Statistical methods for prognostic factor and risk prediction research

Rebecca Louise Whittle

A thesis submitted for the degree of

Doctor of Philosophy

June 2023

Keele University

Table of Contents

Acknowledgments ix

Abstract xi

List of tables xiii

List of figures xvii

List of abbreviations xix

1 Introduction 1

1.1 Overview of the thesis 1

1.2 What is prognosis research? 3

1.2.1 Prognosis research framework 4

1.2.2 Sources of data 10

1.3 Statistical models used for prediction research 11

1.3.1 Linear regression 12

1.3.2 Logistic regression 12

1.4 Developing a prediction model 14

1.4.1 Selecting candidate predictors 15

1.4.2 Sample size 15

1.4.3 Missing data 16

1.4.4 Continuous predictors 18

1.4.5 Selection of final model 18

1.4.6 Model performance 20

1.4.7 Model validation 22

1.5 Why is the timing of measurements important? 25

1.6 Measurement error 27

1.6.1 What is measurement error? 27

1.6.2 Differential and non-differential measurement error 28

1.6.3 Additive and multiplicative measurement error 29

1.6.4 Dependent and independent measurement error 30

1.6.5 Impact of measurement error 30

1.6.6 Measurement error in prediction models 32

1.7 Individual participant data (IPD) meta-analysis for prognosis and prediction studies………… 34

1.7.1 Rationale for IPD meta-analysis 35

1.7.2 Statistical methods for an IPD meta-analysis 36
1.7.3 Issues with IPD meta-analysis 39
1.8 Aims and outline of the thesis 40
2 First-trimester ultrasound measurements and maternal serum biomarkers as prognostic factors in monochorionic twins 45
2.1 Introduction 45
2.2 Medical terms glossary 45
2.3 Clinical rationale for this chapter 47
2.4 Objectives 48
2.4.1 Clinical objectives 48
2.4.2 Methodological objectives 49
2.4.3 Chapter outline 49
2.5 Methods 50
2.5.1 Participants 50
2.5.2 Potential prognostic factors 51
2.5.3 Outcomes 54
2.5.4 Existing prognostic factors 55
2.5.5 Missing data 57
2.5.6 Statistical analysis 58
2.5.7 Sample size 60
2.5.8 Developing a prognostic model and examining instability 61
2.5.9 Sensitivity analysis: Firth's correction 62
2.6 Results 62
2.6.1 Summary of data 62
2.6.2 Primary outcome: Fetal composite adverse outcome 64
2.6.3 Illustration of the issues of developing a prognostic model with this dataset 67
2.6.4 Secondary outcomes 69
2.7 Discussion 74
2.7.1 Summary of clinical findings 74
2.7.2 Strengths and limitations 75
2.7.3 Clinical implications and future research 77
2.7.4 Developing a prognostic model 78
2.7.5 Conclusions 79
3 Measurement error and timing of predictor values used in prediction model research: a systematic review of current practice and reporting 81

3.1 Introduction 81

3.1.1 Chapter rationale 81

3.1.2 Chapter aims 83

3.1.3 Chapter outline 83

3.2 Methods 84

3.2.1 Search strategy 84

3.2.2 Inclusion/Exclusion criteria 87

3.2.3 Selection of articles 88

3.2.4 Extraction of information 88

3.2.5 Categorisation of susceptibility to measurement error 90

3.2.6 Timing of predictors and intended moment of model use 92

3.3 Results Part 1: Description of included studies and quality of reporting 92

3.3.1 Literature search and inclusions 92

3.3.2 PRISMA flowchart 94

3.3.3 Study design and aim 95

3.3.4 Candidate predictors 96

3.3.5 Sample size 97

3.3.6 Missing data 98

3.3.7 Model development 98

3.3.8 Model performance 100

3.3.9 Model validation 101

3.4 Results Part 2: Measurement error and intended moment of model use 101

3.4.1 Measurement error 101

3.4.2 Intended moment of using the model 107

3.5 Discussion 110

3.5.1 Key findings 111

3.5.2 Strengths and limitations 117

3.6 Recommendations 120

3.7 Conclusions and rationale for next chapter 122

4 The effect of measuring time-varying predictors at a different time point to that of the intended moment of use: an illustrative example 125

4.1 Introduction 125

4.1.1 Background 125

4.1.2 Chapter objectives 127

4.2 Methods 128

4.2.1 Datasets 128

4.2.2 Predictor measurement 131
4.2.3 Outcome measure 131
4.2.4 Statistical analysis: predictor-outcome associations 132
4.2.5 Statistical analysis: prognostic model performance 133
4.2.6 Statistical analysis: sensitivity analyses 135
4.3 Results 137
4.3.1 Data description 137
4.3.2 Preliminary analyses 141
4.3.3 Examination of predictor-outcome associations 141
4.3.4 Examination of prognostic model performance 143
4.3.5 Sensitivity analyses 144
4.4 Discussion 146
4.4.1 Principal findings 146
4.4.2 Explanation for the findings 147
4.4.3 Limitations and future research 149
4.5 Recommendations 151
4.6 Conclusions 152
4.7 Direction of the remainder of the thesis 153
5 External validation of prediction models for stillbirth using individual participant data (IPD) meta-analysis: the IPPIC study 155
5.1 Introduction 155
5.1.1 Background 155
5.1.2 Chapter objectives 158
5.2 Methods 159
5.2.1 Data 159
5.2.2 Models to validate 160
5.2.3 Imputation of missing predictor data 161
5.2.4 External validation of models 162
5.2.5 Decision curve analysis 166
5.3 Results 167
5.3.1 Prediction models to be included in validation 167
5.3.2 Characteristics of the IPPIC validation cohorts 171
5.3.3 Linear predictors 177
5.3.4 Performance statistics 182
5.3.5 Calibration plots 191
5.3.6 Net benefit 193

5.4 Discussion 195

5.4.1 Summary of the findings 195

5.4.2 Strengths and Limitations 196

5.4.3 Comparison to existing studies 199

5.4.4 Relevance to clinical care 200

5.4.5 Recommendations for further research 200

5.5 Conclusions 202

5.5.1 Remainder of the thesis 202

6 Calculating the power to examine prognostic factor effects when planning an individual participant data meta-analysis with a binary outcome 205

6.1 Introduction 205

6.1.1 Chapter outline 205

6.1.2 Background 205

6.2 A two-stage approach to estimating a prognostic effect in an IPD meta-analysis with a binary outcome 207

6.2.1 First-stage 208

6.2.2 Second stage 209

6.3 Calculating the variance of an unadjusted prognostic effect estimate for a binary outcome in a single study 210

6.3.1 Binary prognostic factor 211

6.3.2 Continuous prognostic factor 214

6.4 Calculating the power of a potential IPD meta-analysis project to estimate a prognostic effect with a binary outcome 215

6.4.1 Step 1: Estimate the variance of the prognostic factor separately for each study in the planned IPD meta-analysis 216

6.4.2 Step 2: Estimate the variance of the prognostic factor from the planned IPD meta-analysis 220

6.4.3 Step 3: Calculate the power of the planned IPD meta-analysis 221

6.4.4 Adjusting for other covariates 221

6.5 Applied examples 222

6.5.1 Example 1: Prognostic effect of age and sex on gastrointestinal bleeding in patients with cirrhosis and oesophageal varices (Poynard) 223

6.5.2 Example 2: Prognostic effect of age and sex on pain improvement in patients with osteoarthritis (STEER-OA) 231

6.6 Extension: Allowing for heterogeneity 238

6.7 Discussion 241

6.8 Conclusions 244

7 Discussion 247
7.1 Overview of the thesis 247
7.1.1 Summary of the chapters 248
7.1.2 Publications and outputs arising from this thesis 250
7.2 Contributions to applied and methodological research 251
7.3 Further research needs 256
7.4 Impact of measurement error 265
7.5 Machine learning 268
7.6 Conclusions 269
Appendix A 271
Appendix B 277
Appendix C 281
Appendix C1: Binary prognostic factor example 281
Appendix C2: Continuous prognostic factor example 285
Appendix C3: Allowing for heterogeneity 291
Reference List 293

Acknowledgments

I would like to express my gratitude to my lead supervisor, Prof. Richard Riley, for his constant inspiration, enthusiasm, and guidance throughout my PhD. I feel incredibly lucky to have had you as my supervisor. I would also like to thank my other supervisors, both past and present, Dr Joie Ensor, Dr Kym Snell, Prof. George Peat and Dr John Belcher, for their support, encouragement and advice at various points throughout this thesis.

I am grateful to Keele University for funding this work and to all my colleagues and friends in the School of Medicine. I would also like to kindly acknowledge the collaborators involved in the various projects, and the participants who provided their data toward the clinical applications within the thesis.

My deepest thanks go to Prof. Christian Mallen for believing in me and giving me this opportunity, and I am indebted to Dr Sara Muller for her support, compassion, and encouragement, but most of all, friendship.

Thank you to my friends, particularly Vicki and Emma, who have kept me sane whilst trying to finish my PhD with two small children. I am grateful to my family for always supporting me and a huge thanks goes to my in-laws, Sue and Brian, who have provided lots of childcare while I worked on my thesis.

I would like to thank my husband Jon, who without his love and support, this thesis would never have been completed. Thank you for holding the fort and keeping us all going while I have focused on writing, and thank you for reading everything multiple times even though it makes absolutely no sense to you.

Finally, to my children Arthur and Georgie, who never fail to make me smile.

Abstract

Prognosis research is an important part of medical research as it seeks to understand, predict, and improve future outcomes in people with a given disease or health condition. This thesis focuses on the application and development of statistical methods for prognosis research, with a particular focus on the identification of prognostic factors and the performance of risk prediction models.

The first part of the thesis considers the use of a single study for prognostic factor and prediction model research. Prognostic factors of adverse outcome in monochorionic diamniotic twin pregnancies are investigated and difference in nuchal translucency and crown-rump length were found to have prognostic value. The instability of developing a prediction model in small sample sizes is also illustrated. Then, a review of published prediction models is conducted which reveals potential concerns that measurement error may affect the predictors included in many models, and a lack of clarity about the timing of predictor measurements and the intended moment of using the proposed models. Recommendations for improved reporting are provided. A real example is then used to illustrate how displacing the collection of a time-varying predictor from the intended moment of model use leads to substantial differences in the predictor-outcome association, and the subsequent performance of the prediction model.

The second part of the thesis focuses on the synthesis of IPD from multiple studies. An IPD meta-analysis is used to validate existing stillbirth prediction models and demonstrates that the models should not be recommended for clinical practice due to poor predictive performance and insufficient clinical utility. Finally, a novel analytic method is developed to calculate the power of an IPD meta-analysis to examine prognostic factor effects with

binary outcomes, based on published study aggregate data, to help researchers decide on the benefit of the IPD approach in advance of collecting IPD.

List of tables

Table 2.1: Potential prognostic factors of interest 53

Table 2.2: Existing prognostic factors 56

Table 2.3: Maternal and fetal characteristics 63

Table 2.4: Outcomes 64

Table 2.5: Fetal adverse outcome composite - Unadjusted and adjusted results 65

Table 2.6: Adjusted results for the primary outcome (fetal adverse outcome composite) after using Firth's correction....66

Table 2.7: Coefficients from prognostic model for risk of a fetal adverse outcome before and after shrinkage 68

Table 2.8: Twin to twin transfusion syndrome (TTTS) 69

Table 2.9: Antenatally detected growth restriction (per pregnancy)....70

Table 2.10: Antenatal growth restriction (per fetus)....71

Table 2.11: Postnatally detected growth restriction (per pregnancy) 71

Table 2.12: Postnatal growth restriction (per baby)....72

Table 2.13: Single intrauterine fetal death 72

Table 2.14: Double intrauterine fetal death 73

Table 2.15: Maternal composite outcome 73

Table 2.16: Spontaneous preterm birth....74

Table 2.17: Neonatal composite outcome....74

Table 3.1: Search Filter – an adaptation of the search strategy by Geersing et al. (2012) which uses a variation of the Ingui filter (Ingui and Rogers, 2001) 86

Table 3.2: Items to be extracted from each article....89

Table 3.3: Explanation of reasons for measurement error in example predictors at high risk of error 106

Table 4.1: Design of the datasets 130

Table 4.2: Sample characteristics of the datasets 140

Table 4.3: Predictor-outcome association between a one unit increase in pain intensity and an unfavourable outcome 142

Table 4.4: Measures of model performance at the point of care and questionnaire in PROG-RES and POST 144

Table 4.5: Predictor-outcome associations and C-statistics in PROG-RES and POST at the point of care and post-consultation questionnaire when fitting general practice as a random effect 146

Table 5.1: Stillbirth prediction model equations externally validated in the IPPIC datasets 170

Table 5.2: Maternal characteristics and outcomes of IPPIC individual participant datasets used for validation 172

Table 5.3: Number of outcomes in each IPPIC individual participant dataset used for validation 174

Table 5.4a: Number and proportion missing (or not recorded) for each predictor in each dataset used for external validation 175

Table 5.4b: Number and proportion missing (or not recorded) for each predictor in each dataset used for external validation 176

Table 5.5: Summary of linear predictors and predicted probabilities for the models for each study used in the validation 178

Table 5.6: Study specific performance statistics 184

Table 5.7: Summary estimates of performance statistics from meta-analysis 186

Table 6.1: Aggregate data from 4 randomised trials included in the IPD meta-analysis project of Poynard et al. (1991) 224

Table 6.2: Results of the power calculations for the Poynard example for a range of γ values 226

Table 6.3: Results of the power calculations for the sex covariate in the Poynard example for a range of $\hat{\gamma}$ values and different VIFs 229

Table 6.4: Results of the power calculations for the age covariate in the Poynard example for a range of $\hat{\gamma}$ values and different VIFs 229

Table 6.5: Aggregate data from 31 randomised trials included in the IPD meta-analysis project of STEER OA (Holden et al., 2017) 232

Table 6.6: Results of the power calculations for the STEER OA example for a range of $\hat{\gamma}$ values 234

Table 6.7: Results of the power calculations for the sex covariate in the STEER OA example for a range of $\hat{\gamma}$ values and different VIFs 236

Table 6.8: Results of the power calculations for the age covariate in the STEER OA example for a range of $\hat{\gamma}$ values and different VIFs 237

Table 6.9: Comparison of the power in the common-effect model and random-effect model for an assumed OR for age of 1.04 in the Poynard example, considering a range of values for τ 240

List of figures

Figure 1.1: Logit function 14

Figure 1.2: Scatter plots showing the effect of measurement error in a predictor and the outcome in a linear regression model. 32

Figure 4.1: Participant flow - PROG-RES 138

Figure 4.2: Participant flow - POST 139

Figure 4.3: Predicted probability (95% CI) of an unfavourable outcome at 6 months by pain intensity rating estimated from the point of care and questionnaire models (female patient aged 65 (PROG-RES) or 66 (POST) from practice 30 in POST or 4 in PROG-RES). 143

Figure 5.1: Median and range of linear predictors (logit outcome probabilities) from the various prediction models....................181

Figure 5.2: C-statistic – Model 1b 187

Figure 5.3: Calibration slope – Model 1b....................187

Figure 5.4: Calibration-in-the-large – Model 1b 188

Figure 5.5: C-statistic - Model 2 188

Figure 5.6: Calibration slope – Model 2....................189

Figure 5.7: Calibration-in-the-large – Model 2 189

Figure 5.8: C-statistic - Model 3 190

Figure 5.9: Calibration slope – Model 3....................190

Figure 5.10: Calibration-in-the-large – Model 3 191

Figure 5.11: Calibration plot for model 1b - Stork dataset....................192

Figure 5.12: Calibration plot for model 2 - Chie dataset 192

Figure 5.13: Calibration plot for model 3 - Chie dataset 193

Figure 5.14: Decision curves for model 1b - Stork dataset.. 194

Figure 5.15: Decision curves for model 2 - Chie dataset.. 194

Figure 5.16: Decision curves for model 3 - Chie dataset.. 195

Figure 6.1: Results of the power calculations for the Poynard example with sex as the prognostic factor for a range of gamma values (presented as odds ratios) 226

Figure 6.2: Results of the power calculations for the Poynard example with age as the prognostic factor for a range of gamma values (presented as odds ratios) 227

Figure 6.3: Results of the power calculations for the Poynard example with sex as the prognostic factor for different values of the VIF for a range of gamma values (presented as odds ratios) 230

Figure 6.4: Results of the power calculations for the Poynard example with age as the prognostic factor for different values of the VIF for a range of gamma values (presented as odds ratios) 230

Figure 6.5: Results of the power calculations for the STEER OA example with sex as the prognostic factor for a range of gamma values (presented as odds ratios) 234

