-1-3_`var'_1_0", replace

       *Summary statistics
       summ `var', detail
```

```
	local no=r(N)
	local mean=r(mean)
	local sd=r(sd)
	local med=r(p50)
	local min=r(min)
	local max=r(max)
	local skew=r(skewness)
	local kurt=r(kurtosis)

	centile `var', centile(2.5 50 97.5)
	local lci=r(c_1)
	local uci=r(c_3)

	*Save summary statistics to postfile
	post `summary' ("`varname'") (`no') (`mean') (`sd') (`med') (`lci') (`uci') 
(`min') (`max') (`skew') (`kurt')
}

postclose `summary'
preserve
	use `summ_stats', clear
	save "results summ 2-1-3 scenario 1 alpha het 1_0", replace
restore
window manage close graph _all
```

Appendix E3: Including an additional predictor and interaction

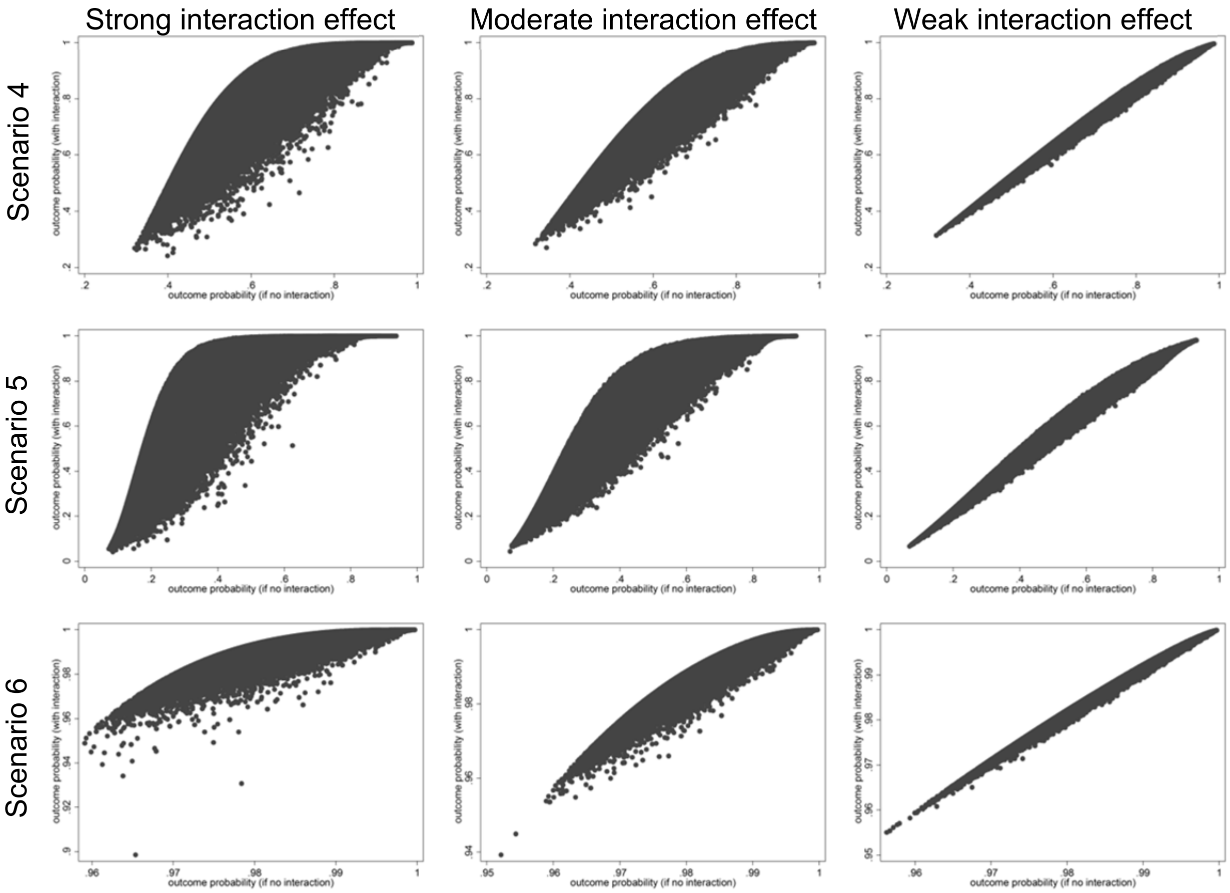


Figure E3.1: Probability of event with and without inclusion of an additional continuous predictor and interaction term for different strength interactions in scenarios 4 to 6.

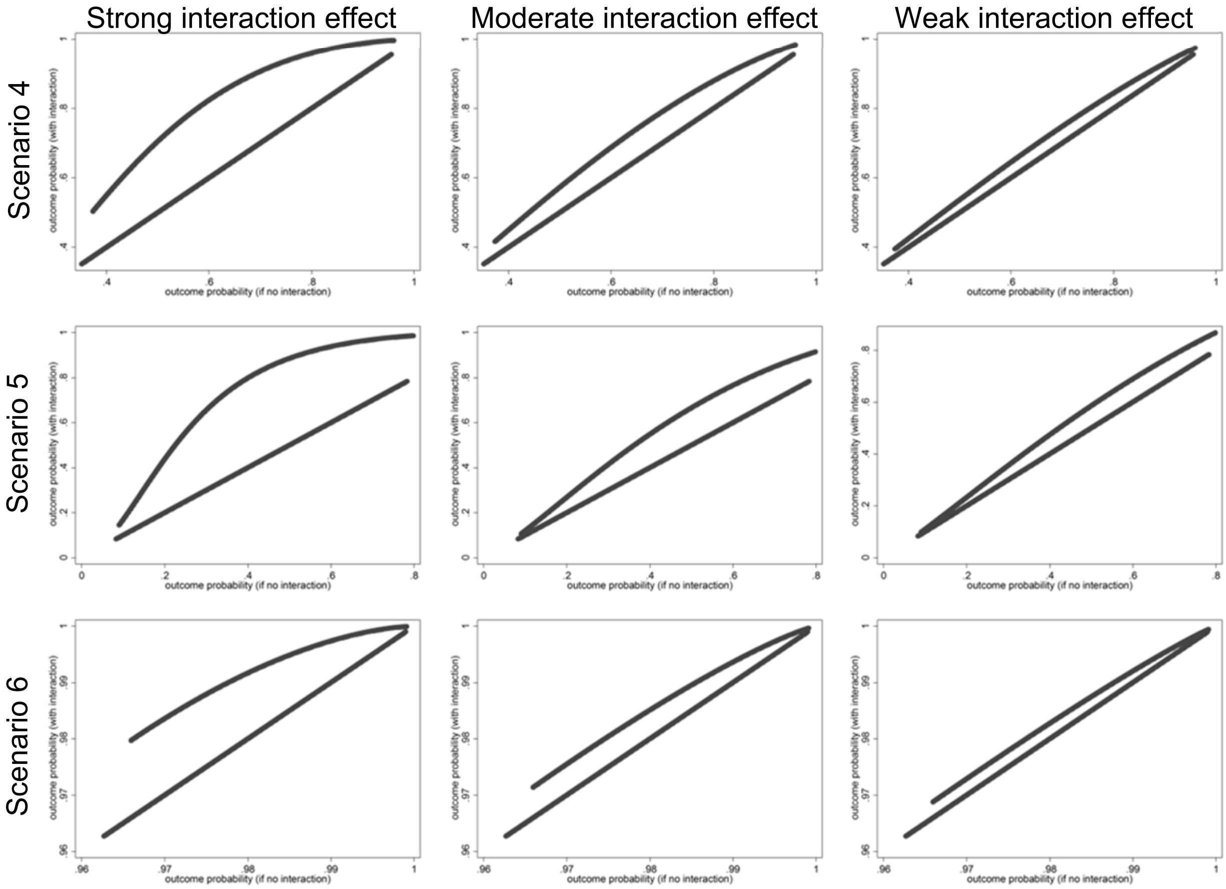


Figure E3.2: Probability of event with and without inclusion of an additional categorical predictor and interaction term for different strength interactions in scenarios 4 to 6.