Figure 6.6: Results of the power calculations for the STEER OA example with age as the prognostic factor for a range of gamma values (presented as odds ratios) 235

Figure 6.7: Results of the power calculations for the STEER OA example with sex as the prognostic factor for different values of the VIF for a range of gamma values (presented as odds ratios) 237

Figure 6.8: Results of the power calculations for the STEER OA example with age as the prognostic factor for different values of the VIF for a range of gamma values (presented as odds ratios) 238

List of abbreviations

AUC	Area under the curve
AC	Abdominal circumference
AESD	Acute encephalopathy with biphasic seizures and reduced diffusion
AFP	Alpha-Fetoprotein
AIC	Akaike's information criterion
aOR	Adjusted odds ratio
BIC	Bayesian information criterion
BMI	Body mass index
BS	Brier score
CHARMS	Checklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies
CI	Confidence interval
CITL	Calibration-in-the-large
CPAP	Continuous positive airway pressure
CRL	Crown-rump length
DCA	Decision curve analysis
dIUFD	Double intrauterine fetal death
DMI	Depth of myometrial invasion
E	Expected
EFW	Estimated fetal weight
EPV	Events per variable
ERCP	Endoscopic retrograde cholangiopancreatography
ET	Endotracheal tube
FGR	Fetal growth restriction
GEE	Generalised estimating equations
GP	General Practitioner
HKSJ	Hartung-Knapp-Sidik-Jokman
IDI	Integrated discrimination improvement
IPD	Individual participant data

IPPIC	International Prediction of Pregnancy Complications Network
IQR	Inter-quartile range
IUFD	Intrauterine fetal death
LASSO	Least absolute shrinkage and selection operator
LP	Linear predictor
MAR	Missing at random
MC	Monochorionic
MCDA	Monochorionic diamniotic
MCMC	Markov chain Monte Carlo
MgSO4	Magnesium sulphate
MICE	Multiple imputation by chained equations
MLE	Maximum likelihood estimation
NAFLD	Non-alcoholic fatty liver disease
NPV	Negative predictive value
NRI	Net classification improvement
NRS	Numerical rating scale
NRS	Numerical rating scale
NT	Nuchal translucency
O	Observed
OR	Odds ratio
PAPP-A	Pregnancy-associated plasma protein A
PH	Proportional hazards
PMR	Polymyalgia rheumatic
POST	The Primary Care Osteoarthritis cluster randomised Trial
PPV	Positive predictive value
PROG-RES	The Prognosis Research observation study
PSA	Prostate specific antigen
PTB	Preterm birth
RCT	Randomised controlled trial
REML	Restricted maximum likelihood

ROC	Receiver operating characteristic
SD	Standard deviation
SE	Standard error
sFlt-1	Soluble fms-like tyrosine kinase-1
SGA	Small for gestational age
sIUFD	Single intrauterine fetal death
SpA	Spondyloarthritis
STEER-OA	Subgrouping and TargetEd Exercise pRogrammes for OsteoArthritis
TAPS	Twin anaemia plycythaemia sequence
TOPS	Twin oligohydramnios-polyhydramnios
TRIPOD	Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis
TTTS	Twin-twin transfusion syndrome
VIF	Variance inflation factor
WHO	World Health Organization

1 Introduction

This chapter aims to provide the foundations of the thesis, by presenting a detailed background to prognosis and risk prediction research, and other relevant topics such as measurement error and individual participant data (IPD) meta-analysis. Fundamental statistical methods are described, establishing core terminology and key concepts. The chapter concludes by outlining the aims and rationale for the thesis, including an overview of the structure of subsequent chapters.

1.1 Overview of the thesis

Predicting a patient's risk of future outcome events is an important part of medical research as it enables optimal treatment, informs clinical decision making and helps patients understand their risk. Prognosis research can be used to help predict future outcomes in patients with a particular disease or health condition (Hemingway et al., 2013) by identifying prognostic factors (predictors) and developing prediction (prognostic) models. Many studies are published each year which examine potential prognostic factors of outcome risk (Riley et al., 2013), and/or develop a prediction model, utilising values of multiple predictors to enable individualised risk (Steyerberg, 2010). Such models are intended "to assist clinicians with their prediction of a patient's future outcome and to enhance informed decision making with the patient" (Steyerberg et al., 2013) and thus the predictions from the models should have optimal performance when being practically implemented - the "intended moment of using the model" (Moons et al., 2014).

However, when developing such models, there are various methodological challenges and issues that need to be considered. For example, measurement error may affect the observed predictor values, which could potentially lead to biased estimates of predictor-outcome associations (Carroll et al., 2006, Gustafson, 2003, Prentice, 1982, Rothman et al., 2008). However, there has been little research into the impact that measurement error in the predictors may have on the predictions made and on the model performance. A recent study has found that measurement error in the predictors can reduce the area under the curve (AUC) and increase the Brier score (Khudyakov et al., 2015), but in general the impact of measurement error in prediction model research is relatively neglected. Hence, previously developed models and models currently being developed without consideration of measurement error, may be unknowingly providing misleading estimates of a patient's risk and the model may not perform as well as expected in practice.

A specific aspect of measurement error in the predictors is whether the predictors used in the model development were generally measured at the same time that the model is intended to be used in practice. The TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis) statement recommends to clearly define when the predictors used in the development of the model were measured (Collins et al., 2015) and states that "all predictors should be measured before or at the study time origin and known at the intended moment the model is intended to be used" (Moons et al., 2015). However, it is a concern that in many prognosis studies the timing of predictor measurement may differ from the intended moment of model implementation, which itself may lead to errors and misleading predictions.

IPD meta-analysis is another area of growing methodological interest for prediction models, where IPD from multiple existing studies are used to increase the quality and quantity of data available. Utilising IPD can improve the ability and power to examine the prognostic effects of a factor, or the development and/or validation of prediction models. However, completing an IPD meta-analysis project is a huge undertaking, requiring a substantial investment in terms of time and funding and can take years to complete. Yet, despite such large investment in the project, the final meta-analysis may still not have sufficient power to detect the effect of interest.

This thesis is focused on the application and development of statistical methods for prognosis research, with a particular emphasis on various methodological aspects including (i) the measurement error that may be present within the predictors used to develop prediction models, (ii) the impact of measuring a time-varying predictor after the intended moment of using the prediction model in practice, and (iii) the use of IPD from multiple studies to evaluate prognostic factor effects with binary outcomes, and for validating existing prediction models.

1.2 What is prognosis research?

Prognosis is the study of the risk (probability, likelihood) of future outcomes and events, so foreseeing, predicting, or estimating the probability of a future outcome. There are many uses for prognosis research, for example, it can be used in weather forecasting (Inness and Dorling, 2013), in sport (Crowder et al., 2002, Hughes and Franks, 2015) or in medicine (Steyerberg, 2010). In medicine, the focus of this thesis, prognosis research is used to understand and improve health outcomes (i.e. death or recurrence of cancer) in particular diseases and clinical conditions, by identifying prognostic factors that are associated with outcome risk and by developing prediction models that estimate an individual's risk of experiencing a particular health outcome (Moons et al., 2009a, Riley et al., 2013, Steyerberg et al., 2013).

Prognosis research is central to medicine and can be used to inform clinical research and clinical practice, as all screening, diagnostic and therapeutic actions aim to improve prognosis (Steyerberg, 2010). Screening can be used to detect early signs of disease, but the usefulness of the screening can be assessed by estimating the improvement in prognosis that is achieved by screening. Diagnostic tests can be used to detect an underlying disease, but then prognosis research is needed to evaluate whether the prognosis of patients whose disease is detected early is better than those who follow the natural course of the disease. Prognosis research is also used for making therapeutic decisions, to identify the best treatment option for a particular patient given his/her clinical condition and characteristics.

1.2.1 Prognosis research framework

Prognosis research consists of four inter-related themes, as described in a series of four articles which were published by the PROGRESS (PROGnosis RESearch Strategy) partnership in 2013 (Hemingway et al., 2013, Hingorani et al., 2013, Riley et al., 2013, Steyerberg et al., 2013), aiming to improve prognosis research. These articles introduce the framework of the four themes of prognosis research, with each article focusing on a specific theme:

1. Fundamental prognosis research (Hemingway et al., 2013): understanding overall outcomes in populations and settings in relation to current diagnostic and treatment practices
2. Prognostic factor research (Riley et al., 2013): discovering and evaluating specific factors (predictors) that are associated with prognosis
3. Prognostic model research (Steyerberg et al., 2013): the development, validation and impact of models that predict individual risk of a future outcome
4. Stratified medicine research (Hingorani et al., 2013): the use of prognostic information to help tailor treatment decisions to an individual or group of individuals with similar characteristics.

A brief description of what each of the aspects of the prognosis research framework is, why it is important and a clinical example, is given below. The focus of this thesis is on prognostic factor research (theme 2) and prognostic model research (theme 3). The terms prediction model and prognostic model are used interchangeably, as are predictor and prognostic factor.

1.2.1.1 Fundamental prognosis research

Fundamental prognosis research aims to describe and summarise future outcomes in people with a specific disease or health condition, in relation to current diagnostic and treatment practices (Hemingway et al., 2013). This could be expressed as an absolute risk (or rate) of a particular endpoint among groups with the same characteristics or in the same clinical setting and is often referred to as an average prognosis (overall risk) in a particular group.

Fundamental prognosis research is needed to inform health care professionals in order to assist clinical decision making and is also needed to assess the population burden of diseases and enable comparisons of effectiveness of health care systems. Fundamental prognosis research may also generate hypotheses for prognostic factor research, the second aspect of prognosis research.

Fundamental prognosis research provides the initial answer to "What is the prognosis of people with a given disease?", for example, what is the overall mortality of people with coronary heart disease? Bhatnagar et al. (2015) aimed to answer this question, investigating the mortality, prevalence, treatment, and costs of cardiovascular disease as a whole and of coronary heart disease in patients in the UK in 2014. They found that in 2012, cardiovascular disease was the most common cause of death in the UK for women, accounting for 28% of all female deaths. They then looked at regional variations, to assess whether the prognosis is different in different parts of the country, which found that mortality from cardiovascular disease varied widely throughout the UK, with the highest age-standardised cardiovascular disease death rates in Scotland (347/100,000) and the North West of England (320/100,000). This could then be compared over time to assess whether the prognosis is worsening or improving, and whether more research is needed to improve the outcome of those with coronary heart disease, particularly in certain areas of the country.

1.2.1.2 Prognostic factor research

A prognostic factor (predictor) is a measure that, among people with a given start point, is associated with a subsequent endpoint (Riley et al., 2013). Prognostic factor research aims to discover and evaluate factors that are associated with prognosis. Prognostic factor research is important as prognostic factors can change how diseases and health conditions are defined, can inform treatment recommendations and individual patient management, can provide a building block for prognostic models and can be potential predictors of treatment response for stratified medicine. They can also be used to monitor disease progression.

Prognostic factor research starts off with an exploration of factors and their association with the outcome, often the factors are identified by biological reasoning, but can be evaluated in hypothesis-free studies which aim to discover previously unsuspected factors. After exploration of prognostic factors, replication and confirmation is required, in multiple independent studies, together with the assessment of the prognostic value of the factor over and above other factors.

To do this, a regression model, usually linear, logistic or survival (detailed in Section 1.3), is fitted to the outcome including the prognostic factor of interest in the model. This provides an unadjusted estimate of the association between the prognostic factor and the outcome. Then to evaluate the prognostic value of the factor, over and above other factors, the model can be adjusted for other known prognostic factors, which will provide an adjustment estimate of the effect size.

Continuing with the example of prognosis in coronary heart disease patients, after finding the average prognosis of particular groups of patients, researchers may want to investigate whether certain individual-level factors are associated with prognosis as this may enable certain factors to be targeted (such as smoking) in order to improve prognosis in individuals (under a causal assumption), or to explain a decline or improvement in prognosis, as Unal et al. (2004) did. They found that coronary heart disease mortality decreased by more than 50% in England and Wales between 1981 and 2000 and approximately 60% of the decrease was attributable to reductions in major prognostic factors, particularly smoking.

Factors that are repeatedly found to be associated with prognosis, in multiple independent studies, could then be considered for use in developing a prediction model which would enable individualised prediction of a patient's future outcome.

1.2.1.3 Prognostic model research

Using a single predictor to estimate a patient's future prognosis rarely gives an adequate estimate (Moons et al., 2009a), and hence to improve the prediction, multiple predictors are combined to develop a model which can be used to estimate the risk of a specific outcome for individual participants. Prognostic model research is the formal combination of multiple predictors for which risks of a specific endpoint can be calculated for individual participants. Other terms often used to describe a prognostic model are prognostic (or prediction) index or rule, risk (or clinical) prediction model and predictive model.

Prognostic models aim to assist clinicians with their prediction of a patient's future outcome, using values of multiple predictors to make an individualised prediction of a future outcome occurring in a particular patient. This can help clinicians to make informed decisions with the patient, in terms of treatment choices. Prognostic models can also be used to help identify participants who are at risk of a poor outcome, to assign priority based on clinical need and to improve the design and analysis of randomised therapeutic trials.

The main use of prognostic models is to inform individuals about the future course of their illness and to guide doctors and patients in making joint decisions about their treatment. But it is important to remember that predicting outcomes is not the same as explaining the cause of the outcome.

A well-known and widely used simple example of a prognostic model is the Nottingham Prognostic Index (NPI) (Haybittle et al., 1982), which predicts the survival probability of women with newly diagnosed breast cancer by combining information on tumour grade, number of involved lymph nodes and tumour size. The formula for the NPI is:

$$\mathrm{NPI} = \big(0.2 \times \mathrm{tumour\ diameter(cm)}\big) + \mathrm{lymph\ node\ stage} + \mathrm{tumour\ grade}$$

Lymph node stage is coded as 1 = no nodes affected, 2 = ≤3 glands affected and 3 = >3 glands affected. Tumour grade is scored as 1, 2 or 3. The values of these predictors from a specific person can be inputted to calculate the NPI score, with a lower score suggesting a good outcome. The risk was categorised as high (>4.4), medium (2.8 to 4.4) and low (<2.8) and survival curves can be plotted for these risk groups, which allows patients and clinicians to visually observe the risk.

The development of prognostic models is discussed in more detail in Section 1.4.

1.2.1.4 Stratified medicine research

Stratified medicine research aims to identify those who will have the most clinical benefit or least harm from a specific treatment to allow the targeting of treatments dependent on certain characteristics (Hingorani et al., 2013). Prognosis research, particularly the development of prognostic models, is a fundamental component of stratified medicine. It enables the assessment of priorities for stratified medicine, through fundamental prognosis research and prognostic factor research. The use of prognostic models in stratified medicine can help identify those at the greatest risk, which are those who will have the largest absolute benefit from the treatment if the relative treatment effect is the same for all patients. Alternatively, statistical modelling can be used to identify those who will have the greatest benefit from the treatment due to the presence of individual factors that are predictive of an improved treatment response, i.e. when the relative treatment effect is shown to be different for different subgroups of patients in a randomised trial.

An example of how stratified medicine is used in practice is the primary care management of lower back pain to improve patient outcomes and economic benefits (Hill et al., 2011) by targeting treatment based on the patients prognosis (low, medium or high risk). This is estimated from a prognostic model which was developed to identify patient subgroups for initial treatment (Hill et al., 2008).

1.2.2 Sources of data

Different types of studies can be used for prediction research, such as retrospective studies, prospective studies, registry data and case-control studies. Each type of study has its advantages and disadvantages, but the study design seen as the best for prediction research is a prospective cohort study (Moons et al., 2014).

A retrospective study is simple and has low costs, but patients are identified retrospectively which means that previously recorded data needs to be relied on and this may lead to selection bias if any information is missing or incorrectly recorded. Similarly, the recording of the predictors and the outcome needs to have been reliable and the predictors available to be used to develop the prediction model and for prognostic factor research will be limited to those that were recorded in the study. Sample size and generalisability may also be an issue if the data being used is a single centre study.

Prospective studies allow better identification of included participants, as the inclusion and exclusion criteria can be specified prior to recruiting participants. Predictors, outcomes, and time-points can also be better defined prior to the study commencing, meaning that prospective cohort studies are more desirable than retrospective studies for prognostic factor research and for developing a prediction model.

Prediction models are commonly developed using data from registries, where data collection is prospective but its primary purpose is not for prediction research (Steyerberg, 2010). Again, using registry data for prediction research is relatively simple and low cost, but there are usually no pre-defined assessments made and outcome measurement is not assessed in line with a protocol. Advantages of registry data are that it can often be linked to many other sources to gain additional information, they include large sample sizes and have a wide representation of patients.

A nested case-control study can be an efficient option for prediction research when an outcome is relatively rare (Lee and Krischer, 2017), but this study design is seldom used in prediction research.

1.3 Statistical models used for prediction research

Below, statistical models and methods for prognostic factor and prediction model development are introduced, starting with linear and logistic regression models.

1.3.1 Linear regression

While continuous outcomes are common in medical research, and often receive the most attention in regression modelling literature (Steyerberg, 2010), they are not quite as common in clinical prediction models, but are included here for completeness. A linear regression model can be used when the outcome is continuous. The linear regression model can be written as

$$Y_j = \alpha + \boldsymbol{\beta X_j} + error \qquad \textbf{(1.1)}$$

Where j represents each person, α refers to the intercept and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_k)$ is the set of k regression coefficients that relate the predictors $\boldsymbol{X_j} = (X_{j1}, X_{j2}, \dots, X_{jk})$ to the outcome Y. The error is calculated as the observed Y minus the predicted Y $(\hat{Y})$. This error is assumed to have normal distribution and be independent of $\boldsymbol{X}$.

The regression coefficients, β_i, represent the increase in the estimated outcome, given a one unit increase in the value of X_i. Therefore, after estimation, the estimated outcome Y_j for patient j, is related to a linear combination of the predictors, and can be calculated by inputting patient j's individual values of $\boldsymbol{X}$ into **(1.1)**. Often penalisation or shrinkage is needed to improve model estimation, to help address the problem of overfitting to the dataset at hand.

1.3.2 Logistic regression

The logistic regression model is the most commonly used model for binary outcomes when the follow-up is relatively short and thus complete for (most) patients (Harrell, 2001, Steyerberg et al., 2013). The binary outcome, $Y_j = 0$ or 1, for person j, is linked via a logit-transformation of the outcome event probability to a linear combination of a set of k predictors, $\boldsymbol{X_j} = (X_{j1}, X_{j2}, \dots, X_{jk})$, and regression coefficients, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_k)$. The logit function (Figure 1.1) is used to restrict predictions to the interval $< 0, 1 >$. If we define the binary outcome to be $Y_j = 0$ or 1, with $Y_j = 1$ meaning the outcome of interest occurs for patient j, and α to be the estimated intercept, the logistic regression model is written as a linear function in the logistic transformation:

$$\begin{aligned} logit\left(\text{Prob}(Y_j = 1)\right) &= log\left(\frac{\text{Prob}(Y_j = 1)}{1 - \text{Prob}(Y_j = 1)}\right) \\ &= \alpha + \boldsymbol{\beta X_j}. \end{aligned} \qquad \textbf{(1.2)}$$

The regression parameters, $\boldsymbol{\beta}$, are usually estimated by the method of (penalised) maximum likelihood to give the fitted model:

$$logit\left(\text{Prob}(Y_j = 1)\right) = \hat{\alpha} + \boldsymbol{\hat{\beta} X_j}. \qquad \textbf{(1.3)}$$

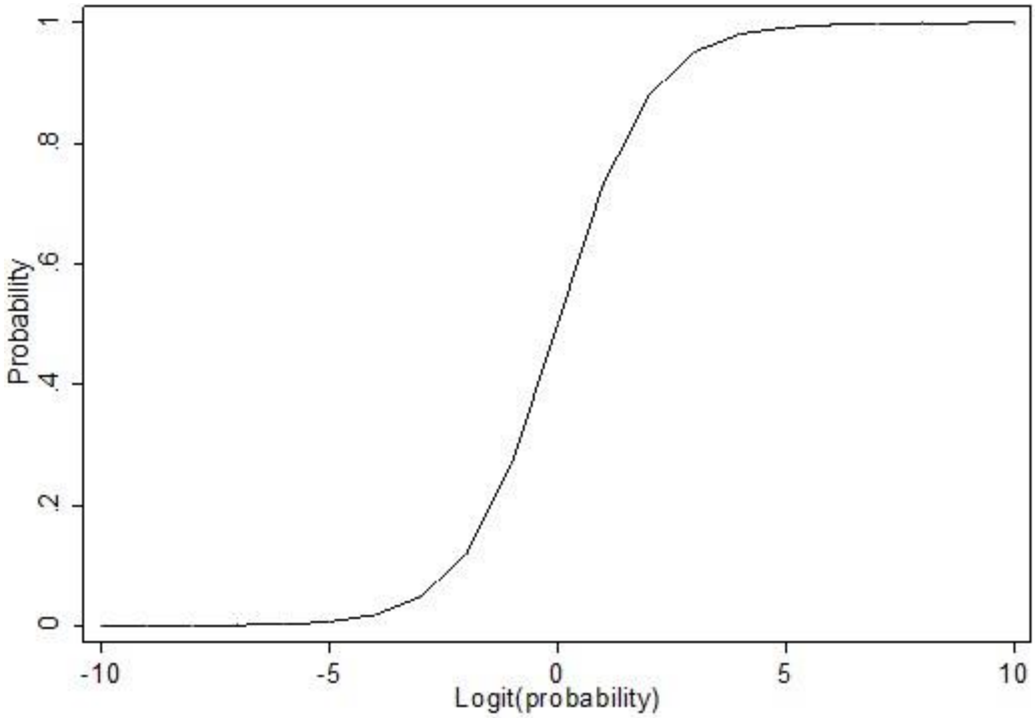
The regression parameters can be written in terms of odds ratios (OR), which are usually reported instead of the coefficient. The odds ratio for variable i, comparing two individuals who differ by one unit, interpreted as the estimated increase in the odds given the one-unit increase, is:

$$\text{Odds Ratio}(\text{X}_\text{i}) = \frac{\exp(\hat{\beta}_i[X_i + 1])}{\exp\,(\hat{\beta}_i X_i)} = \exp\,(\hat{\beta}_i[X_i + 1] - \hat{\beta}_i X_i) = \exp(\hat{\beta}_i)\,. \qquad \textbf{(1.4)}$$

Based on the fitted model, an estimated probability for a new patient, j, can be calculated by back transforming and replacing the X's with patient j's individual predictor values in the following equation:

$$\text{Prob}(Y_j = 1) = \frac{\exp(\hat{\alpha} + \widehat{\boldsymbol{\beta}}\boldsymbol{X_j})}{1 + \exp(\hat{\alpha} + \widehat{\boldsymbol{\beta}}\boldsymbol{X_j})}. \qquad \textbf{(1.5)}$$

Figure 1.1: Logit function



1.4 Developing a prediction model

There are several things to consider when developing a prediction model such as what the candidate predictors are, how missing data will be handled, how continuous predictors will be dealt with and how the final model will be chosen. After the model has been developed, the performance of the model should be assessed, and the model should be validated. These topics are discussed in detail below.

1.4.1 Selecting candidate predictors

Before a prediction model can be developed, the predictors to be used in the development of the model (candidate predictors) need to be identified. Potential predictors can be patient demographics, type and severity of disease, history characteristics, comorbidities, physical function status or subject health status and quality of life measures. There are two main methods to identifying predictors: using factors that are already known to be predictors, for example from systematic reviews, or using data-driven methods to test which predictors are (significantly) associated with the outcome.

Ideal candidate predictors are clearly defined, are measured reliably and as they would be in usual clinical practice and are simple to measure (Moons et al., 2012b, Steyerberg, 2010).

1.4.2 Sample size

The number of candidate predictors also needs to be considered, especially when the sample size is small. The statistical power to develop a prediction model is often based on the number of outcome events per the number of candidate predictors. When the number of participants experiencing an outcome relative to the number of predictors being considered is small, overfitting typically occurs (Steyerberg, 2010). Historically, it has been recommended that there are at least 10 outcome events per candidate predictor (Peduzzi et al., 1996) as a general rule of thumb. For studies using a continuous outcome, a recommended rule of thumb is 20 participants per candidate predictor (Harrell, 2001, Moons et al., 2014), although a suggestion of 2 participants per predictor has been made (Austin and Steyerberg, 2015). However, more recent research has shown this is not a good way of determining sample size. Ogundimu et al. (2016) has shown that the sample size does not only depend on the events per the number of candidate predictors, but also the prevalence of binary predictors. Further, Riley et al. (2020) state that "the sample size should be at least large enough to minimise model overfitting and to target sufficiently precise model predictions". Riley and colleagues advise that the actual required sample size is context specific and depends not only on the number of events relative to the number of candidate predictor parameters but also on the total number of participants, the outcome proportion and the expected predictive performance of the model (Riley et al., 2020). Hence, even if a dataset is used which has greater than 10 EPV, if a rare binary predictor is included then this can still create imprecise estimates.

Penalisation techniques, such as uniform shrinkage estimated via bootstrapping, or penalised regression methods such as ridge regression, the least absolute shrinkage and selection operator (lasso), and elastic net (Hoerl and Kennard, 1970, Hoerl and Kennard, 2000, Tibshirani, 1996, Zou and Hastie, 2005), are recommended to address overfitting. Yet, shrinkage and penalty terms are estimated with uncertainty from the development data set (Riley et al., 2021a).

1.4.3 Missing data

The next thing to consider when developing a prediction model is how to handle the data, for example, how to deal with any missing data.

Missing data can be a problem in all kinds of medical studies, prediction modelling being no exception, particularly as data from existing registries are increasingly being used in prediction modelling studies and these databases are especially prone to missing data (Moons et al., 2014). Missing data can occur when participants do not respond to questions, when equipment/technology fails, when participants withdraw from a study before the end, or because of data entry issues. The amount and type of missing data can have a big impact on the model development and the accuracy of the predictions from the model. The method in which to handle the missing data is also an important consideration (Kang, 2013).

A common but potentially naive approach to missing data is complete case analysis. When complete case analysis is performed, all participants with a missing value for any variable are deleted, but this can leave a non-random subset of the participants if the data are not missing completely at random (which they rarely are) (Royston et al., 2009), and hence yields invalid predictive performance and biased predictor-outcome associations (Donders et al., 2006, Harrell, 2001).

Multiple imputation is acknowledged as usually the preferred method of handling missing data, if the data is missing at random (Janssen et al., 2010). Missing at random means that the data are missing independently of any unobserved data, and so are related to observed covariates and outcomes. Multiple imputation creates multiple copies of the dataset, replacing the missing values by imputed values (Sterne et al., 2009), which are drawn from a posterior distribution. Standard statistical methods are applied to each of the imputed datasets and the results are averaged to give an overall estimate. Multiple imputation techniques are available in several commonly used techniques, but care needs to be taken as the validity of the results depends on the modelling being completed carefully and appropriately (Sterne et al., 2009).

1.4.4 Continuous predictors

The handling of continuous predictors within the prediction model needs to be considered before developing the model. Dichotomising or categorising continuous data should be avoided (Royston et al., 2006), as this loses information and hence reduces the power to detect genuine predictors and their relationship with outcome (Cohen, 1983).

Continuous predictors may also have a non-linear relationship with the outcome. Ideally, the linearity of the relationship should be assessed, and a suitable transformation performed if non-linearity is apparent. Two ways of executing a transformation is to use fractional polynomials (Royston et al., 1999) or to use restrictive cubic splines (Durrleman and Simon, 1989). Briefly, fractional polynomials provide flexible parametrisation for continuous variables by transforming the variable for different values of powers, or combination of powers, from a predefined set (–2, –1, –0.5, 0, 0.5, 1, 2, 3). Restrictive cubic splines split up the range of predictor values and fits a separate curve to each segment, defined so that the resulting overall curve is smooth and continuous.

1.4.5 Selection of final model

Once the candidate predictors have been selected and the methods of handling the data have been decided, the modelling of the data can commence to develop the prediction model.

A prediction model could be developed for a continuous outcome, such as pain rating, in which case a linear regression model could be used. More commonly they are developed for a binary event, for example death or diagnosis. The two main methods used are a logistic regression model or a time-to-event model (survival analysis). The logistic regression model was introduced in Section 1.3.

The method of selection of predictors to be included in the multivariable prediction model can be a source of bias in the model development (Moons et al., 2014). There is no consensus on the best method for selection of the final predictors, though some recommendations have been made. A popular method is to use automatic selection procedures, including forward selection, backward selection, and stepwise selection. If automatic selection methods are used, then backward elimination is preferred to forward elimination, as this method starts with the full model and removes predictors deemed to have little predictive value after full adjustment, rather than starting with an empty model and building up (Royston and Sauerbrei, 2008), potentially missing important predictors. The worst approach is inclusion based on significance in univariable (unadjusted) analysis (Royston et al., 2009, Sun et al., 1996). Two concerns when using this form of model selection are firstly that this is likely to introduce error, as the correlation between the predictors is not properly controlled for, and secondly that predictors are not included in the final model because they were not significant in the univariable model simply due to chance.

Predictor selection based on significance can produce optimism due to overfitting, and the variables selected highly depend on the significance level used. An alternative approach is to include all the candidate predictors within the final model, which can reduce the potential for overfitting, but it can often be impractical to include all candidate predictors. Another approach is to use a penalised regression, such as the lasso or elastic net, which includes variable selection in the model estimation (Tibshirani, 1996, Zou and Hastie, 2005).

1.4.6 Model performance

The discrimination and calibration of a prediction model are two of the main measures used when assessing the performance of a model.

1.4.6.1 Discrimination

Discrimination is how well a model separates those who experience an event from those who do not experience an event. Discrimination can be presented by the receiver operating characteristic (ROC) curve for logistic regression models, and quantified using measures such as the concordance statistic (C-statistic), net reclassification improvement (NRI), and integrated discrimination improvement (IDI).

The most common of these is the C-statistic, which is equivalent to the area under the curve (AUC) for a logistic regression model (Steyerberg, 2010) and has also been extended to the Cox PH model setting (Harrell et al., 1996). The C-statistic is the probability that for any randomly selected pair of individuals, one who experiences the outcome of interest and one who does not, the model assigns a higher probability to the individual who experiences the outcome. In the survival setting, the Harrell's C-statistic is the probability that for any randomly selected pair of individuals, the model assigns a higher probability to the individual who survives longer. A C-statistic of 0.5 indicates that the model is no better than chance, and a value of 1 indicates that the model perfectly classifies the individuals.

1.4.6.2 Calibration

The calibration of a prediction model is the amount of agreement between the observed outcomes and the predictions (Steyerberg, 2010). Calibration is the ability of the model to accurately predict the absolute risk level across groups of similar individuals (Crowson et al., 2013).

Calibration can be graphically assessed by plotting the predicted probabilities against the observed probabilities and fitting a calibration slope. The calibration slope is a smooth non-linear line fitted between these predicted and observed probabilities on the logit scale. The calibration slope would be equal or very close to 1 for good calibration. However, a slope <1 indicates overfitting of the model, whereas a slope >1 indicates underfitting.

Another measure of calibration is the 'calibration-in-the-large'. For a logistic regression model this is the difference between the mean number of predicted events and the mean number of observed events. It can be calculated by regressing the observed outcome on the predicted probabilities, which provides a measure of effect size and a confidence interval as well as a p-value and does not require the data to be grouped, hence is preferred over using the Hosmer-Lemeshow test (Van Calster et al., 2019).

The ratio of expected (E) to observed (O) events can also be used to assess calibration. For logistic regression models this can be calculated by dividing the number of expected events by the number of observed events. The ratio should be close to one if the model calibrates well.

1.4.6.3 Other measures

Overall performance statistics can also be used such as R^2, which measures the proportion of explained variation, or the Brier score which measures the overall model fit. The Brier score is the average squared difference between the observed outcome and the predicted probability (Harrell, 2001). Other measures used to assess model performance at a particular threshold are sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV). Sensitivity is the true positive rate whereas specificity is the true negative rate, and PPV and NPV are the proportions of positive and negative results that were a true positive or true negative result, respectively. The clinical utility can also be examined using decision curves, which is a weighted function of sensitivity, specificity, and outcome prevalence (Vickers and Elkin, 2006).

1.4.7 Model validation

When a performance measure is calculated in the same dataset as the model was developed in, this is known as the apparent predictive performance, and this tends to be biased (Moons et al., 2014). Hence, the performance of a prediction model should not be solely evaluated in the development dataset but should also be evaluated in an independent dataset. This is known as external validation. However, there are also ways to perform validation within the development data, a process known as internal validation.

1.4.7.1 Internal validation

There are several approaches to internal validation. One method is to split a large dataset into two samples: a training/development sample and a validation sample (Picard and Berk, 1990). Data are often split randomly, which means that the data are likely to only differ due to chance (Altman and Royston, 2000). Another approach would be to continue enrolling participants into the study after the development sample has been recruited and enrol these additional participants into a validation sample. Data splitting is simple, but means that the sample size for the development of the model is smaller than it could be, which could mean that the parameter estimates are not as precise as they could be if all the data were used (Steyerberg, 2010) and is not providing a truly independent dataset for validation so does not give a true indication of how the model will perform in other populations. Hence, splitting the data set at a single point for validation is not considered a useful technique for validation of a prediction model.