Appendix E4: Histograms for distribution of performance statistics

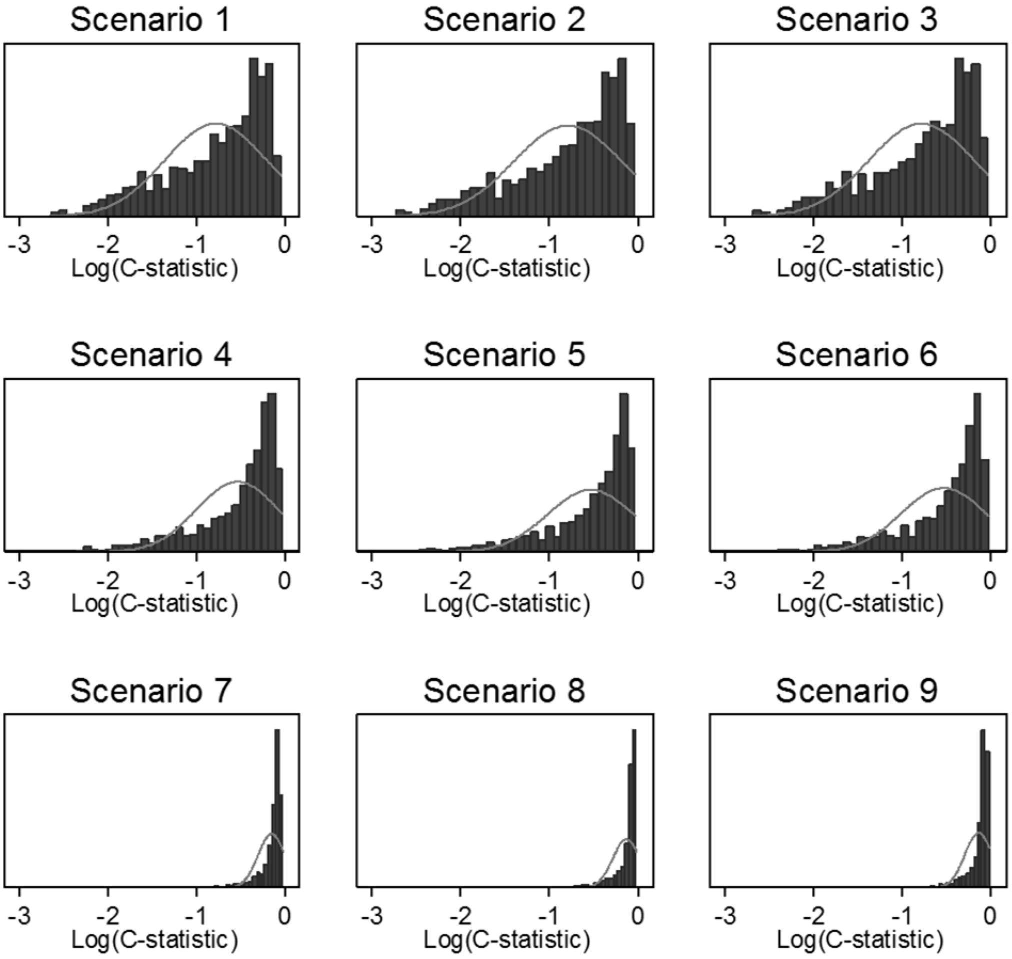


Figure E4.1: Histograms for natural log of C-statistic in all scenarios when heterogeneity in *β* was large (setting 7: σ_β=0.07).

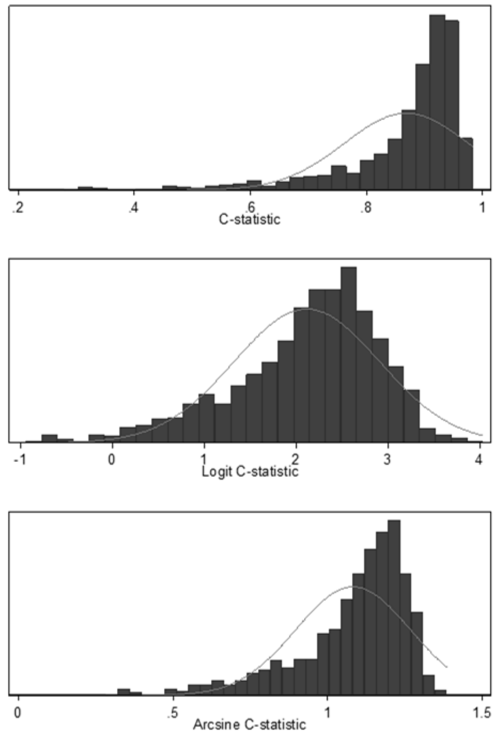


Figure E4.2: Histograms for C-statistic on original scale and using logit and arcsine transformations for scenario 7 when heterogeneity in *β* was large (setting 7: σ_β=0.07).

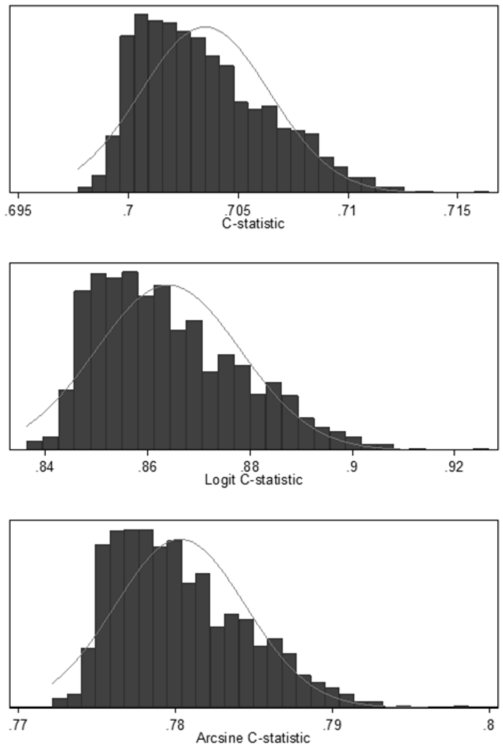


Figure E4.3: Histograms for C-statistic on original scale and using logit and arcsine transformations for scenario 4 when heterogeneity in *α* was large (setting 4: σ_α=1.0).