Rather than splitting at a single point, cross validation could be used which splits the data multiple times at different points, using different parts of the data for the development and validation each time. This can be done by leaving a single observation out of the analysis each time and then predicting the outcome for that individual by using the model developed for the remaining participants, then summarising the performance for each time this is repeated. Another method of cross validation is to use 5 or 10-fold cross validation, by dividing the data into parts (e.g. 5 or 10), developing the model in all but one of these parts and testing the model in the other part, repeating for each part and then averaging the model performance (Steyerberg, 2010).

Internal validation could also be performed using resampling techniques such as bootstrapping to estimate the amount of optimism in the developed model which can then be used to shrink the estimated regression coefficients in the prediction model (Efron, 1983, Efron and Tibshirani, 1994). Bootstrap samples of the same size of the original sample are drawn with replacement and a prediction model is developed in each bootstrapped sample. This model is then evaluated in the bootstrapped sample and the original sample, and the difference in the performance is the optimism of the model, which is averaged over the number of bootstrap samples to calculate the best estimate of optimism. Optimism adjusted performance measures can then be derived. In particular, the optimism adjusted calibration slope is often used as a uniform shrinkage factor to penalise predictor effects post-estimation. Bootstrapping is preferred for internal validation as it does not require any of the data to be excluded.

1.4.7.2 External validation

Internal validation of a prediction model is necessary to account for overfitting within the development data but does not provide information on the generalisability or transportability of the model. Hence, external validation should also be performed (Moons et al., 2012a). External validation evaluates the model developed in a completely external dataset to that it was developed in, but with a similar patient population to the development data. The external dataset could be collected at a different time or at a different location to that of the original data.

A framework developed by Debray and colleagues (Debray et al., 2015) proposes three steps to external validation studies:

1. Investigate extent of relatedness; investigate how related the individuals from the validation sample are with the development sample.
2. Assess model performance; assess the performance observed when the existing prediction model is tested in the development and validation sample.
3. Interpretation of model validation results; examine how well the model reproduces the target population of the development sample or how well the model transports to a different but related target population.

To increase the generalisability of the model, external validation of the model should be performed in multiple independent external datasets including participants from different settings and different populations, to assess how well the model works in different scenarios.

Although external validation is viewed to be an essential part of the model development (Altman et al., 2009), still very few studies do externally validate the prediction model, as will be discussed further in a review of recently published prediction models in Chapter 3.

1.5 Why is the timing of measurements important?

The TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis) statement recommends to clearly define when the predictors used in the development of the model were measured (Collins et al., 2015), and states that "all predictors should be measured before or at the study time origin and known at the intended moment the model is intended to be used" (Moons et al., 2015). This is because if a prediction model is to be used to predict the probability of a patient experiencing a future outcome, then including a predictor that will not be known at the time that the model will be applied will be of little value. This is particularly true if the predictor is time-dependent and not measurable at baseline (point of using the prediction model), as the use of these predictors can produce biased estimates and impact the study's conclusions (van Walraven et al., 2004). The same issue applies for prognostic factor research: factors under investigation should be measured at or before the start-point when prognosis is clinically relevant.

For example, a prediction model could be developed using the last recorded blood pressure measurement in the patient's primary care record, but if the clinician then has the patient's current blood pressure available when making a prediction at the time of the patient consultation, this means there is a time difference between the measure used to develop the model and the measure used when implementing the model, potentially creating a prediction model which is sub-optimal when applied in practice.

Another example is if researchers collect baseline information on predictiors from participants three weeks after their consultation with their general practitioner, and use this baseline information to develop a prediction rule for the outcome of the consultation (measured six months after the consultation). If this model is to be applied at the point of care (the consultation), the predictor values may be different to those that were measured at the baseline and hence the model may not perform as well as expected when used at the point of care as a result of this.

The impact of measuring a time-varying predictor after the intended moment of model use will be examined in a real example in Chapter 4.

1.6 Measurement error

1.6.1 What is measurement error?

Measurement error is a difference between the measured values of a variable and the true values of the variable, or if the variable is categorical, the classification to an incorrect category. All variables are generally measured with some degree of error (Armstrong, 1998).

The term measurement error is generally used when the variable of interest is continuous, whereas when the variable is categorical, the error is commonly referred to as misclassification. Mismeasurement is a term that has been used to denote error from both continuous and categorical variables (Gustafson, 2003). Within this thesis, the term measurement error will be used generally, and misclassification will be used when specifically referring to categorical variables.

Measurement error is common within clinical studies, particularly observational studies (Guolo, 2008). Some of the main reasons that measurement error may occur are:

- Fluctuations in human samples/biological variability (e.g. blood pressure changing in a short period of time)
- Inaccuracy of measurement instruments (e.g. if scales being used to measure weight were not calibrated correctly)
- Imperfect recall (e.g. when asking a patient to report previous symptoms, they may not remember correctly)
- Cost/resource limitations (e.g. using family history instead of genetic testing for diseases such as breast cancer)

- Subjective nature of measures (e.g. patients could report pain levels differently)
- Laboratory or measurer error (e.g. inaccuracy of blood tests analysed in a laboratory)
- Timing error (e.g. using most recently recorded values recorded in health records rather than current values)

1.6.2 Differential and non-differential measurement error

Non-differential measurement error occurs when everyone has the same probability of measurement error or misclassification, so the distribution of the variable measured with error depends only on the actual variable and not on the outcome variable. If $\boldsymbol{X}$ is the true unmeasured variable, $\boldsymbol{W}$ is the error-prone measure of $\boldsymbol{X}$, $\boldsymbol{Z}$ is a precisely measured variable and $\boldsymbol{Y}$ is the outcome variable, then the measurement error in $\boldsymbol{W}$ is non-differential if no additional information on $\boldsymbol{Y}$ is contained in $\boldsymbol{W}$ other than what is available in $\boldsymbol{X}$ and $\boldsymbol{Z}$, so the conditional distribution of $\boldsymbol{Y}$ given $(\boldsymbol{W}, \boldsymbol{X}, \boldsymbol{Z})$ is the same as the distribution of $\boldsymbol{Y}$ given $(\boldsymbol{X}, \boldsymbol{Z})$. When the conditional distribution is not the same, so the error depends on the actual value of other variables (Rothman et al., 2008), the error is differential. Non-differential misclassification is present if all participants have the same probability of being misclassified (Sorahan and Gilthorpe, 1994). Measurement error is typically non-differential, for example, if the association between blood pressure and myocardial infarction is being studied, but blood pressure is measured with error due to fluctuations in the patient and inaccuracy of the blood pressure instrument, no more information will be gained about the patient’s likelihood of a myocardial infarction from the measure of blood pressure recorded than the patient’s true measure of blood pressure.

There are feasible situations in which the error can be classified as differential, for example, when a surrogate measure is used in place of the true measure, there may still be an association between $\boldsymbol{W}$ and $\boldsymbol{Y}$, even after adjusting for $\boldsymbol{X}$. Another example where the measurement error may be differential could be in a case-control study, where recall bias may lead to different error in the cases and controls (Armstrong, 1998).

The parameters of a model given the true measures, even when those true measures are not observable, can typically be estimated when non-differential measurement error is present, but not when the measurement error is differential (Carroll et al., 2006).

1.6.3 Additive and multiplicative measurement error

Additive measurement error is when a measured (surrogate) variable $\boldsymbol{W}$ is equal to an unobservable (true) variable $\boldsymbol{X}$ plus some error, $\boldsymbol{U}$:

$$\boldsymbol{W} = \boldsymbol{X} + \boldsymbol{U} \qquad \textbf{(1.6)}$$

In this model, the additive measurement error is non-differential, unbiased and normally distributed. The measurement error is said to be unbiased if $E[\boldsymbol{W}|\boldsymbol{X}] = \boldsymbol{X}$ (Gustafson, 2003).

Multiplicative measurement error occurs when the amount of error in $\boldsymbol{W}$ is proportional to the actual value of $\boldsymbol{X}$:

$$\boldsymbol{W} = \boldsymbol{U} \times \boldsymbol{X} \qquad \textbf{(1.7)}$$

Separate methods of analysis for additive and multiplicative measurement are not needed (Gustafson, 2003).

1.6.4 Dependent and independent measurement error

Dependent error is error that depends on the errors of other measured variables, otherwise it is known as independent (or nondependent) (Rothman et al., 2008). Dependent errors most often occur when the explanatory variables and the outcome are recorded in the same way (Lash and Fink, 2003), e.g. by questionnaire or interview, because if a particular patient is likely to over exaggerate a response, their other responses are also likely to be over exaggerated and hence the errors of one variable are dependent on the errors of the other variables. Dependent errors can logically be avoided by using separate sources for collecting information on explanatory variables and the outcome (Kristensen, 1992).

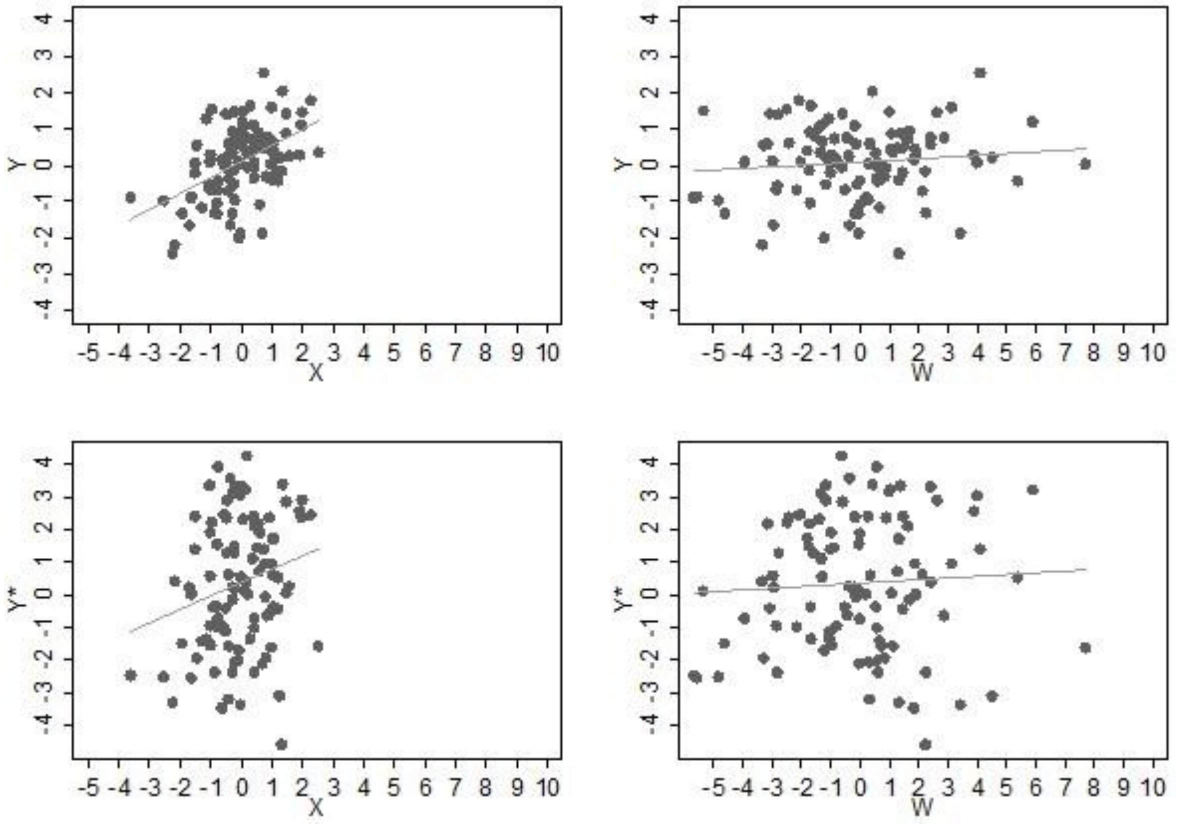
1.6.5 Impact of measurement error

Analysis which does not account for measurement error, i.e. treats $\boldsymbol{W}$ as if it is equal to $\boldsymbol{X}$, is referred to as naïve (Gustafson, 2003). In assessing measurement error, careful attention must be given to the type and nature of the error (Carroll et al., 2006) because different types of error have different impacts on the inferential results and the different available techniques to correct for the error (Guolo, 2008).

There are three main effects of classical measurement error being present in predictor values (Carroll et al., 2006). Firstly, it may lead to biased estimates of the parameters derived from statistical models (rather than just less precise inferences (Gustafson, 2003)). Secondly, it creates a loss of power for detecting relationships, and thirdly, it masks the features of the data, meaning that it is harder to spot relationships via graphical methods.

Measurement error in the outcome does not greatly bias the parameter estimates, just adds uncertainty (Gustafson, 2003), whereas measurement error in the predictor does bias the parameter estimate, however, not always necessarily towards the null (Fosgate, 2006). This can be seen in a linear regression model using scatter plots in Figure 1.2. The top-left plot shows the relationship between a predictor $\boldsymbol{X}$ and the outcome $\boldsymbol{Y}$, randomly simulated from a bivariate normal distribution both with mean 0 and standard deviation 1, and a correlation coefficient equal to 0.5. The line on the plot shows the line of best fit. In the top-right plot, $\boldsymbol{W}$ was observed in place of $\boldsymbol{X}$, where $\boldsymbol{W}$ was obtained by adding random normally distributed error to $\boldsymbol{X}$. We can see here that by adding error to the predictor, the slope of the regression line is flatter, hence the parameter estimate is biased. Whereas in the bottom-left plot, error has been added to the outcome to create $\boldsymbol{Y*}$, and we can see that the addition of this error does not change the estimate of the slope greatly, just increases the uncertainty. The bottom-right plot shows the relationship between the predictor measured with error, $\boldsymbol{W}$, and the outcome measured with error, $\boldsymbol{Y*}$. We can see that the regression line is similar to that when regressing $\boldsymbol{Y}$ on $\boldsymbol{W}$, with more uncertainty. Hence, throughout this thesis, the focus will be on measurement error in the predictors rather than the outcome.

Figure 1.2: Scatter plots showing the effect of measurement error in a predictor and the outcome in a linear regression model.



1.6.6 Measurement error in prediction models

It has been reported that there may not be a need to account for the measurement error within the prediction model. Carroll et al. (2006) state that if a true predictor (X) is measured with error and this error-prone predictor (W) is used to develop a prediction model to predict a participants outcome (Y), then if it is this same error-prone measure (W) that is available in practice when implementing the prediction model, rather than the true predictor (X), then there is little issue with using W to develop the prediction model. Although, a problem may arise under two circumstances:

1. When a prediction model is developed in one population but is intended to be used to predict the outcome in another population.
2. When a surrogate for X is used to develop the prediction model, but the true value of X will be available when applying the model.

Another viewpoint could be that a prediction model should provide the most accurate estimate possible, and if a predictor used in the development of a model is measured with error then the estimates of the predictor outcome associations will be sub-optimal. Even if the same predictor is then measured with this same error when applying the model in practice, this does not necessarily mean that the probability of outcome being estimated from the model is a true estimate of a patient's predicted probability, as using a predictor measured with error when implementing the model may not cancel out the bias, but could potentially underestimate a participants probability of future outcome.

Therefore, measurement error within prediction models could produce sub-optimal predictions of a patient's future outcome. Indeed, measurement error in prediction models has been shown to reduce the AUC and increase the Brier score (Khudyakov et al., 2015), but this study focused on the gain in prediction performance from using error-free predictors instead of error-prone predictors, rather than the gain in prediction performance from accounting for the measurement error in the model when the true error-free values are not known. The study also only evaluated the scenario where only one error-prone predictor was included in the prediction model.

Another study found that both random and systematic error in self-reported health data influences the calibration, discrimination and predicted risks (Rosella et al., 2012). This study assessed the impact of random and systematic error in self-reported height and weight on the performance of a model used to predict diabetes. The authors found that random error reduced the calibration and discrimination, and biased the predicted risk upwards, whereas systematic error reduced the calibration and biased the predicted risk in the direction of the bias but had no effect on the discrimination.

However, in general the impact of measurement error in prediction model research is relatively neglected.