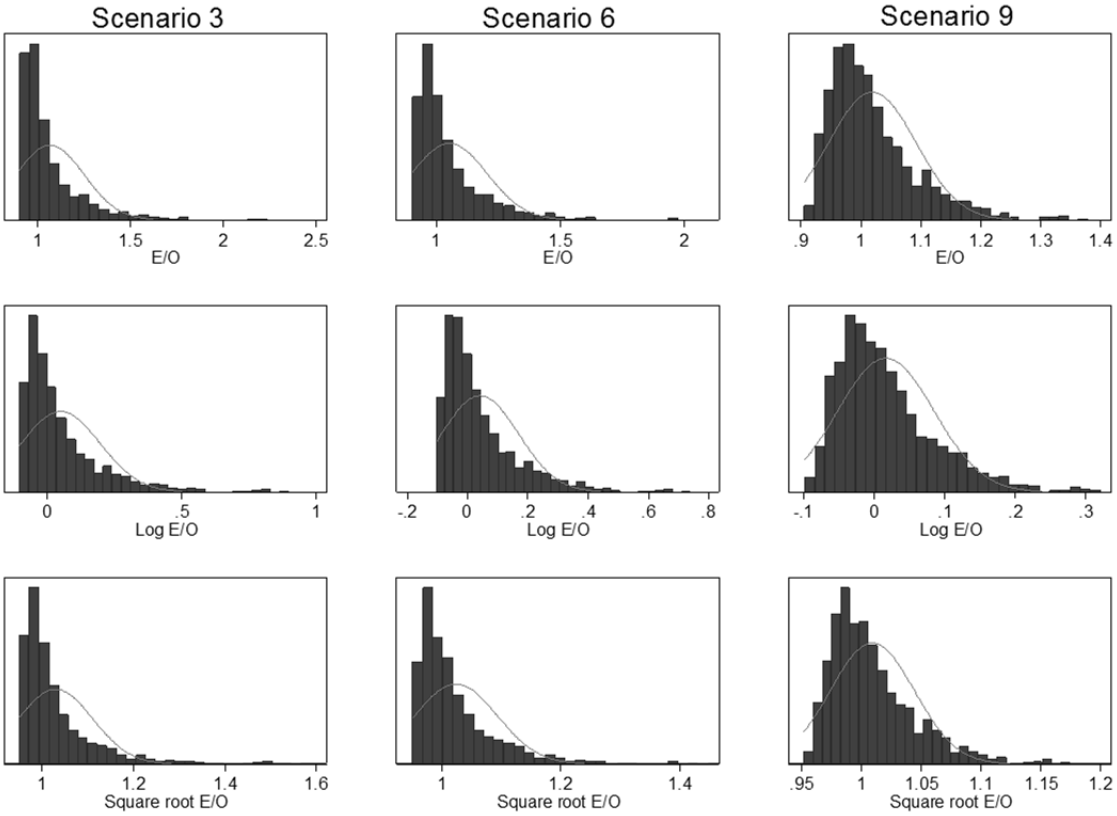


Figure E4.4: Histograms for E/O on original scale and using natural log and square root transformations for scenarios 3, 6 and 9 when heterogeneity in *α* was large (setting 4: σ_α=1.0).

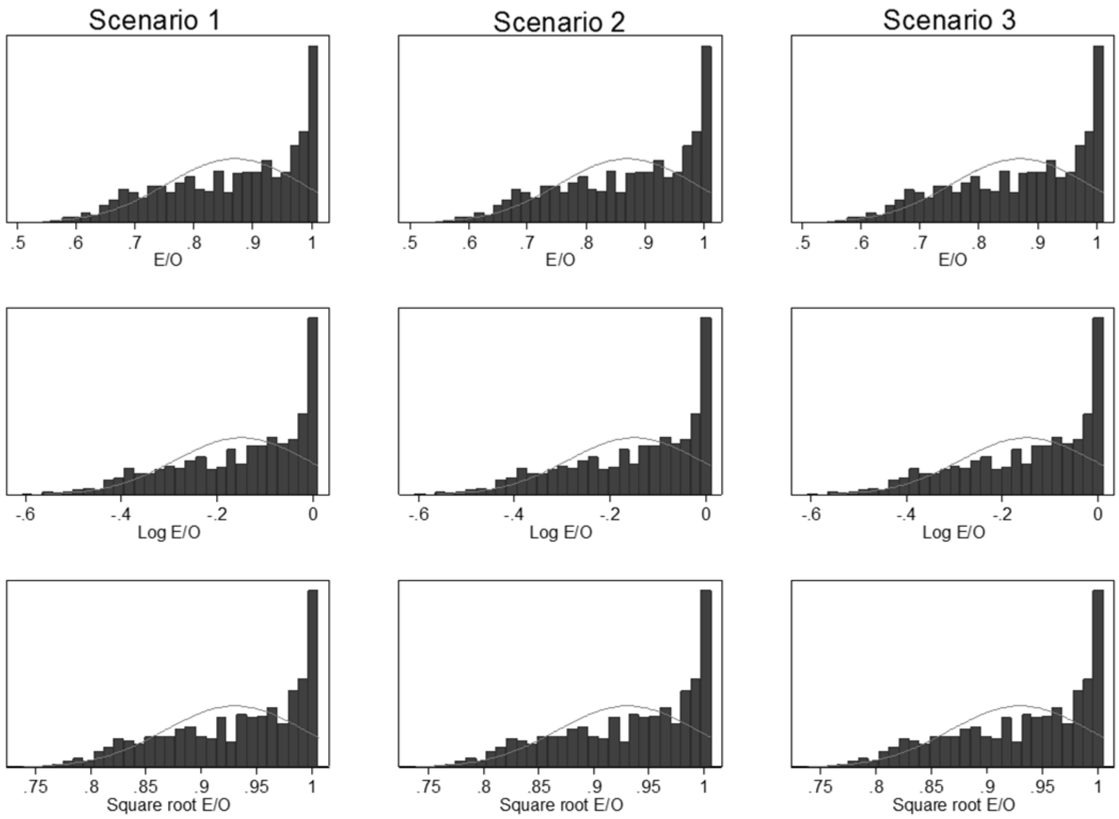


Figure E4.5: Histograms for E/O on original scale and using log and square root transformations for scenarios 1, 2 and 3 (weak predictor) when heterogeneity in β was large (setting 7: σ_β=0.07).

LIST OF REFERENCES

1. 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. British Medical Journal. 2013;346.

2. Riley RD, Hayden JA, Steyerberg EW, Moons KGM, Abrams K, Kyzas PA, et al. Prognosis Research Strategy (PROGRESS) 2: Prognostic Factor Research. Plos Medicine. 2013;10(2).

3. Steyerberg EW, Moons KGM, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. Plos Medicine. 2013;10(2).

4. Hingorani AD, van der Windt DA, Riley RD, Abrams K, Moons KGM, Steyerberg EW, et al. Prognosis research strategy (PROGRESS) 4: Stratified medicine research. British Medical Journal. 2013;346.