1.7 Individual participant data (IPD) meta-analysis for prognosis and prediction studies

Multiple studies are often conducted to investigate the same prognostic factors; however, they often have conflicting findings and are of variable quality. Similarly, multiple prediction models are frequently developed with the aim of predicting the same outcome. This motivates the need for evidence synthesis: the combination of data from multiple studies to provide an overall summary of current knowledge (Riley et al., 2021b). There is an increasing interest in synthesis of prognostic factor effects and the synthesis of data for the external performance of prediction models (Debray et al., 2017, Riley et al., 2013, Steyerberg et al., 2013). The statistical technique used to combine quantitative data obtained from multiple research studies is known as meta-analysis. Traditionally, most meta-analyses have used aggregate data extracted from study publications, but there is a growing demand for meta-analyses that utilise the IPD (Harbord et al., 2008, Macaskill, 2004, Macaskill et al., 2010, Rutter and Gatsonis, 2001), the raw patient level data recorded for each participant in a research study, to calculate and synthesise the effect of interest from each study. IPD meta-analysis projects offer novel opportunities for the development of clinical prediction models and can allow the performance of existing models to be externally validated across different populations. Existing prediction models often show poor predictive performance when tested or applied in other populations or settings than used for model development, however, IPD meta-analysis can allow researchers to update or tailor the existing model equation to improve the performance in particular populations or settings (Riley et al., 2016).

1.7.1 Rationale for IPD meta-analysis

Compared to using aggregate data (e.g. prognostic factor estimates) from publications, IPD projects can potentially provide substantial improvements to the quantity and quality of data available (Riley et al., 2021b). The quality of the data is improved as detailed checks are used to ensure the completeness, validity, and internal consistency of data items for each study.

There is also a greater ability to standardise outcome and covariate definitions across the studies, and it can support more flexible and sophisticated analyses than possible with aggregate data as there is no need to be restricted by the original analysis methods of the study. IPD meta-analyses can allow continuous variables that may have previously been categorised to be analysed on their continuous scale, and likewise, potential non-linear relationships can be examined.

Another advantage of using IPD over aggregate data is that unpublished studies can be included, as can any outcomes that were not reported for published studies, or participants who were inappropriately excluded from the original analyses (Macaskill, 2004, Macaskill et al., 2010, Nikoloulopoulos, 2017). This can help evade potential reporting biases (Higgins et al., 2009), increase the quantity of information available for analyses, and therefore boost the statistical power to detect genuine effects (Riley et al., 2011).

1.7.2 Statistical methods for an IPD meta-analysis

There are two approaches to analysing an IPD meta-analysis; two-stage and one-stage (Burke et al., 2017, Simmonds et al., 2005). The two-stage approach first uses standard regression analysis in each study separately to obtain the aggregate data needed for the second stage. The second stage then uses standard common-effect or random-effects meta-analysis models to synthesis the aggregate data and produce summary results and forest plots (Riley et al., 2021b).

The one-stage approach analyses the IPD from all the studies in one model whilst allowing for the clustering of participants within studies (Abo-Zaid et al., 2013). The one-stage approach requires a hierarchical (multilevel) regression model appropriate to the type of outcome data being synthesised. The one-stage approach utilises a more exact statistical likelihood than the two-stage meta-analysis approach, which results in better statistical properties, particularly when the included studies have few participants or outcome events (Altman et al., 2007, Macaskill et al., 2010, Putter et al., 2010).

However, the results of a one-stage and two-stage IPD meta-analysis are usually similar, with the two-stage performing just as well as the one-stage approach in most situations unless the outcome event is sparse (Burke et al., 2017). The two-stage approach is often preferred (Chu and Cole, 2006, Leeflang et al., 2012), as the second-stage uses well-known meta-analysis methods that are relatively straightforward, is often computationally faster than the one-stage approach and naturally separates the within-study information from across-study information.

1.7.2.1 Meta-analysis assuming a common-effect model

Consider the aim to summarise a prognostic factor effect or the performance of a prediction model. In the second stage of the two-stage approach, the estimates (e.g. log odds ratios or calibration slopes) obtained in the first stage are combined using either a common-effect model or random-effect model. The common-effect model assumes the prognostic effect or model performance is the same in every study.

Define $\hat{\gamma}_i$ to be the estimate of a particular effect of interest, either the prognostic factor effect or measure of model performance, where $i = 1$ to I studies, and let $\hat{S}_i^2$ be the associated variance of $\hat{\gamma}_i$. The common-effect meta-analysis model assumes that the true effect, γ, is the same in all studies, and that $\hat{\gamma}_i$ are estimates of this common effect as given by (Riley et al., 2013, Whitehead and Whitehead, 1991):

$$\hat{\gamma}_i \sim N(\gamma, S_i^2) \qquad \textbf{(1.8)}$$

This assumes that the effect estimates from the first stage are normally distributed and that their variances are known. Maximum likelihood (ML) estimates can be used to fit the model in **(1.8)** which leads to the following analytic solutions for the summary effect estimate ($\hat{\gamma}$) and its variance ($\mathrm{var}(\hat{\gamma})$):

$$\hat{\gamma} = \frac{\sum_{i=1}^{I} \hat{\gamma}_i w_i}{\sum_{i=1}^{I} w_i} \qquad \textbf{(1.9)}$$

$$\operatorname{var}(\hat{\gamma}) = \frac{1}{\sum_{i=1}^{I} w_i} \qquad \textbf{(1.10)}$$

It can be seen from equation **(1.9)** that the summary result, $\hat{\gamma}$, is a weighted average with the weight of each study, w_i, defined by:

$$\mathrm{w_i} = \frac{1}{S_i^2} \qquad \textbf{(1.11)}$$

Hence, the meta-analysis gives more weight to the estimates from studies with the smallest S_i^2 values (i.e., those with more precise effect estimates, which is generally the studies with the largest number of participants or outcome events).

1.7.2.2 Meta-analysis assuming a random-effects model

The assumption of a common effect across all included studies is usually inappropriate, as the prognostic factor effects or model performance will usually differ across studies (Higgins et al., 2009). This is known as between-study heterogeneity. To allow for unexplained between-study heterogeneity in the parameter of interest, the γ_i can be made random (Riley et al., 2009), so the true effects are allowed to be different but are assumed to be from a particular distribution (Levis et al., 2017). The meta-analysis then needs to summarise this distribution of the γ_i. If we extend the model in **(1.8)**, the meta-analysis model with random effects is:

$$\hat{\gamma}_i \sim N(\gamma_i, S_i^2) \qquad \textbf{(1.12)}$$

$$\gamma_i \sim N(\gamma, \tau^2) \qquad \textbf{(1.13)}$$

As with the model in **(1.8)**, the estimates of S_i^2 are assumed to be known. The between-study variance of the true prognostic effect (or model performance parameter) is denoted by τ^2. If τ^2 is equal to zero, then there is no between-study heterogeneity in the parameter of interest, and the model in **(1.13)** reduces to the common-effect meta-analysis as in **(1.8)**.

The ML estimate solution for $\hat{\gamma}$ is a weighted average of the treatment effect estimates where:

$$\hat{\gamma} = \frac{\sum_{i=1}^{I} \hat{\gamma}_i w_i^*}{\sum_{i=1}^{I} w_i^*} \tag{1.14}$$

and

$$\mathrm{var}(\hat{\gamma}) = \frac{1}{\sum_{i=1}^{I} w_i^*} \tag{1.15}$$

where

$$\mathrm{w_i} = \frac{1}{S_i^2 + \hat{\tau}^2} \tag{1.16}$$

Each study's weight (w_i^*) now depends on the sum of the two estimated variances: the variance of the study's prognostic effect estimate (S_i^2) and the between-study variance of the prognostic effect (τ^2). Hence, τ^2 must also be estimated in order to derive the summary prognostic effect estimate ($\hat{\gamma}$) when assuming random prognostic effects, and this is best done using restricted maximum likelihood (REML) (Riley et al., 2021b).

1.7.3 Issues with IPD meta-analysis

Despite the benefits of IPD meta-analysis over traditional aggregate data meta-analysis, it still faces some challenges. IPD meta-analysis projects can be incredibly time consuming and costly in terms of obtaining, cleaning and analysing the data, requiring more time and resources than for conventional aggregate data (Macaskill, 2004, Macaskill et al., 2010, Nikoloulopoulos, 2017). Negotiating and maintaining relationships with investigators from different countries, settings and disciplines can also take considerable time and effort (Kuss et al., 2014, Nikoloulopoulos, 2017). Yet despite the extensive efforts to obtain IPD from all identified authors, it may still be unavailable for some studies, leading to an availability bias in the analyses (Ahmed et al., 2012). Similarly, publication bias may also be an issue in identifying suitable studies for inclusion. Issues from primary study deficiencies may also remain, such as differences in outcome definitions, differences in methods of measurement or differences in the predictors available in each study. Another issue that could arise when undertaking an IPD meta-analysis are continuous predictors may have been categorised by the study authors, making it impossible to harmonise the continuous predictors across studies.

1.8 Aims and outline of the thesis

The broad aim of the thesis is to apply and develop statistical methods for prognosis and risk prediction research. In particular, the thesis aims to:

- Apply statistical methods for prognosis and prediction research in novel clinical examples, to provide new findings about prognostic factors and prediction models.
- Investigate the added prognostic value of potential prognostic factors for the development of complications in monochorionic (MC) twin pregnancies, to improve knowledge of complications in MC twin pregnancies.
- Review recent prediction model articles to ascertain the potential for measurement error in the predictors used and how often it was acknowledged and/or accounted for.

- Review recent prediction model articles to establish whether the predictors in the models were generally measured at the same time that the model is intended to be used in practice.
- Illustrate methodological issues for prediction model development when predictors are measured at a different time point to the intended moment of use of the model.
- Externally validate existing prediction models across several population groups for predicting stillbirth, using IPD meta-analysis.
- Propose a method for estimating the power of an IPD meta-analysis project for evaluating potential prognostic factors.

The thesis has seven chapters. Chapters 2 to 4 focus on statistical approaches and issues in prognosis research conducted in a single study. Chapters 5 and 6 focus on the use of IPD for prognosis research studies. An outline of the chapters is given below.

Chapter 2 *Prognostic value of first-trimester ultrasound measurements and serum biomarkers for adverse outcomes in monochorionic twins*

Chapter 2 presents an applied prognostic factor study which investigates the added prognostic value of two ultrasound measurements and three serum biomarkers for the development of complications later in monochorionic twin pregnancies, to showcase key statistical approaches for examining potential prognostic factors. The chapter also aims to highlight how developing a prediction model may not always be appropriate. The work arising from this chapter has been published in Diagnostic and Prognostic Research (Mackie et al., 2019), for which I contributed to the study design, conducted all statistical analyses, contributed to the interpretation and reporting of the results and drafted the methods and results sections of the manuscript.

Chapter 3 *Measurement error and timing of predictor values used in prediction model research: a systematic review of current practice and reporting*

Chapter 3 provides a systematic literature search to identify recent prediction models with the aim of evaluating the potential for measurement error of values of included predictors within the models, and to observe if and how authors accounted for any measurement error. The review also examined whether the timing of the predictor measurements was clearly stated, and if so, its relation to the intended moment of use of the prediction model. The work arising from this chapter has been published in Journal of Clinical Epidemiology (Whittle et al., 2018), for which I led the completion of the review and drafted the initial manuscript. This work has also been presented at the 38th International Society of Clinical Biostatics Conference (ISCB).

Chapter 4 *The effect of measuring time-varying predictors at a different time point to that of the intended moment of use: an illustrative example*

Chapter 4 illustrates the effect that measuring a time-varying predictor after the intended moment of using a prediction model has on the predictor-outcome associations and model performance. The direction and magnitude of predictor-outcome associations of a multivariable prediction model were compared under two scenarios: using a time-varying predictor of interest, ascertained by the treating physician at the point of care (i.e. the intended moment of use) and using the same predictor, but ascertained by a self-complete questionnaire mailed several days after the point of care. The work arising from this chapter has been published Diagnostic and Prognostic Research (Whittle et al., 2017), for which I designed the study in collaboration with colleagues, prepared and analysed the data, and drafted the initial manuscript. This work has also been presented at the 37th International Society of Clinical Biostatics Conference (ISCB), and the Young Statisticians Meeting 2016.

Chapter 5 *External validation of prediction models for stillbirth using individual participant data (IPD) meta-analysis: the IPPIC study*

Chapter 5 utilises individual participant data from multiple studies (the IPPIC study) to externally validate existing prediction models that have been developed across several population groups for predicting stillbirth. The predictive performance of three previously identified prediction models are assessed and compared using discrimination and calibration statistics. Decision curve analysis is used to assess the clinical utility of the prediction models, and the model performance is pooled and summarised across data sets using a two-stage IPD meta-analysis. The results arising from this chapter have been published in Ultrasound in Obstetrics & Gynaecology (Allotey et al., 2022) for which I am a joint co-author having conducted all statistical analyses, contributed to the interpretation and reporting of results and drafted the methods and results sections of the manuscript.

Chapter 6 *Calculating the power to examine prognostic factor effects when planning an individual participant data meta-analysis with a binary outcome*

Chapter 6 derives a method to estimate the power of a planned IPD meta-analysis project, in advance of collecting the IPD, for a project which aims to synthesise the IPD to examine the effect of a (potential) prognostic factor on a binary outcome. The chapter modifies previously published methods that calculated the power of an IPD project to identify a treatment-covariate interaction. Extensions are provided for adjusting the power for the presence of other correlated adjustment factors and for allowing for heterogeneity between studies. The work arising from this chapter is currently being written up for submission to Research Synthesis Methods.

Chapter 7 *Discussion*

Chapter 7 contains an overview of the principal findings from the thesis, a discussion of the strengths and weaknesses of the work completed, implications for future studies developing prediction models and carrying out prognostic factor studies and recommendations for further research.

2 First-trimester ultrasound measurements and maternal serum biomarkers as prognostic factors in monochorionic twins

2.1 Introduction

This chapter presents a prognostic factor study in a real clinical dataset, to showcase key statistical approaches for examining potential prognostic factors. Logistic regression models are fitted to assess the association between potential prognostic factors and adverse outcomes in monochorionic twin pregnancies, after adjusting for the effect of previously identified prognostic factors. This chapter also highlights how it might not always be sensible to develop a prediction model, even when the C-statistic from fitting a multivariable model is apparently promising. The clinical findings of this chapter were published in Diagnostic and Prognostic Research (Mackie et al., 2019), for which I undertook all aspects of statistical analysis and the interpretation and reporting of results. The findings of the work also contributed to the rationale for a paper showcasing the issues of instability of prediction models in small datasets, for which I am a co-author (Riley et al., 2021a).

2.2 Medical terms glossary

A glossary of the medical terms relating to pregnancy used throughout this chapter which may not be commonly known are given below.

Aneuploidy screening Screening tests that are conducted during the 12-week ultrasound to give information on the baby’s risk of certain chromosome disorders (such as Down Syndrome).

Concordant Twins inheriting the same genetic characteristic are known as concordant.

Congenital abnormalities Structural or functional anomalies that occur during intrauterine life.

Crown-rump length The measurement of the length of the fetus from the top of the head (crown) to the bottom of the buttocks (rump).

Dichorionic twins A form of multiple gestation in which each twin has a separate placenta (blood supply) and amniotic sac.

Fetal biometry A measurement taken during a standard ultrasound.

Fetal growth restriction Babies that are smaller and lighter than they should be for the number of weeks of pregnancy.

Fetal/amniotic sac A thin-walled sac that surrounds the fetus during pregnancy.

Fetoplacental doppler assessment An assessment of the blood flow going to the baby and within its cord, heart, and brain.

Monochorionic diamniotic twins The product of a single fertilized egg, resulting in genetically identical offspring, each with their own amniotic sac.

Monochorionic monoamniotic twins identical twins that not only share a placenta, but also share the same amniotic sac.

Nuchal translucency	The amount of fluid behind a baby's neck in the first trimester of pregnancy measured during an ultrasound.
Parity	The number of times a woman has previously given birth.
TAPS	A rare condition that occurs when there are unequal blood counts between the twins in the womb, meaning one twin is not receiving the appropriate amount of oxygen and nutrients it needs to develop properly.
Twin-twin transfusion syndrome	A prenatal condition in which twins share unequal amounts of the placenta's blood supply resulting in the two fetuses growing at different rates.

2.3 Clinical rationale for this chapter

Multiple pregnancies are at an increased risk of adverse outcomes, with monochorionic diamniotic (MCDA) twins being at higher risk of pregnancy loss and morbidity compared to dichorionic twins (Hack et al., 2007). This is due to MCDA twins sharing a single placenta. In 10-15% of MCDA twin pregnancies, twin-twin transfusion syndrome (TTTS) occurs because of unbalanced anastomoses (Moldenhauer and Johnson, 2015), which subsequently increases the risk of fetal growth restriction (FGR) in either one or both fetuses (Oepkes and Sueters, 2017). International guidelines recommend intensive antenatal surveillance to detect adverse outcomes complicating monochorionic (MC) twins, principally TTTS and FGR. This involves regular monitoring via ultrasound scans from 16 weeks gestation at two weekly intervals to evaluate the liquor volume in each fetal sac, fetal biometry and often fetoplacental Doppler assessment (Khalil et al., 2016) (ACOG, 2016, Neilson and Kilby, 2008, NICE, 2011). Such obstetric surveillance requires ultrasonographic expertise and health economic resources, it is time-consuming and targets all MC twins as a 'high-risk population'. Additionally, this intensive surveillance may increase maternal anxiety and affect mental health. If it was possible to predict which MC twin pregnancies were at higher risk of developing complications, it would allow clinicians to stratify care, and those at higher risk could undergo more frequent surveillance or be assessed earlier in a tertiary referral centre. This motivates the need for prognostic factor research and the potential development of prediction models in this field, to help identify those at most risk of poor outcomes.