5. Adams ST, Leveson SH. Clinical prediction rules. British Medical Journal. 2012;344.

6. Wells PS, Anderson DR, Rodger M, Ginsberg JS, Kearon C, Gent M, et al. Derivation of a simple clinical model to categorize patients probability of pulmonary embolism: increasing the models utility with the SimpliRED D-dimer. Thrombosis and haemostasis. 2000;83(3):416-20.

7. Teramukai S, Ochiai K, Tada H, Fukushima M, Japan Multinational Trial Organization OC. PIEPOC: a new prognostic index for advanced epithelial ovarian cancer--Japan Multinational Trial Organization OC01-01. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2007;25(22):3302-6.

8. Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? British Medical Journal. 2009;338.

9. Sauerbrei W, Hollaender N, Riley RD, Altman DG. Evidence-based assessment and application of prognostic markers: The long way from single studies to meta-analysis. Communications in Statistics-Theory and Methods. 2006;35(7):1333-42.

10. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in Medicine. 1996;15(4):361-87.

11. Altman DG. Prognostic Models: A Methodological Framework and Review of Models for Breast Cancer. Cancer Investigation. 2009;27(3):235-43.

12. Steyerberg EW. Clinical prediction models: A practical approach to development, validation, and updating. New York, NY: Springer; 2010.

13. QResearch [28/03/15]. Available from: www.qresearch.org.

14. Haybittle JL, Blamey RW, Elston CW, Johnson J, Doyle PJ, Campbell FC, et al. A prognostic index in primary breast cancer. Br J Cancer. 1982;45(3):361-6.

15. Todd JH, Dowle C, Williams MR, Elston CW, Ellis IO, Hinton CP, et al. Confirmation of a prognostic index in primary breast cancer. Br J Cancer. 1987;56(4):489-92.

16. Kollias J, Murphy CA, Elston CW, Ellis IO, Robertson JF, Blamey RW. The prognosis of small primary breast cancers. European journal of cancer. 1999;35(6):908-12.

17. Hearne BJ, Teare MD, Butt M, Donaldson L. Comparison of Nottingham Prognostic Index and Adjuvant Online prognostic tools in young women with breast cancer: review of a single-institution experience. BMJ open. 2015;5(1):e005576.

18. Hemingway H, Riley RD, Altman DG. Ten steps towards improving prognosis research. British Medical Journal. 2009;339.

19. Royston P, Moons KGM, Altman DG, Vergouwe Y. Prognosis and prognostic research: Developing a prognostic model. British Medical Journal. 2009;338.

20. Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. British Medical Journal. 2009;338.

21. Moons KGM, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. British Medical Journal. 2009;338.

22. Harrell FE. Regression Modeling Strategies with Applications to Linear Models, Logistic Regression, and Survival Analysis. New York: Springer; 2001.

23. Battle C, Hutchings H, Lovett S, Bouamra O, Jones S, Sen A, et al. Predicting outcomes after blunt chest wall trauma: development and external validation of a new prognostic model. Critical care. 2014;18(3):R98.

24. Collett D. Modelling survival data in medical research. London: Chapman and Hall; 1994.

25. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association. 1958;53(282):457-81.

26. Cox DR. Regression Models and Life-Tables. Journal of the Royal Statistical Society Series B-Statistical Methodology. 1972;34(2):187-220.

27. StataCorp. Stata Survival Analysis and Epidemiological Tables Reference Manual: Release 13. College Station, Texas: Stata Press; 2013 2013.

28. Schoenfeld D. Partial residuals for the proportional hazards regression model. Biometrika. 1982;69(1):239-41.

29. Grambsch PM, Therneau TM. Proportional Hazards Tests and Diagnostics Based on Weighted Residuals. Biometrika. 1994;81(3):515-26.

30. Etezadiamoli J, Ciampi A. Extended hazard regression for censored survival data with covariates: A spline approximation for the baseline hazard function. Biometrics. 1987;43(1):181-92.

31. Rosenberg PS. Hazard function estimation using B-splines. Biometrics. 1995;51(3):874-87.

32. Royston P. Flexible parametric alternatives to the Cox model, and more. Stata Journal. 2001;1(1):1-28.

33. Lambert PC, Royston P. Further development of flexible parametric models for survival analysis. Stata Journal. 2009;9(2):265-90.

34. Royston P, Lambert PC. Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model. College Station, Texas: Stata Press; 2011 8/15/2011.

35. Royston P, Sauerbrei W. Multivariable modeling with cubic regression splines: A principled approach. Stata Journal. 2007;7(1):45-70.

36. Rutherford MJ, Crowther MJ, Lambert PC. The use of restricted cubic splines to approximate complex hazard functions in the analysis of time-to-event data: a simulation study. J Stat Comput Sim. 2015;85(4):777-93.

37. Akaike H. A new look at the statistical model identification. Ieee Transactions on Automatic Control. 1974;19(6):716-23.

38. Schwarz G. Estimating the dimension of a model. Annals of Statistics. 1978;6(2):461-4.

39. StataCorp. Stata Base Reference Manual: Release 12. College Station, Texas: Stata Press; 2011 2011.

40. Royston P, Parmar MKB. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Statistics in Medicine. 2002;21(15):2175-97.

41. Baade PD, Royston P, Youl PH, Weinstock MA, Geller A, Aitken JF. Prognostic survival model for people diagnosed with invasive cutaneous melanoma. BMC cancer. 2015;15(1):27.

42. Collins GS, Reitsma JB, Altman DG, Moons KG, for the members of the Tg. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): The TRIPOD Statement. European urology. 2014.

43. Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis .2. Accuracy and precision of regression estimates. Journal of Clinical Epidemiology. 1995;48(12):1503-10.

44. 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.

45. Rubin DB. Multiple Imputation for Nonresponse in Surveys. New York: Wiley; 1987 1987.

46. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Statistics in Medicine. 2011;30(4):377-99.

47. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Statistics in Medicine. 2006;25(1):127-41.

48. Altman DG, Lausen B, Sauerbrei W, Schumacher M. Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. J Natl Cancer Inst. 1994;86(11):829-35.

49. Royston P, Sauerbrei W. Building multivariable regression models with continuous covariates in clinical epidemiology - With an emphasis on fractional polynomials. Methods of Information in Medicine. 2005;44(4):561-71.

50. Royston P, Sauerbrei W. Multivariable model-building - A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables. Chichester: John Wiley; 2008 2008.

51. Sauerbrei W, Royston P. Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. Journal of the Royal Statistical Society Series A-Statistics in Society. 1999;162:71-94.

52. Sauerbrei W, Royston P, Binder H. Selection of important variables and determination of functional form for continuous predictors in multivariable model building. Stat Med. 2007;26(30):5512-28.

53. Mantel N. Why stepdown procedures in variable selection. Technometrics. 1970;12(3):621-&.

54. 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 Med Res Methodol. 2014;14:40.