2.4 Objectives

2.4.1 Clinical objectives

The pre-specified primary clinical objective was to assess if there was an association between a pre-determined list of potential prognostic factors and a fetal composite adverse outcome, to improve knowledge of complications of MC twin pregnancies. In particular, the aim was to investigate whether ultrasound measurements and serum biomarkers add prognostic value over and above standard clinical characteristics that are already routinely measured. Secondary objectives were to investigate whether these factors were associated with other, secondary outcomes. The aim here was not to develop a clinical prediction model to use in practice, but to explore the relationship between each potential prognostic factor and the outcome, whilst adjusting for previously identified prognostic factors.

2.4.2 Methodological objectives

Alongside the clinical objectives, for this thesis the statistical objective of the chapter was to highlight how developing a prognostic model with the data at hand may not always be a sensible idea. Clinical collaborators suggested that, following the prognostic factor evaluations, it was important to develop a prognostic model to predict the occurrence of a fetal adverse outcome in women with a MCDA twin pregnancy, including all the prognostic factors of interest and the previously identified prognostic factors. However, a methodological concern was that the sample size to do this was insufficient. To demonstrate this, the instability of developing a prognostic model with the data available is examined, by using bootstrapping to (i) estimate the amount of overfitting and to calculate a shrinkage factor, (ii) estimate the uncertainty in the shrinkage factor, and (iii) assess the lack of stability in model fit and model predictions. Although a multivariable model including all the prognostic factors of interest was fitted, again the aim was not to develop a prediction model to be used in practice, but to emphasize that developing a prediction model may not be appropriate by estimating the amount of instability such a model would have.

2.4.3 Chapter outline

The outline of this chapter is as follows. Section 2.4 begins by defining the data used, the prognostic factors of interest and the outcomes that were considered. This is followed by a description of the statistical methods. Section 2.5 presents the results of the analyses of the prognostic factor study, culminating with the results from fitting a prognostic model and the optimism and instability associated with this model in Section 2.5.3. The chapter concludes with a discussion of the results found in Section 2.6.

2.5 Methods

2.5.1 Participants

An existing multicentre, international cohort of MC twin pregnancies formed by routinely prospectively collected data was provided by colleagues from the University of Birmingham, who also provided the clinical rationale for this project. A protocol was published prior to analysis (Mackie et al., 2017). All women with a MCDA twin pregnancy in the West Midlands and North Thames regions who had undergone first trimester aneuploidy screening between October 2014 and September 2015, or women with a MCDA twin pregnancy who booked at the Royal Prince Alfred Hospital, Sydney between June 2011, and April 2016, and for whom a first trimester blood sample was stored, were eligible for inclusion. Chronicity had to have been determined in the first trimester based on: a single placental mass, a thin inter-twin membrane, and the presence of the 'T' sign, absence of Lambda sign (Sepulveda et al., 1996). If the twins were a different sex, or postnatally the pregnancy was diagnosed as dichorionic based on placental assessment, these pregnancies were excluded. These women were booked at 29 different secondary and tertiary care maternity units (28 in the UK, 1 in Australia), depending on geographical area and were under consultant-led care due to the high-risk nature of multiple pregnancies. Women were not eligible for inclusion if they had a miscarriage prior to 14 weeks gestation, a monochorionic monoamniotic pregnancy, a higher order multiple pregnancy, or the pregnancy was affected by serious structural or congenital anomalies, whether concordant or discordant, as the aetiology of their adverse outcomes, such as growth restriction, preterm birth or intrauterine demise would be different to pregnancies not affected by structural or congenital anomalies, and would therefore increase heterogeneity within the cohort. Where outcome data were missing due to the time-period and setting diversity, women were not contacted for further details, thus these pregnancies were excluded. Pregnancies were cared for according to local and national guidelines. Postnatal outcome data until discharge from hospital were retrospectively collected from hospital notes. No further follow-up data was collected.

2.5.2 Potential prognostic factors

The ultrasound and biomarkers to be examined for their prognostic ability were pre-defined before data collection and analysis, following advice from clinical collaborators at the University of Birmingham.

2.5.2.1 Ultrasound measurements

Nuchal translucency (NT) and crown-rump length (CRL) are measured as standard practice during routine ultrasound scans in women who consent to first trimester aneuploidy screening in the UK or Australia (FASP, 2015) (FMF, 2004). These were performed by sonographers and fetal medicine doctors in the local units who were approved by the Fetal Medicine Foundation to perform these scans. NT discordance (%) was calculated as the smallest NT subtracted from the largest NT, divided by the largest NT, and multiplied by 100. CRL discordance (%) was calculated as per NT discordance. These measurements were treated as continuous variables within analyses (Royston et al., 2006).

2.5.2.2 <u>Biomarker measurements</u>

Three serum biomarkers (AFP, sFlt-1, PlGF) were measured on stored maternal serum samples that were initially analysed for β-hCG and PAPP-A. The β-hCG and PAPP-A are measured as standard practice in women who consent to participate in the UK or Australian aneuploidy screening programme and hence these measurements were not repeated for this study.

Further details on all the potential prognostic factors of interest are given in Table 2.1.

Table 2.1: Potential prognostic factors of interest

Factor	How measured	When measured	Biological role and clinical use	Lower limit of detection	Upper limit of detection
Nuchal translucency	By accredited sonographers as part of the National Aneuploidy screening programme	11^{+2} weeks–14^{+1} weeks, which is calculated based on the CRL which must be between 45-84mm	Measurement of lymphatic fluid at the back of the fetus's neck. High measurement indicates high risk of chromosomal or cardiac problems. Discrepancy between twins may indicate adverse outcome.	0.1cm	Infinite
Crown-rump length	By accredited sonographers as part of the National Aneuploidy screening programme	11^{+2} weeks–14^{+1} weeks, which is calculated based on the CRL which must be between 45-84mm	Measurement of the length of the baby used to date the pregnancy. Discrepancy between twins may indicate adverse outcome.	1mm	Infinite
sFlt-1	Clinical laboratory with experience of performing these assays for pre-eclampsia prediction	Sample taken at 10^{+0}–14^{+1} weeks gestation, stored at -80°C and analysed 6 years 5 months to 0 years 5 months later	Produced by the syncytiotrophoblast and prevents angiogenesis. High level indicates pre-eclampsia in singleton pregnancies prior to the appearance of signs and symptoms.	10pg/mL	85000pg/mL
PlGF	Clinical laboratory with experience of performing these assays for pre-eclampsia prediction	Sample taken at 10^{+0}–14^{+1} weeks gestation, stored at -80°C and analysed 6 years 5 months to 0 years 5 months later	Produced by the syncytiotrophoblast and promotes angiogenesis. Low level indicates pre-eclampsia in singleton pregnancies prior to the appearance of signs and symptoms.	3pg/mL	10000pg/mL
AFP	Clinical laboratory accredited to perform testing as part of National Aneuploidy screening programme	Sample taken at 10^{+0} -14^{+1} weeks gestation, stored at -80°C and analysed 6 years 5 months to 0 years 5 months later	Human function is unknown, but abundant in fetuses. Low level indicates increased aneuploidy risk in twin pregnancies in the 2nd trimester as part of the Quad test.	1U/mL	1000U/mL

2.5.3 Outcomes

The primary outcome for the study was a fetal adverse outcome composite defined as at least one of the following: TTTS, antenatally detected growth restriction, postnatally detected growth restriction, twin anaemia plycythaemia sequence (TAPS), twin oligohydramnios-polyhydramnios (TOPS) or intrauterine fetal death (IUFD).

Individual complications were also examined as secondary outcomes, as well as a neonatal composite outcome and a maternal composite outcome. Secondary outcomes were defined as:

I. TTTS: defined and staged as per Quintero criteria (Quintero et al., 1999). Pregnancies affected by TTTS with concurrent growth restriction were not included in the antenatal or postnatally detected growth restriction groups

II. Antenatally detected fetal growth restriction: abdominal circumference (AC) or estimated fetal weight (EFW) <10th centile in either/both fetus(es) and/or growth discordance >20% recorded at least twice over ≥ 2-week period

III. Postnatally detected growth restriction: birthweight <9th centile on the World Health Organization Growth Charts (RCPCH, 2016)

IV. IUFD: sub-classified as either single IUFD (sIUFD) or double IUFD (dIUFD). The pregnancy was considered a miscarriage if the pregnancy loss occurred at 14-24weeks and a stillbirth if ≥24 weeks.

V. Spontaneous preterm birth (PTB): between 24 and 34 weeks gestation. Iatrogenic PTB delivery was not included.

VI. Neonatal composite outcome: neonatal death, respiratory distress syndrome, assisted ventilation (Continuous positive airway pressure [CPAP] or endotracheal [ET] tube) for

>24 hours, intraventricular haemorrhage or other brain injury, necrotising enterocolitis, neonatal encephalopathy, chronic lung disease, severe jaundice requiring phototherapy, severe infection e.g. septicaemia, meningitis, exchange transfusion, cardiac impairment, neurological impairment.

VII. Maternal composite outcome: gestational diabetes mellitus, severe infection, hypertensive disorders (pregnancy induced hypertension requiring medication, pre-eclampsia, eclampsia or HELLP syndrome), placental abruption, venous thromboembolism, disseminated intravascular coagulopathy, High-Dependency or Intensive Care Unit admission, cerebrovascular event, renal or liver failure, pulmonary oedema, massive obstetric haemorrhage (>2L EBL), acute fatty liver.

2.5.4 Existing prognostic factors

The aim was to examine the *added* prognostic value of the ultrasound and biomarker variables, and therefore the multivariable analyses adjusted for the prognostic effect of standard clinical information considered (by the clinical collaborators) to be existing prognostic factors: maternal BMI, age, smoking status, ethnicity, parity, and mode of conception (Table 2.2). The neonatal outcome was also adjusted for gestational age at delivery and steroid and antenatal magnesium sulphate administration. The existing prognostic factors were forced into the multivariable model (i.e. no variable selection was used), as is recommended (Riley et al., 2019b).

Table 2.2: Existing prognostic factors

Prognostic factor	How measured	When measured	Categorical or continuous data	Outcomes it may affect
Maternal body mass index (BMI)	Height and weight measured and calculated according to: kg/m^2	Booking	Continuous	All
Maternal age	From date of birth	Booking	Continuous	All
Maternal smoking status	Maternal reporting as documented in notes	Booking	Categorical: Never smoked (ref. group) Current smoker Ex-smoker	All
Maternal ethnicity	Maternal reporting as documented in notes	Booking	Categorical: White (ref. group) Other/mixed Oriental South Asian African-Caribbean	All
Parity	Maternal reporting as documented in notes	Booking	Categorical: Nulliparous(ref. group) Multiparous 1 Multiparous 2+	All
Mode of conception	As documented in the notes	Booking	Categorical: Natural conception (ref. group) Assisted conception	All
Gestational age at delivery	As documented in the notes	Delivery	Continuous	Neonatal composite
Steroid administration	As documented in the notes	Throughout pregnancy	Categorical: Yes steroids No steroids (ref. group)	Neonatal composite
Magnesium Sulphate (MgSO4) administration	As documented in the notes	Throughout pregnancy	Categorical: Yes MgSO4 No MgSO4 (ref. group)	Neonatal composite; Spontaneous preterm birth

Other existing prognostic factors that were unable to be adjusted for were social deprivation (due to not having an adequate measure of this) and tocolysis (as insufficient data were present to adjust for this).

2.5.5 Missing data

Multiple imputation was performed to replace missing predictor (prognostic factor) values using a chained equation approach, with predictive mean matching for continuous variables, based on maternal age, BMI, log(AFP), log(sFlt-1), log(PIGF), ethnicity (categorised as white vs. non-white), assisted conception, steroid use, magnesium sulphate (MgSO4), smoking status, country of antenatal care, parity, gestation at delivery, NT discordance (%), CRL discordance (%) and fetal adverse outcome composite. Ten imputed datasets were created for missing maternal BMI and smoking status that were then combined across all datasets using Rubin's rule to obtain final model estimates. An old rule of thumb regarding an appropriate number of imputations was that 3 to 10 imputations would typically suffice (Rubin, 1987). Hence, in this work, 10 imputed datasets were created following this rule of thumb to err on the side of caution. However, it has been shown that this advice only ensured the precision and replicability of the point estimates, but not the estimates of the standard error, and White et al. (2011) suggest that the number of imputations should be greater than or equal to the percentage of missing observations in order to ensure an adequate level of reproducibility, which has since gained in popularity and is now the accepted rule of thumb. The work in this chapter was carried out towards the beginning of my PhD programme and therefore, if I was carrying out this study again, I would now be following this advice and creating at least 19 imputed datasets (the greatest percentage of missing data that is being imputed).

The neonatal outcome contained 14 missing observations, but imputation was not considered necessary, as the missing outcome was deemed plausibly missing at random conditional on the set of factors included in the multivariable model. Hence, these pregnancies were not included in the neonatal composite outcome but were included in the other outcomes.

2.5.6 Statistical analysis

Maternal and fetal characteristics at baseline were summarised using frequencies and percentages for categorical variables and mean and standard deviation (SD) or median and inter-quartile range (IQR) for continuous variables (depending on the normality of the variable). Univariable logistic regression models were fitted individually to each of the five predictors of interest (NT, CRL, AFP, sFlt-1 and PIGF) to assess their unadjusted association (odds ratio) with the primary outcome (pregnancy composite outcome) and with each of the 10 secondary outcomes (neonatal composite outcome, maternal antenatal and postnatal composite outcome, antenatally detected growth restriction (per pregnancy), antenatally detected growth restriction (per twin), postnatally detected growth restriction (per pregnancy), postnatally detected growth restriction (per twin), TTTS, single IUFD, double IUFD and spontaneous preterm birth).

Multivariable logistic regression models were fitted, with random effects as necessary, to examine the independent prognostic value of each of the five factors separately, with adjustment for standard characteristics already deemed likely to be prognostic of adverse outcome: maternal BMI, maternal age, maternal smoking status, maternal ethnicity, parity, mode of conception. The spontaneous preterm birth outcome was additionally adjusted for administration of MgSO4. The neonatal composite outcome was also adjusted for gestation at delivery, administration of steroids, and MgSO4. For both single and double IUFD, it was not possible to adjust for all the previously identified prognostic factors due to a lack of convergence, hence, these models were both adjusted for maternal age, BMI, ethnicity, and mode of conception only.

For outcomes relating to individual babies (rather than per pregnancy), a random intercept term at the level of the mother was included in the logistic regression models where it was sensibly estimable, to account for clustering of multiple babies per women. Only the association with baby specific prognostic factors, NT and CRL, were estimated for these outcomes. The individual baby's value of NT and CRL were used for these outcomes in place of the discordance.

Clustering by hospital was also adjusted for by putting a random effect on the intercept which allowed for heterogeneity in baseline risk across hospitals.

The three serum biomarkers (AFP, PlGF and sFlt-1) were log transformed as they were highly skewed, and all continuous prognostic factors were included in the models as a linear term. All potential prognostic factor measurements remained as continuous variables during analysis, and no cut-offs were applied.

As part of a sensitivity analysis, fractional polynomials were used to assess for the possibility of a non-linear relationship between each of the prognostic factors and the primary outcome (Royston et al., 1999) (using the *fp* command in Stata). Fractional powers (-2, -1, -0.5, 0, 0.5, 1, 2, 3) of NT discordance, CRL discordance, log(AFP), log(PlGF) and log(sFlt-1) were considered individually, adjusting for the previously identified prognostic factors. For simplicity, fractional polynomials were considered in the complete case data only.

To gauge the potential increase in discrimination performance of a prognostic model that includes each potential prognostic factor in addition to existing factors, the change in apparent C-statistic (increase in the area under the curve) for each outcome was calculated (i.e., the difference in apparent C-statistic for models with standard characteristic including or excluding each factor). No adjustment for potential model overfitting was made during the calculation, as this was only for illustration of the potential impact of including the factors.

The Akaike's Information Criterion (AIC) was not considered here (as in Chapter 4), as the AIC is generally used to compare the fit of multiple models and the aim in this chapter was to investigate the individual relationships of the potential prognostic factors rather than compare the models.

2.5.7 Sample size

The dataset used for the analyses in this chapter was an existing cohort of fixed size, and hence no sample size calculation was completed. However, due to the rarity of MCDA twins, this represents the largest available data to be used for assessing prognostic factors. Retrospectively, a power calculation was conducted to estimate the size of odds ratio that the study would have 80% power to detect. For the primary outcome, composite fetal adverse outcome, there was 80% power to detect an odds ratio of 1.034 for the association with the percentage discordance in CRL. For log(sFlt-1) there was an 80% power to detect an odds ratio of 3.49.

2.5.8 Developing a prognostic model and examining instability

To examine the instability of developing a prognostic model in this available dataset, a multivariable logistic regression model was fitted to the primary outcome (including all the predictors of interest and all adjustment factors), followed by a bootstrap investigation. No variable selection process was used, and apparent model performance was summarised using the C-statistic (discrimination) and the calibration slope (calibration). Then, bootstrapping was used to assess the amount of overfitting and the uncertainty in the estimation of shrinkage needed. One hundred random bootstrap samples were used, with replacement, from the original data (before imputation) and the imputation procedure was applied to each bootstrapped dataset. The same model as above was then fitted to each imputed bootstrapped datasets, and within each bootstrap, imputed results were again combined using Rubin's rules. In each bootstrapped dataset, the apparent performance was calculated by calculating the C-statistic and the calibration slope. The test performance of each these models was also calculated by fitting the bootstrapped model in the original imputed dataset (i.e. estimating the predicted probabilities in the original imputed data using the coefficients of the model from the bootstrapped data). The optimism of each model was then calculated as the mean value of the difference between the apparent and test performance measures in each of the bootstrapped models. The shrinkage factor could then be estimated as the optimism in the calibration slope subtracted from the mean apparent performance measure. The estimated shrinkage factor was applied uniformly to the coefficients of the original model and an updated intercept was calculated using the new coefficients after shrinkage. The uncertainty in the shrinkage factor was also summarised by the 95% interval of values based on the bootstrap procedure. An optimism adjusted C-statistic was also calculated.

2.5.9 Sensitivity analysis: Firth's correction

The use of Firth's correction for dealing with sparse events has been gaining recommendations for prognostic factor research, specifically to address potential upward bias in odds ratios. Hence, the analyses for the primary outcome were repeated using Firth's correction, which is a penalisation method aiming for unbiased estimates of odds ratios for each predictor (whereas, the uniform shrinkage is aimed at calibration of the prognostic model as a whole).

2.6 Results

2.6.1 Summary of data

A total of 177 pregnancies (354 babies) were included in the study. The maternal and fetal characteristics are summarised in Table 2.3. The average age of the mother was 30.4 years (SD 5.4) and had an average BMI of 24.9 (SD 5.4). Most of the mothers had never smoked (77.4%), were white (64.7%) and were first time mothers (60.5%). The median gestational age at delivery was 35.4 weeks (IQR 33.0, 36.6). The numbers of missing values for each of the baseline characteristics and adjustment factors are also given in Table 2.3. There was no or minimal missing data in most of the factors, except for BMI which had 19% missing.

The number (%) of each of the outcome events that occurred are given in Table 2.4. Values are given per pregnancy unless otherwise stated. There were 55 (31.1%) participants who did not experience an adverse outcome event and delivered two healthy babies after 34 weeks gestation. The primary outcome, fetal adverse outcome composite occurred in 94 (53.1%) of the included pregnancies.

Table 2.3: Maternal and fetal characteristics

	Total cohort (n=177, 354 babies)	**Missing; N (%)**
Maternal age; mean (SD) years	30.38 (5.43)	0
Maternal BMI; mean (SD) kg/m^2	24.87 (5.41)	16 (19.04)
Maternal Smoking status; n (%)		13 (7.34)
Never smoked	127 (77.44)	
Current smoker	12 (7.32)	
Ex-smoker	25 (15.24)	
Maternal ethnicity; n (%)		4 (2.26)
White	112 (64.74)	
Mixed	10 (5.78)	
Oriental	22 (12.72)	
South Asian	19 (10.98)	
African-Caribbean	10 (5.78)	
Parity; n (%)		0
0	107 (60.45)	
1	48 (27.12)	
2	18 (10.17)	
3	2 (1.13)	
4	2 (1.13)	
Assisted conception; n (%)	24 (13.95)	5 (2.82)
Gestational age at delivery; median (IQR) weeks	35.43 (33.00, 36.57)	0
Steroid administration; n (%)	125 (71.02)	1 (0.56)
Magnesium sulphate administration; n (%)	12 (6.82)	1 (0.56)
Nuchal translucency (NT); median (IQR) % discordance	11.76 (5.55, 21.15)	0
Crown-rump length (CRL); median (IQR) % discordance	4.22 (1.75, 7.02)	0
AFP; median (IQR) U/mL	29.29 (23.40, 41.50)	1 (0.56)
sFlt-1; median (IQR) pg/mL	2163 (1645, 2945.5)	1 (0.56)
PlGF; median (IQR) pg/mL	60.45 (40.89, 89.02)	1 (0.56)

Table 2.4: Outcomes

Outcome	N (%)
Uncomplicated monochorionic diamniotic twin pregnancy, delivered >34 weeks gestation	55 (31.07)
Fetal adverse outcome composite	94 (53.11)
Twin-twin transfusion syndrome	23 (12.99)
Antenatal growth restriction	41 (23.16)
Antenatal growth restriction (per fetus)	73 (20.6)
Postnatal growth restriction	43 (24.29)
Postnatal growth restriction (per baby)	54 (15.25)
Intrauterine fetal death (single)	11 (6.21)
Intrauterine fetal death (double)	12 (6.78)
Spontaneous preterm birth	12 (6.78)
Maternal antenatal and postnatal composite	46 (25.99)
Neonatal composite	91 (26.76)

2.6.2 Primary outcome: Fetal composite adverse outcome

Table 2.5 gives the unadjusted and adjusted odds ratios (OR) and 95% confidence intervals (CI) for the association between each of the prognostic factors and the primary outcome, fetal composite adverse outcome. The adjusted results show the added value of the prognostic factors over and above standard previously identified factors, along with the added discrimination of adding each particular prognostic factor of interest to the model with all of the standard adjustments factors included.

There was evidence of an unadjusted association between the pregnancy composite outcome and both the percentage discordance between the baby's nuchal translucency (OR 1.03 95% CI 1.01, 1.05) and percentage discordance between the baby's crown-rump lengths (OR 1.16 95% CI 1.06, 1.27). These associations both remained after adjustment with an estimated 3% (aOR 1.03 95%CI 1.01, 1.06) increase in the odds of an unfavourable outcome for each 1% increase in NT discordance, and an estimated 17% (aOR 1.17 95% CI 1.07, 1.29) increase in the odds of an unfavourable outcome for each 1% increase in CRL discordance. The baseline C-statistic for the model including only standard characteristics was 0.59, and the increase in the C-statistic when each prognostic factor was additionally included was quite high with the NT discordance increasing the C-statistic by 0.045 and the addition of CRL discordance increasing the C-statistic by 0.103. There was no evidence of any associations between the serum biomarkers and the primary outcome, though confidence intervals were wide and so firm conclusions about prognostic value are not possible.

Table 2.5: Fetal adverse outcome composite - Unadjusted and adjusted results

	Unadjusted prognostic effect	**Adjusted* prognostic effect**	**Change in C-statistic**
	OR (95% CI)	***aOR (95% CI)***	
NT (% discordance)	1.03 (1.01, 1.05)	1.03 (1.01, 1.06)	0.045
CRL (% discordance)	1.16 (1.06, 1.27)	1.17 (1.07, 1.29)	0.103
log(AFP)	1.91 (0.93, 3.94)	2.08 (0.94, 4.59)	0.026
log(sFlt-1)	1.12 (0.52, 2.40)	1.03 (0.42, 2.50)	<0.001
log(PlGF)	0.73 (0.44, 1.22)	0.65 (0.37, 1.13)	0.014

*Adjusted for maternal BMI, maternal age, maternal smoking status, maternal ethnicity, parity, and mode of conception

2.6.2.1 Sensitivity analysis: Firth's correction

Table 2.6 gives the adjusted odds ratios (95% CIs) for the relationships between each potential prognostic factor and the primary outcome, after using Firth's correction. Previously identified relationships remained, albeit with the odds ratios being marginally shrunk in comparison to before using Firth's correction. The attenuation in the estimates was greater for the prognostic factors that had a larger odds ratio to begin with, i.e. for log(AFP) the odds ratio reduced from 2.08 before Firth's correction to 1.96 after. However, the conclusions of the analysis did not change.

Table 2.6: Adjusted results for the primary outcome (fetal adverse outcome composite) after using Firth's correction

	Before Firth's correction		**After Firth's correction**
	Unadjusted prognostic effect	**Adjusted* prognostic effect**	**Adjusted* prognostic effect**
	OR (95% CI)	***aOR (95% CI)***	***aOR (95% CI)***
NT (% discordance)	1.03 (1.01, 1.05)	1.03 (1.01, 1.06)	1.03 (1.01, 1.05)
CRL (% discordance)	1.16 (1.06, 1.27)	1.17 (1.07, 1.29)	1.16 (1.06, 1.27)
log(AFP)	1.91 (0.93, 3.94)	2.08 (0.94, 4.59)	1.96 (0.92, 4.23)
log(sFlt-1)	1.12 (0.52, 2.40)	1.03 (0.42, 2.50)	1.03 (0.44, 2.43)
log(PlGF)	0.73 (0.44, 1.22)	0.65 (0.37, 1.13)	0.67 (0.39, 1.14)

*Adjusted for maternal BMI, maternal age, maternal smoking status, maternal ethnicity, parity, and mode of conception

2.6.2.2 Sensitivity analysis: Fractional polynomials

When accounting for a potential non-linear relationship between each of the prognostic factors and a fetal adverse outcome, the best fitting relationships (defined by the deviance) were: a squared relationship for NT discordance, a linear relationship for CRL discordance, a cubic relationship for log(afP) and a fractional power of -2 for both log(PlGF) and log(sFlt). Hence, some potential non-linear relationships were identified, nevertheless, the differences in the AIC were very small for all the models that allowed for non-linear associations compared to the models assuming a linear association, with the greatest difference in AIC being 0.958 (for log(afp)). Hence, further research, with larger sample sizes, is needed to examine non-linearity for these continuous factors.

2.6.3 Illustration of the issues of developing a prognostic model with this dataset

The results of the multivariable model for the primary outcome, fetal adverse outcome composite, are presented in Table 2.7. The model included all the potential prognostic factors of interest from the prognostic factor study (NT % discordance, CRL % discordance, log(AFP), log(sFlt-1) and log(PlGF)), and all the previously identified existing prognostic factors from Section 2.4.4 (age, BMI, smoking status, ethnicity, parity, and assisted conception). The C-statistic from the multivariable model was 0.741, which appears to demonstrate a reasonably well discriminating model. However, using bootstrapping to estimate and then adjust for the amount of optimism in the model, gave an optimism-adjusted C-statistic of 0.670, which is substantially smaller, and indicates that there was a large amount of optimism due to overfitting.

After bootstrapping, the shrinkage factor was calculated to be 0.654 (95% CI 0.387, 0.921). Therefore, the best estimate is that the overfitting needs to be corrected by shrinking the beta coefficients (log odds ratios) in the multivariable model by 34.6%. This is a large shrinkage factor, indeed guidance that came out around the same time as this work suggests aiming for a shrinkage of 0.9 or above (Riley et al., 2019a). The coefficients of the multivariable model after applying the shrinkage factor are also given in Table 2.7, which are shown to be attenuated.

However, the use of a shrinkage factor does not resolve the issue of small sample size for model development. There is still instability in the shrunken prognostic model, as highlighted by the uncertainty in the shrinkage factor. The mean is 0.654, but the 95% range from bootstrapping is 0.387 to 0.921, suggesting there could be a lot of error in the shrinkage factor applied, and thus even after adjusting for overfitting, the model is unlikely to be robust for practice, as we cannot be confident what the actual predictor effects should be.

Due to the time-consuming nature of imputing data within bootstraps, a pragmatic approach of using 100 random bootstrap samples was taken to examine the instability. However, as a sensitivity analysis, the analysis was re-run using 500 bootstrap samples, which produced an optimism adjusted c-statistic of 0.665 and a shrinkage factor of 0.628 (95% CI 0.381, 0.876), so overall very similar to when using 100 bootstrap samples.

Table 2.7: Coefficients from prognostic model for risk of a fetal adverse outcome before and after shrinkage

		Before shrinkage	**After shrinkage**	
		Coefficient	Coefficient	Odds Ratio
Predictor	*Scale*			
NT	% discordance	0.028 (-0.001, 0.057)	0.0182	1.018
CRL	% discordance	0.188 (0.077, 0.300)	0.123	1.131
AFP	log	0.891 (-0.022, 1.805)	0.583	1.792
sFlt-1	log	-0.203 (-1.225, 0.820)	-0.133	0.876
PlGF	log	-0.292 (-0.903, 0.319)	-0.191	0.826
Age	years	0.042 (-0.028, 0.112)	0.027	1.028
BMI	kg/m^2	0.017 (-0.060, 0.094)	0.011	1.011
Smoking status	1=current	0.659 (-0.795, 2.112)	0.431	1.539
Ethnicity	1=non-white	-0.026 (-0.793, 0.742)	-0.017	0.983
Parity	1=nulliparous	0.662 (-0.094, 1.419)	0.433	1.542
Assisted conception	1=yes	-0.084 (-1.189, 1.022)	-0.055	0.947
Constant		*-3.505 (-12.934, 5.924)*	*-4.481*	

2.6.4 Secondary outcomes

2.6.4.1 Twin-twin transfusion syndrome (TTTS)

Unadjusted and adjusted odds ratios (95% CIs) for the association between each prognostic factor and TTTS, and the increase in C-statistic after adding each factor to the baseline model, are given in Table 2.8. There was evidence of an association between percentage discordance between the babies nuchal translucency and TTTS, with an estimate 6% increase (aOR 1.06 95%CI 1.03, 1.10) in the odds of TTTS for each percent increase in NT discordance. The log transformed AFP was associated with TTTS both when the model was unadjusted and when adjusted (aOR 3.24 (1.00, 10.48)), however the confidence interval is very wide, such that the actual magnitude of prognostic effect is uncertain. The log transformed PlGF was associated with TTTS, with an estimated 58% decrease (aOR 0.42 95% CI: 0.19, 0.93) in the odds of TTTS for each unit increase in log(PlGF), though again with a wide confidence interval. The baseline C-statistic (for the model with only standard prognostic characteristics) was 0.617, and the increase in the C-statistic when each prognostic factor was included was quite high for NT (0.137), AFP (0.067) and PlGF (0.074).

Table 2.8: Twin to twin transfusion syndrome (TTTS)

	Unadjusted prognostic effect	**Adjusted* prognostic effect**	**Change in C-statistic**
	OR (95% CI)	***aOR (95% CI)***	
NT (% discordance)	1.05 (1.02, 1.08)	1.06 (1.03, 1.10)	0.137
CRL (% discordance)	1.07 (0.96, 1.20)	1.09 (0.97, 1.23)	0.032
log(AFP)	3.04 (1.05, 8.78)	3.24 (1.00, 10.48)	0.067
log(sFlt-1)	1.91 (0.62, 5.88)	1.64 (0.44, 6.03)	0.006
log(PlGF)	0.43 (0.20, 0.91)	0.42 (0.19, 0.93)	0.074

*Adjusted for maternal BMI, maternal age, maternal smoking status, maternal ethnicity, parity, and mode of conception.

2.6.4.2 Antenatally detected growth restriction

Unadjusted and adjusted odds ratios (95% CIs) for the association between each prognostic factor and antenatally detected growth restriction, per pregnancy, along with change in C-statistic from adding each factor to a baseline model with just the standard characteristics, are given in Table 2.9. The results are given for antenatally detected growth restriction, per fetus in Table 2.9. There was evidence for an unadjusted association between the percentage discordance in CRL and antenatal growth restriction within the pregnancy, and this association remained after adjustment for the previously identified predictors, with an estimated 20% (aOR: 1.20 95% CI: 1.08, 1.34) increase in the odds of antenatal growth restriction for each 1% increase in CRL discordance. The increase in the C-statistic when adding CRL discordance was fairly large, at 0.119 (baseline C-statistic 0.616). There was no clear evidence of a relationship between the individual babies CRL and antenatal growth restriction in the individual baby (aOR 1.11 95% CI 0.97, 1.26).