55. Royston P, Altman DG. External validation of a Cox prognostic model: principles and methods. BMC Med Res Methodol. 2013;13:33.

56. Altman DG, Royston P. What do we mean by validating a prognostic model? Statistics in Medicine. 2000;19(4):453-73.

57. Steyerberg EW, Harrell FE, Borsboom GJJM, Eijkemans MJC, Vergouwe Y, Habbema JDF. Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. Journal of Clinical Epidemiology. 2001;54(8):774-81.

58. Moons KG, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart. 2012;98(9):683-90.

59. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015;162(1):W1-73.

60. Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol. 2015;68(3):279-89.

61. Pennells L, Kaptoge S, White IR, Thompson SG, Wood AM. Assessing Risk Prediction Models Using Individual Participant Data From Multiple Studies. American Journal of Epidemiology. 2014;179(5):621-32.

62. van Klaveren D, Steyerberg EW, Perel P, Vergouwe Y. Assessing discriminative ability of risk models in clustered data. Bmc Medical Research Methodology. 2014;14.

63. Schuetz P, Koller M, Christ-Crain M, Steyerberg E, Stolz D, Muller C, et al. Predicting mortality with pneumonia severity scores: importance of model recalibration to local settings. Epidemiology and infection. 2008;136(12):1628-37.

64. 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. 2010;21(1):128-38.

65. Vergouwe Y, Moons KGM, Steyerberg EW. External Validity of Risk Models: Use of Benchmark Values to Disentangle a Case-Mix Effect From Incorrect Coefficients. American Journal of Epidemiology. 2010;172(8):971-80.

66. Steyerberg EW, Borsboom GJJM, van Houwelingen HC, Eijkemans MJC, Habbema JDF. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. Statistics in Medicine. 2004;23(16):2567-86.

67. Pencina MJ, D'Agostino RB. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. Stat Med. 2004;23(13):2109-23.

68. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat Med. 2011;30(10):1105-17.

69. Royston P, Parmar MKB, Sylvester R. Construction and validation of a prognostic model across several studies, with an application in superficial bladder cancer. Statistics in Medicine. 2004;23(6):907-26.

70. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. Statistics in Medicine. 2004;23(5):723-48.

71. Pencina MJ, D'Agostino RB, Sr., D'Agostino RB, Jr., Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med. 2008;27(2):157-72; discussion 207-12.

72. Pencina MJ, D'Agostino RB, Sr., Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. Stat Med. 2011;30(1):11-21.

73. Steyerberg EW, Pencina MJ. Reclassification calculations for persons with incomplete follow-up. Ann Intern Med. 2010;152(3):195-6; author reply 6-7.

74. Leening MJ, Vedder MM, Witteman JC, Pencina MJ, Steyerberg EW. Net reclassification improvement: computation, interpretation, and controversies: a literature review and clinician's guide. Ann Intern Med. 2014;160(2):122-31.

75. Hilden J, Gerds TA. A note on the evaluation of novel biomarkers: do not rely on integrated discrimination improvement and net reclassification index. Stat Med. 2014;33(19):3405-14.

76. Kerr KF, Wang Z, Janes H, McClelland RL, Psaty BM, Pepe MS. Net reclassification indices for evaluating risk prediction instruments: a critical review. Epidemiology. 2014;25(1):114-21.

77. Pepe MS, Janes H, Li CI. Net risk reclassification p values: valid or misleading? J Natl Cancer Inst. 2014;106(4):dju041.

78. Vickers AJ, Pepe M. Does the net reclassification improvement help us evaluate models and markers? Ann Intern Med. 2014;160(2):136-7.

79. Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. Ann Intern Med. 2006;144(3):201-9.

80. Grobman WA, Stamilio DM. Methods of clinical prediction. American journal of obstetrics and gynecology. 2006;194(3):888-94.

81. Hemingway H. Prognosis research: Why is Dr. Lydgate still waiting? Journal of Clinical Epidemiology. 2006;59(12):1229-38.

82. Peat G, Riley RD, Croft P, Morley KI, Kyzas PA, Moons KG, et al. Improving the transparency of prognosis research: the role of reporting, data sharing, registration, and protocols. PLoS Med. 2014;11(7):e1001671.

83. 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.

84. Mallett S, Royston P, Waters R, Dutton S, Altman DG. Reporting performance of prognostic models in cancer: a review. Bmc Medicine. 2010;8.

85. Bouwmeester W, Zuithoff NPA, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, et al. Reporting and Methods in Clinical Prediction Research: A Systematic Review. Plos Medicine. 2012;9(5).

86. McMinn DJW, Snell KIE, Daniel J, Treacy RBC, Pynsent PB, Riley RD. Mortality and implant revision rates of hip arthroplasty in patients with osteoarthritis: registry based cohort study. British Medical Journal. 2012;344.

87. Hip replacements - Alternatives. NHS Choices [Internet]. 2012 5/9/2012. Available from: www.nhs.uk/Conditions/Hip-replacement/Pages/Alternatives.aspx.

88. Hip replacement - Why it is done. NHS Choices [Internet]. 2012 5/9/2012. Available from: www.nhs.uk/Conditions/Hip-replacement/Pages/Why-it-should-be-done.aspx.

89. National Joint Registry of England and Wales. 8th Annual Report 2011. 2011 9/2011. Report No.

90. Total Hip Replacement. AAOS OrthoInfo [Internet]. 2011 5/16/2012. Available from: http://orthoinfo.aaos.org/topic.cfm?topic=A00377.

91. Dheerendra S, Khan W, Saeed MZ, Goddard N. Recent developments in total hip replacements: cementation, articulation, minimal-invasion and navigation. J PerioperPract. 2010;20(4):133-8.

92. McCaskie AW, Barnes MR, Lin E, Harper WM, Gregg PJ. Cement pressurisation during hip replacement. J Bone Joint SurgBr. 1997;79(3):379-84.

93. Hip implants. AAOS OrthoInfo [Internet]. 2007 5/18/2012. Available from: http://orthoinfo.aaos.org/topic.cfm?topic=A00355.

94. Kim Y-H, Oh S-H, Kim J-S. Primary total hip arthroplasty with a second-generation cementless total hip prosthesis in patients younger than fifty years of age. Journal of Bone and Joint Surgery. 2003;85(1):109-14.

95. Treacy RB, McBryde CW, Pynsent PB. Birmingham hip resurfacing arthroplasty. A minimum follow-up of five years. J Bone Joint SurgBr. 2005;87(2):167-70.

96. Coulter G, Young DA, Dalziel RE, Shimmin AJ. Birmingham hip resurfacing at a mean of ten years: results from an independent centre. J Bone Joint SurgBr. 2012;94(3):315-21.

97. Daniel J, Pynsent PB, McMinn DJW. Metal-on-metal resurfacing of the hip in patients under the age of 55 years with osteoarthritis. Journal of Bone and Joint Surgery-British Volume. 2004;86B(2):177-84.

98. Schemper M, Smith TL. A note on quantifying follow-up in studies of failure time. Control Clin Trials. 1996;17(4):343-6.

99. Kalbfleisch JD, Prentice RL. Likelihood Construction and Further Results. The Statistical Analysis of Failure Time Data. 2nd Edition ed. New Jersey: Wiley; 2002.

100. Hess KR. Graphical methods for assessing violations of the proportional hazards assumption in Cox regression. Statistics in Medicine. 1995;14(15):1707-23.

101. National Joint Registry of England and Wales. 7th Annual Report 2010. 2010 9/2010. Report No.

102. National Joint Registry of England and Wales. 9th Annual Report 2012. 2012 9/2012. Report No.

103. Kendal AR, Prieto-Alhambra D, Arden NK, Carr A, Judge A. Mortality rates at 10 years after metal-on-metal hip resurfacing compared with total hip replacement in England: retrospective cohort analysis of hospital episode statistics. Bmj-British Medical Journal. 2013;347.

104. Abdulkarim A, Ellanti P, Motterlini N, Fahey T, O'Byrne JM. Cemented versus Uncemented fixation in Total Hip Replacement: A Systematic Review & Meta-Analysis. Irish Journal of Medical Science. 2013;182:S356-S7.

105. Kleinbaum DG, Klein M. The Cox Proportional Hazards Model and its Characteristics. Survival Analysis: A Self-Learning Text. 2nd ed. New York, NY: Springer; 2005. p. 83-129.

106. Lunn M, Mcneil N. Applying Cox Regression to Competing Risks. Biometrics. 1995;51(2):524-32.

107. Rosenbaum PR, Rubin DB. The Central Role of the Propensity Score in Observational Studies for Causal Effects. Biometrika. 1983;70(1):41-55.

108. Dehejia RH, Wahba S. Propensity score-matching methods for nonexperimental causal studies. Review of Economics and Statistics. 2002;84(1):151-61.

109. Winkelmayer WC, Kurth T. Propensity scores: help or hype? Nephrology Dialysis Transplantation. 2004;19(7):1671-3.

110. Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. American Journal of Epidemiology. 2003;158(3):280-7.

111. Jackson C. Flexsurv: Flexible parametric survival models. R package. Cran R Project [Internet]. 2013 10/28/2013. Available from: http://cran.r-project.org/web/packages/flexsurv/flexsurv.pdf.

112. Bredel M, Scholtens DM, Harsh GR, Bredel C, Chandler JP, Renfrow JJ, et al. A Network Model of a Cooperative Genetic Landscape in Brain Tumors. Jama-Journal of the American Medical Association. 2009;302(3):261-75.

113. Chen HY, Yu SL, Chen CH, Chang GC, Chen CY, Yuan A, et al. A five-gene signature and clinical outcome in non-small-cell lung cancer. New England Journal of Medicine. 2007;356(1):11-20.

114. Crijns APG, Fehrmann RSN, de Jong S, Gerbens F, Meersma GJ, Klip HG, et al. Survival-Related Profile, Pathways, and Transcription Factors in Ovarian Cancer. Plos Medicine. 2009;6(2):181-93.

115. Parimon T, Au DH, Martin PJ, Chien JW. A risk score for mortality after allogeneic hematopoietic cell transplantation. Annals of Internal Medicine. 2006;144(6):407-14.

116. Patel AA, Chen MH, Renshaw AA, D'Amico AV. PSA failure following definitive treatment of prostate cancer having biopsy Gleason score 7 with tertiary grade 5. Jama-Journal of the American Medical Association. 2007;298(13):1533-8.

117. Stebbing J, Sanitt A, Nelson M, Powles T, Gazzard B, Bower M. A prognostic index for AIDS-associated Kaposi's sarcoma in the era of highly active antiretroviral therapy. Lancet. 2006;367(9521):1495-502.

118. Tice JA, Cummings SR, Smith-Bindman R, Ichikawa L, Barlow WE, Kerlikowske K. Using clinical factors and mammographic breast density to estimate breast cancer risk: Development and validation of a new predictive model. Annals of Internal Medicine. 2008;148(5):337-W75.

119. Hippisley-Cox J, Coupland C, Robson J, Sheikh A, Brindle P. Predicting risk of type 2 diabetes in England and Wales: prospective derivation and validation of QDScore. British Medical Journal. 2009;338.

120. Sattar N, McConnachie A, Shaper AG, Blauw GJ, Buckley BM, de Craen AJ, et al. Can metabolic syndrome usefully predict cardiovascular disease and diabetes? Outcome data from two prospective studies. Lancet. 2008;371(9628):1927-35.

121. Moylan CA, Brady CW, Johnson JL, Smith AD, Tuttle-Newhall JE, Muir AJ. Disparities in Liver Transplantation Before and After Introduction of the MELD Score. Jama-Journal of the American Medical Association. 2008;300(20):2371-8.

122. Liu PY, Swerdloff RS, Christenson PD, Handelsman DJ, Wang C. Rate, extent, and modifiers of spermatogenic recovery after hormonal male contraception: an integrated analysis. Lancet. 2006;367(9520):1412-20.

123. Hippisley-Cox J, Coupland C. Predicting risk of osteoporotic fracture in men and women in England and Wales: prospective derivation and validation of QFractureScores. British Medical Journal. 2009;339.

124. Daly CA, De Stavola B, Fox KM. Predicting prognosis in stable angina - results from the Euro heart survey of stable angina: prospective observational study. British Medical Journal. 2006;332(7536):262-5.
125. Fox KAA, Dabbous OH, Goldberg RJ, Pieper KS, Eagle KA, Van de Werf F, et al. Prediction of risk of death and myocardial infarction in the six months after presentation with acute coronary syndrome: prospective multinational observational study (GRACE). British Medical Journal. 2006;333(7578):1091-4.
126. Montalvo G, Avanzini F, Anselmi M, Prandi R, Ibarra S, Marquez M, et al. Diagnostic evaluation of people with hypertension in low income country: cohort study of "essential" method of risk stratification. British Medical Journal. 2008;337(7672).
127. Buckley BS, Simpson CR, McLernon DJ, Murphy AW, Hannaford PC. Five year prognosis in patients with angina identified in primary care: incident cohort study. British Medical Journal. 2009;339.
128. Rassi A, Rassi A, Little WC, Xavier SS, Rassi SG, Rassi AG, et al. Development and validation of a risk score for predicting death in Chagas' heart disease. New England Journal of Medicine. 2006;355(8):799-808.
129. de Ruijter W, Westendorp RGJ, Assendelft WJJ, den Elzen WPJ, de Craen AJM, le Cessie S, et al. Use of Framingham risk score and new biomarkers to predict cardiovascular mortality in older people: population based observational cohort study. British Medical Journal. 2009;338.
130. Denes P, Larson JC, Lloyd-Jones DM, Prineas RJ, Greenland P. Major and minor ECG abnormalities in asymptomatic women and risk of cardiovascular events and mortality. Jama-Journal of the American Medical Association. 2007;297(9):978-85.
131. Gaziano TA, Young CR, Fitzmaurice G, Atwood S, Gaziano JM. Laboratory-based versus non-laboratory-based method for assessment of cardiovascular disease risk: the NHANES I Follow-up Study cohort. Lancet. 2008;371(9616):923-31.
132. Melander O, Newton-Cheh C, Almgren P, Hedblad B, Berglund G, Engstrom G, et al. Novel and Conventional Biomarkers for Prediction of Incident Cardiovascular Events in the Community. Jama-Journal of the American Medical Association. 2009;302(1):49-57.
133. Zethelius B, Berglund L, Sundstrom J, Ingelsson E, Basu S, Larsson A, et al. Use of multiple biomarkers to improve the prediction of death from cardiovascular causes. New England Journal of Medicine. 2008;358(20):2107-16.
134. Cook NR, Buring JE, Ridker PM. The effect of including C-reactive protein in cardiovascular risk prediction models for women. Annals of Internal Medicine. 2006;145(1):21-9.
135. Ingelsson E, Schaefer EJ, Contois JH, McNamara JR, Sullivan L, Keyes MJ, et al. Clinical utility of different lipid measures for prediction of coronary heart disease in men and women. Jama-Journal of the American Medical Association. 2007;298(7):776-85.
136. Ridker PM, Buring JE, Rifai N, Cook NR. Development and validation of improved algorithms for the assessment of global cardiovascular risk in women - The Reynolds Risk Score. Jama-Journal of the American Medical Association. 2007;297(6):611-9.
137. Paynter NP, Chasman DI, Buring JE, Shiffman D, Cook NR, Ridker PM. Cardiovascular Disease Risk Prediction With and Without Knowledge of Genetic Variation at Chromosome 9p21.3. Annals of Internal Medicine. 2009;150(2):65-U16.
138. 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. British Medical Journal. 2007;335(7611):136-41.
139. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Minhas R, Sheikh A, et al. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. British Medical Journal. 2008;336(7659):1475-+.
140. Schnabel RB, Sullivan LM, Levy D, Pencina M, Massaro JM, D'Agostino RB, et al. Development of a risk score for atrial fibrillation (Framingham Heart Study): a community-based cohort study. Lancet. 2009;373(9665):739-45.

141. Sekhri N, Feder GS, Junghans C, Eldridge S, Umaipalan A, Madhu R, et al. Incremental prognostic value of the exercise electrocardiogram in the initial assessment of patients with suspected angina: cohort study. British Medical Journal. 2008;337.

142. Lauer MS, Pothier CE, Magid DJ, Smith SS, Kaftan MW. An externally validated model for predicting long-term survival after exercise treadmill testing in patients with suspected coronary artery disease and a normal electrocardiogram. Annals of Internal Medicine. 2007;147(12):821-8.

143. Kalbfleisch JD, Prentice RL. Relative Risk (Cox) Regression Models. The Statistical Analysis of Failure Time Data. 2nd Edition ed. New Jersey: Wiley; 2002.

144. Collins GS, Mallett S, Omar O, Yu LM. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. BMC Med. 2011;9:103.

145. 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. J Clin Epidemiol. 2013;66(3):268-77.

146. Copas JB, Long TY. Estimating the Residual Variance in Orthogonal Regression with Variable Selection. Statistician. 1991;40(1):51-9.

147. Derksen S, Keselman HJ. Backward, Forward and Stepwise Automated Subset-Selection Algorithms - Frequency of Obtaining Authentic and Noise Variables. British Journal of Mathematical & Statistical Psychology. 1992;45:265-82.

148. Concato J, Peduzzi P, Holford TR, Feinstein AR. Importance of events per independent variable in proportional hazards analysis .1. Background, goals, and general strategy. Journal of Clinical Epidemiology. 1995;48(12):1495-501.

149. Debray TPA, Moons KGM, Ahmed I, Koffijberg H, Riley RD. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. Statistics in Medicine. 2013;32(18):3158-80.

150. Ingui BJ, Rogers MA. Searching for clinical prediction rules in MEDLINE. Journal of the American Medical Informatics Association : JAMIA. 2001;8(4):391-7.

151. 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):1251-61.

152. CancerStats Key Facts - Pancreatic Cancer. Cancer Research UK [Internet]. 2013 2/12/2013. Available from: http://publications.cancerresearchuk.org/downloads/Product/CS_KF_PANCREAS.pdf.

153. Shuster JJ. Median follow-up in clinical trials. J ClinOncol. 1991;9(1):191-2.

154. Bramhall SR, Rosemurgy A, Brown PD, Bowry C, Buckels JAC. Marimastat as first-line therapy for patients with unresectable pancreatic cancer: A randomized trial. Journal of Clinical Oncology. 2001;19(15):3447-55.

155. Bramhall SR, Schulz J, Nemunaitis J, Brown PD, Baillet M, Buckels JAC. A double-blind placebo-controlled, randomised study comparing gemcitabine and marimastat with gemcitabine and placebo as first line therapy in patients with advanced pancreatic cancer. British Journal of Cancer. 2002;87(2):161-7.

156. Stocken DD, Hassan AB, Altman DG, Billingham LJ, Bramhall SR, Johnson PJ, et al. Modelling prognostic factors in advanced pancreatic cancer. British Journal of Cancer. 2008;99(6):883-93.

157. Stocken DD. Statistical modelling for the prognostic classification of patients with pancreatic cancer for optimisation of treatment allocation 2010.

158. White IR, Royston P. Imputing missing covariate values for the Cox model. Statistics in Medicine. 2009;28(15):1982-98.

159. Graham JW, Olchowski AE, Gilreath TD. How many imputations are really needed? - Some practical clarifications of multiple imputation theory. Prevention Science. 2007;8(3):206-13.

160. 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. Stat Med. 2015;34(11):1841-63.

161. Lachin JM, Matts JP, Wei LJ. Randomization in Clinical-Trials - Conclusions and Recommendations. Controlled Clinical Trials. 1988;9(4):365-74.

162. Scott NW, McPherson GC, Ramsay CR, Campbell MK. The method of minimization for allocation to clinical trials: a review. Controlled Clinical Trials. 2002;23(6):662-74.

163. Kahan BC, Morris TP. Improper analysis of trials randomised using stratified blocks or minimisation. Statistics in Medicine. 2012;31(4):328-40.

164. Raab GM, Day S, Sales J. How to select covariates to include in the analysis of a clinical trial. Controlled Clinical Trials. 2000;21(4):330-42.

165. Sauerbrei W. The use of resampling methods to simplify regression models in medical statistics. Journal of the Royal Statistical Society Series C-Applied Statistics. 1999;48:313-29.

166. Vergouwe Y, Royston P, Moons KGM, Altman DG. Development and validation of a prediction model with missing predictor data: a practical approach. Journal of Clinical Epidemiology. 2010;63(2):205-14.

167. Wood AM, White IR, Royston P. How should variable selection be performed with multiply imputed data? Statistics in Medicine. 2008;27(17):3227-46.

168. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the Yield of Medical Tests. Jama-Journal of the American Medical Association. 1982;247(18):2543-6.

169. Newson RB. Comparing the predictive powers of survival models using Harrell's C or Somers' D. Stata Journal. 2010;10(3):339-58.

170. Chemotherapy for pancreatic Cancer. Pancreatic Cancer UK [Internet]. 2013 8/7/2013. Available from: http://www.pancreaticcancer.org.uk/information-and-support/treatment/chemotherapy.

171. Excellence NIfHaC. Guidance on the use of gemcitabine for the treatment of pancreatic cancer. London: 2001 5/2001. Report No.: TA25.

172. Copas JB. Regression, Prediction and Shrinkage. Journal of the Royal Statistical Society Series B-Methodological. 1983;45(3):311-54.

173. Jackson D, Riley R, White IR. Multivariate meta-analysis: Potential and promise. Statistics in Medicine. 2011;30(20):2481-98.

174. Riley RD, Abrams KR, Lambert PC, Sutton AJ, Thompson JR. An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. Statistics in Medicine. 2007;26(1):78-97.

175. Riley RD. Multivariate meta-analysis: the effect of ignoring within-study correlation. Journal of the Royal Statistical Society Series A-Statistics in Society. 2009;172:789-811.

176. Hua H. Survival modelling in mathematical and medical statistics: University of Birmingham; 2015.

177. Glass GV. Primary, Secondary, and Meta-Analysis of Research. American Educational Research Association. 1976;5:3-8.

178. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Fixed-Effect Model. Inroduction to Meta-Analysis. Chichester: John Wiley & Sons; 2009. p. 63-7.

179. Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. Methods for Meta-Analysis in Medical Research. Chichester: John Wiley; 2000 2000.

180. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Statistics in Medicine. 2002;21(11):1539-58.

181. Deeks JJ, Higgins JPT, Altman DG, (editors). Chapter 9: Analysing data and undertaking meta-analyses. In: Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions Version 510 (updated March 2011). The Cochrane Collaboration2011. p. Available from www.cochrane-handbook.org.

182. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. British Medical Journal. 2003;327(7414):557-60.

183. Rucker G, Schwarzer G, Carpenter JR, Schumacher M. Undue reliance on I-2 in assessing heterogeneity may mislead. Bmc Medical Research Methodology. 2008;8.
184. Riley RD, Higgins JPT, Deeks JJ. Interpretation of random effects meta-analyses. British Medical Journal. 2011;342:964-7.
185. Dersimonian R, Laird N. Metaanalysis in Clinical-Trials. Controlled Clinical Trials. 1986;7(3):177-88.
186. White IR. Multivariate random-effects meta-analysis. Stata Journal. 2009;9(1):40-56.
187. Higgins JPT, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. Journal of the Royal Statistical Society Series A-Statistics in Society. 2009;172:137-59.
188. Riley RD, Price MJ, Jackson D, Wardle M, Gueyffier F, Wang J, et al. Multivariate meta-analysis using individual participant data. J Research Synthesis Methods. 2014.
189. Riley RD, Thompson JR, Abrams KR. An alternative model for bivariate random-effects meta-analysis when the within-study correlations are unknown. Biostatistics. 2008;9(1):172-86.
190. Reitsma JB, Glas AS, Rutjes AWS, Scholten RJPM, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. Journal of Clinical Epidemiology. 2005;58(10):982-90.
191. Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JAC. A unification of models for meta-analysis of diagnostic accuracy studies. Biostatistics. 2007;8(2):239-51.
192. Alexandersson A. Graphing confidence ellipses: An update of ellip for Stata 8. Stata Journal. 2004;4(3):242-56.
193. Riley RD, Abrams KR, Sutton AJ, Lambert PC, Thompson JR. Bivariate random-effects meta-analysis and the estimation of between-study correlation. Bmc Medical Research Methodology. 2007;7.
194. 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. 2012;98(9):691-8.
195. Ahmed I, Debray TP, Moons KG, Riley RD. Developing and validating risk prediction models in an individual participant data meta-analysis. BMC Med Res Methodol. 2014;14:3.
196. Vach W. Calibration of clinical prediction rules does not just assess bias. J Clin Epidemiol. 2013;66(11):1296-301.
197. Lee KJ, Thompson SG. Flexible parametric models for random-effects distributions. Stat Med. 2008;27(3):418-34.
198. van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. Statistics in Medicine. 2002;21(4):589-624.
199. Thompson SG, Higgins JPT. How should meta-regression analyses be undertaken and interpreted? Statistics in Medicine. 2002;21(11):1559-73.
200. Oudega R, Moons KG, Hoes AW. Ruling out deep venous thrombosis in primary care. A simple diagnostic algorithm including D-dimer testing. Thrombosis and haemostasis. 2005;94(1):200-5.
201. Hosmer DW, Lemeshow, S. Assessing the Fit of the Model. Applied Logistic Regression. 2nd ed. New York: Wiley; 2000. p. 143-202.
202. Trikalinos TA, Trow P, Schmid CH. Simulation-Based Comparison of Methods for Meta-Analysis of Proportions and Rates. Rockville (MD)2013.
203. Austin PC, Steyerberg EW. Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. BMC Med Res Methodol. 2012;12:82.
204. Wyatt JC, Altman DG. Prognostic Models - Clinically Useful Or Quickly Forgotten - Commentary. British Medical Journal. 1995;311(7019):1539-41.
205. Liao L, Mark DB. Clinical prediction models: are we building better mousetraps? Journal of the American College of Cardiology. 2003;42(5):851-3.

206. Baker SG, Sargent DJ. Designing a randomized clinical trial to evaluate personalized medicine: a new approach based on risk prediction. J Natl Cancer Inst. 2010;102(23):1756-9.
207. Vickers AJ, Cronin AM. Traditional statistical methods for evaluating prediction models are uninformative as to clinical value: towards a decision analytic framework. Seminars in oncology. 2010;37(1):31-8.
208. von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Bmj. 2007;335(7624):806-8.
209. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Bmj. 2003;326(7379):41-4.
210. McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM, et al. REporting recommendations for tumour MARKer prognostic studies (REMARK). European journal of cancer. 2005;41(12):1690-6.
211. Bossuyt PM. STARD statement: still room for improvement in the reporting of diagnostic accuracy studies. Radiology. 2008;248(3):713-4.
212. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. European heart journal. 2014;35(29):1925-31.
213. Croft P, Altman DG, Deeks JJ, Dunn KM, Hay AD, Hemingway H, et al. The science of clinical practice: disease diagnosis or patient prognosis? Evidence about "what is likely to happen" should shape clinical practice. BMC Med. 2015;13:20.
214. Rizopoulos D, Lesaffre E. Introduction to the special issue on joint modelling techniques. Stat Methods Med Res. 2014;23(1):3-10.
215. Rizopoulos D, Takkenberg JJ. Tools & techniques--statistics: Dealing with time-varying covariates in survival analysis--joint models versus Cox models. EuroIntervention : journal of EuroPCR in collaboration with the Working Group on Interventional Cardiology of the European Society of Cardiology. 2014;10(2):285-8.