Table 2.9: Antenatally detected growth restriction (per pregnancy)

	Unadjusted prognostic effect	**Adjusted* prognostic effect**	**Change in C-statistic**
	OR (95% CI)	***OR (95% CI)***	
NT (% discordance)	1.01 (0.99, 1.03)	1.01 (0.99, 1.04)	0.014
CRL (% discordance)	1.17 (1.06, 1.30)	1.20 (1.08, 1.34)	0.119
log(AFP)	1.55 (0.67, 3.55)	2.10 (0.82, 5.40)	0.032
log(sFlt-1)	1.25 (0.50, 3.13)	1.47 (0.49, 4.35)	0.011
log(PlGF)	0.91 (0.50, 1.66)	0.88 (0.44, 1.76)	-0001

*Adjusted for maternal BMI, maternal age, maternal smoking status, maternal ethnicity, parity, and mode of conception.

Table 2.10: Antenatal growth restriction (per fetus)

	Unadjusted prognostic effect	**Adjusted* prognostic effect**	**Change in C-statistic**
	OR (95% CI)	***OR (95% CI)***	
NT	0.62 (0.15, 2.56)	0.67 (0.13, 3.35)	0.007
CRL	1.12 (0.99, 1.27)	1.11 (0.97, 1.26)	0.004

*Adjusted for maternal BMI, maternal age, maternal smoking status, maternal ethnicity, parity, and mode of conception.

2.6.4.3 Postnatally detected growth restriction

Unadjusted and adjusted odds ratios (95% CIs) for the association between each prognostic factor and postnatally detected growth restriction, and change in C-statistics, are given in Table 2.11 for the outcome per pregnancy and Table 2.11 for the outcome per baby. There was no clear evidence of associations between any of the prognostic factors and postnatally detected growth restriction.

Table 2.11: Postnatally detected growth restriction (per pregnancy)

	Unadjusted prognostic effect	**Adjusted* prognostic effect**	**Change in C-statistic**
	OR (95% CI)	***OR (95% CI)***	
NT (% discordance)	1.01 (0.98, 1.03)	1.01 (0.98, 1.03)	<0.001
CRL (% discordance)	1.05 (0.95, 1.15)	1.04 (0.94, 1.15)	0.012
log(AFP)	0.88 (0.39, 1.98)	0.88 (0.35, 2.20)	<-0.001
log(sFlt-1)	0.63 (0.25, 1.59)	0.65 (0.22, 1.88)	0.006
log(PlGF)	1.55 (0.84, 2.85)	1.62 (0.80, 3.29)	0.017

*Adjusted for maternal BMI, maternal age, maternal smoking status, maternal ethnicity, parity, and mode of conception.

Table 2.12: Postnatal growth restriction (per baby)

	Unadjusted prognostic effect	**Adjusted* prognostic effect**	**Change in C-statistic**
	OR (95% CI)	***OR (95% CI)***	
NT	0.81 (0.39, 1.68)	0.75 (0.36, 1.55)	0.003
CRL	0.96 (0.90, 1.01)	0.97 (0.91, 1.02)	0.002

*Adjusted for maternal BMI, maternal age, maternal smoking status, maternal ethnicity, parity, and mode of conception.

2.6.4.4 Intrauterine fetal death

The results for the associations between each prognostic factor and intrauterine fetal death are given in Table 2.13 for single IUFD and in Table 2.14 for double IUFD. The clustering of babies within mothers for IUFD was not able to be accounted for using a random effects model due to small numbers and minimal variation within the clusters which created convergence issues; hence, a fixed-effect model was used. There was evidence for an unadjusted and adjusted association between log(PlGF) and both single (aOR 0.34 95%CI 0.12, 0.98) and double intrauterine fetal death (aOR 0.18 95%CI 0.05, 0.58). Discordance in CRL was also found to be associated with single intrauterine death (aOR 1.19 95% CI: 1.01, 1.40). The baseline C-statistic was 0.625 and 0.783 for single IUFD and double IUFD, respectively.

Table 2.13: Single intrauterine fetal death

	Unadjusted prognostic effect	**Adjusted* prognostic effect**	**Change in C-statistic**
	OR (95% CI)	***OR (95% CI)***	
NT (% discordance)	1.01 (0.98, 1.05)	1.02 (0.98, 1.06)	<-0.001
CRL (% discordance)	1.17 (1.00, 1.36)	1.19 (1.01, 1.40)	0.085
log(AFP)	0.68 (0.16, 2.87)	0.80 (0.17, 3.90)	-0.004
log(sFlt-1)	1.27 (0.27, 6.04)	1.79 (0.30, 10.64)	<-0.001
log(PlGF)	0.35 (0.13, 0.97)	0.34 (0.12, 0.98)	0.057

*Adjusted for maternal BMI, maternal age, maternal ethnicity, and mode of conception.

Table 2.14: Double intrauterine fetal death

	Unadjusted prognostic effect	Adjusted* prognostic effect	Change in C-statistic
	OR (95% CI)	***OR (95% CI)***	
NT (% discordance)	1.02 (0.98, 1.06)	1.02 (0.98, 1.06)	-0.005
CRL (% discordance)	1.06 (0.91, 1.24)	1.12 (0.94, 1.33)	0.019
log(AFP)	1.33 (0.34, 5.21)	0.97 (0.18, 5.33)	<-0.001
log(sFlt-1)	4.13 (0.92, 18.58)	8.21 (1.02, 66.24)	0.035
log(PlGF)	0.23 (0.08, 0.63)	0.18 (0.05, 0.58)	0.080

* Adjusted for maternal BMI, maternal age, maternal ethnicity, and mode of conception.

2.6.4.5 Maternal antenatal and postnatal composite

The unadjusted and adjusted odds ratios (95% CIs) for the associations between each prognostic factor and the maternal composite outcome are given in Table 2.15. No clear evidence of associations were found. The baseline C-statistic for the model including only the standard characteristics was 0.728.

Table 2.15: Maternal composite outcome

	Unadjusted prognostic effect	Adjusted* prognostic effect	Change in C-statistic
	OR (95% CI)	***OR (95% CI)***	
NT (% discordance)	1.01 (0.99, 1.03)	1.01 (0.98, 1.03)	0.001
CRL (% discordance)	0.96 (0.87, 1.06)	0.97 (0.87, 1.07)	-0.001
log(AFP)	0.56 (0.25, 1.26)	0.55 (0.21, 1.42)	0.010
log(sFlt-1)	1.36 (0.57, 3.27)	1.26 (0.44, 3.58)	-0.004
log(PlGF)	0.78 (0.44, 1.39)	0.70 (0.34, 1.42)	0.004

*Adjusted for maternal BMI, maternal age, maternal smoking status, maternal ethnicity, parity, and mode of conception.

2.6.4.6 Spontaneous preterm birth

No clear evidence of associations were found between any of the prognostic factors and spontaneous preterm birth (Table 2.16). The baseline C-statistic for the model including only the standard characteristics was 0.676.

Table 2.16: Spontaneous preterm birth

	Unadjusted prognostic effect	**Adjusted* prognostic effect**	**Change in C-statistic**
	OR (95% CI)	***OR (95% CI)***	
NT (% discordance)	1.00 (0.64, 1.04)	0.99 (0.95, 1.04)	-0.013
CRL (% discordance)	0.93 (0.77, 1.11)	0.92 (0.76, 1.11)	-0.004
log(AFP)	0.96 (0.24, 3.81)	0.76 (0.15, 3.80)	-0.009
log(sFlt-1)	0.38 (0.08, 1.84)	0.30 (0.05, 1.90)	0.026
log(PlGF)	0.69 (0.26, 1.82)	0.70 (0.25, 1.98)	0.006

*Adjusted for maternal BMI, maternal age, maternal ethnicity, parity and magnesium sulphate.

2.6.4.7 Neonatal composite

No clear evidence of associations were found between the neonatal composite outcome and either NT discordance or CRL discordance, as shown in Table 2.17. The neonatal outcome had missing data for 14 babies, hence only 340 babies were included in this analysis. The baseline C-statistic for the model including only the standard characteristics was 0.790.

Table 2.17: Neonatal composite outcome

	Unadjusted prognostic effect	**Adjusted* prognostic effect**	**Change in C-statistic**
	OR (95% CI)	***OR (95% CI)***	
NT	0.93 (0.30, 2.89)	1.07 (0.33, 3.50)	<0.001
CRL	0.99 (0.91, 1.08)	1.00 (0.91, 1.09)	-0.001

*Adjusted for maternal BMI, maternal age, maternal smoking status, maternal ethnicity, parity, mode of conception, gestation at delivery, administration of steroids, and magnesium sulphate.

2.7 Discussion

2.7.1 Summary of clinical findings

The study provides evidence that an increasing percentage difference in NT and CRL is associated with a fetal adverse outcome composite, including after adjustment for standard prognostic factors defined by maternal variables. Increasing inter-twin CRL discordance was also associated with IUFD and antenatally detected growth restriction, whilst an increasing discordance in inter-twin NT was associated with the development of TTTS.

2.7.2 Strengths and limitations

This study has the benefit of investigating the prognostic values of inter-twin NT and CRL percentage discordance as a continuous variable whereas other studies dichotomised the data using non-validated 'cut-offs of abnormality' which loses important information (often equivalent to throwing away one third of the data) (Altman and Royston, 2006).

Although associations were found between a few of the prognostic factors and outcomes, the issue of multiple testing needs to be considered, as we would expect 1 in every 20 tests performed to be found to be significant simply by chance.

Overfitting may also be an issue due to having a small sample size. However, the clinical focus here was to examine the relationships between the potential prognostic factors and the outcomes of interest rather than determine the performance of the models predicting these outcomes. Furthermore, no predictor selection was performed based on statistical significance to reduce the potential for overfitting.

Due to the relative scarcity of MCDA twins in the UK and Australian general obstetric populations, sample size was an issue within this study. There was large uncertainty in many of the estimates, due to the small number of outcomes occurring, and so the results should be interpreted with caution. To address this, post-publication, this chapter includes findings from logistic regression using Firth's correction to deal with sparse data bias for the primary outcome. This found the overall conclusions did not change, however the odds ratios were marginally reduced.

The risk of the composite fetal adverse outcome (94/177 pregnancies, 53.1%)) demonstrates how high-risk MCDA twin pregnancies are, and how important research in this area is, particularly given the potentially fatal outcome of these complications. Although the use of a composite outcome is less desirable than individual outcomes, they were created to enable meaningful statistical analysis within this relatively specialised area of obstetrics. The conditions included in the primary outcome of the composite fetal adverse outcome are all monitored in the same way: with at least 2-weekly ultrasound scans, and the potential sequelae of TTTS, TAPS and growth restriction are the same: IUFD. The conditions were examined individually as well, but if viewed pragmatically, the clinical action for being higher-risk for one of the conditions in the fetal adverse outcome composite group could be similar for all conditions, for example increased monitoring, and a lower threshold to refer to a tertiary fetal medicine centre. As the study included women who underwent antenatal care at 28 different UK maternity units, the results are generalisable to the UK obstetric population and possibly to other high-income countries, however the study should be repeated in other cohorts to account for different obstetric populations. The results may be less applicable to those in developing countries, who lack the resources to conduct prognostic testing and such intensive antenatal surveillance.

A major strength of this study is that the potential prognostic factors evaluated were evaluated by appropriately accredited NHS laboratories, and the assays are readily accessible, easily and reliably measurable on an automated platform, and only require a small amount of maternal blood thus presenting no risk to the mother or fetuses. The ultrasound measurements used are easily calculable and although ultrasound may be subject to inter-operator variability, all health care professionals performing these assessments require additional certification, which is a national programme in the UK and thus the variability should be negligible. Consequently any prognostic factors with sufficient predictive ability would be clinically useful and feasible to measure at a national level, as with first trimester aneuploidy screening. However, any potential measurement error in the prognostic factors was ignored because information on measurement error was not collected, i.e. from repeated biomarker/ultrasound measurements, and was not available from another source, e.g. from a standalone test study.

2.7.3 Clinical implications and future research

The aim of this study was to identify individual prognostic factors of complications in MC twin pregnancies. The RCOG MC twin pregnancy guidance recommends that "screening for TTTS by first trimester nuchal translucency measurements should not be offered" (Kilby 2016) and the findings of this study support that NT % discordance, and individual NT measurements alone should not be used to predict TTTS.

Although the changes in individual biomarkers do not accurately predict outcome, and their individual predictive ability was thought to be too low to justify combination in a prognostic model, as illustrated in Section 2.5.3, the findings are exciting from a pathophysiological perspective as they suggest that physiological changes occur before the appearance of the ultrasound signs of polyhydramnios and oligohydramnios, and IUFD. This supports that first trimester prognostic factors may exist, and warrant further investigation. Interestingly no potential prognostic factors affected both growth restriction and TTTS supporting that they have different pathological mechanisms. The lack of animal models and scarcity of MCDA twin pregnancies makes the pathophysiology difficult to investigate. No longitudinal studies have been performed prospectively recruiting women with MCDA pregnancies in the first trimester, prior to the appearance of the clinical signs of MC twin complications, and comparing those who subsequently develop a complication. This would help determine whether the differences in the biomarkers are because the biomarker is abnormal earlier in pregnancy, or that it does not increase.

2.7.4 Developing a prognostic model

This chapter also has methodological implications, as it provides an example of developing a prognostic model using a dataset where some of the predictors are not very prevalent and outcomes are quite sparse. The results demonstrate large potential optimism in the model coefficients and the model's predictive performance, such that overfitting is a major problem, with the optimism-adjusted C-statistic much lower than the apparent C-statistic. Further, adjusting for this overfitting using a uniform shrinkage is shown unlikely to be reliable, as the shrinkage factor is estimated with large imprecision. This finding convinced clinical collaborators that developing a prognostic model was not sensible using this dataset, and motivated subsequent methodology work to show the issue of penalisation and shrinkage methods in small sample sizes (Riley et al., 2021a).

2.7.5 Conclusions

This study has demonstrated the potential prognostic ability of individual first trimester ultrasound measurements and maternal serum biomarkers. Currently, there are no established prognostic models for predicting adverse outcome in MC twins. This study has identified potential individual prognostic factors in the first trimester (fetal biometric and maternal serum biomarkers) that show promise but require further robust evaluation in a larger, prospective series of MC twin pregnancies, so that their usefulness both individually and in combination can be defined (Riley et al., 2013). When larger datasets are available, these markers could potentially be combined with standard prognostic variables to form a prognostic model ready for internal and external validation.

Statistically, this chapter also highlighted how it may not always be beneficial to develop a prognostic model, even in a situation where some of the individual factors are identified as having prognostic ability. In particular, overfitting may not be reliably addressed when using penalisation methods if the sample size is small.

Some of the variables examined in this chapter could be subject to measurement error, and the impact of this error on their prognostic value is unclear, or indeed if it should be addressed. Therefore, in the following chapter, a systematic review of prognostic models is performed to ascertain how susceptible to measurement error the predictors used in the final models are and how often the measurement error was acknowledged or accounted for within the development of the models.

3 Measurement error and timing of predictor values used in prediction model research: a systematic review of current practice and reporting

3.1 Introduction

3.1.1 Chapter rationale

The previous chapter demonstrated a prognostic factor research study, where five biomarkers were investigated for their independent prognostic value. However, one of the limitations noted was that potential measurement error in the biomarker values was ignored, because such error information (e.g. from repeated biomarker values per individual, of a biomarker assumed to be in a stable state) was not collected and was not available from another source (e.g. a standalone re-test study). Moreover, it is not clear whether or how measurement error should be examined in prognosis and risk prediction model research. Ideally, such studies should include measurement of predictors that reflects how they will be measured in practice, which may or may not involve measurement error. Hence, whether this is the case in current practice, or if measurement error is even considered, is not clear.

To address this, in this current chapter, a systematic review of recently published prediction models is performed to ascertain how susceptible to measurement error the predictors used in the final models are and how often the measurement error was acknowledged or accounted for within the development of the models. A brief summary of methods that might be used to account for measurement error were discussed in Chapter 1.

A particular aspect of measurement error in the predictors is timing error, so whether the predictors used in the model development were measured at the moment the model is intended to be used in practice. When time-dependent predictors are not able to be measured at `baseline' this creates time-dependent bias, which has been shown to often have an impact on the estimates of key predictors and study conclusions (van Walraven et al., 2004). Additionally, the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis) statement recommends to clearly define when the predictors used in the development of the model were measured (Collins et al., 2015) and states that "all predictors should be measured before or at the study time origin and known at the intended moment the model is intended to be used" (Moons et al., 2015). Nevertheless, for a range of practical and ethical reasons, researchers may design prognosis studies that collect time-varying predictor information after the intended moment of use.

Therefore, this review investigates whether the timing of predictor measurement and intended moment of model use is clearly reported in articles developing clinical prediction models, and if they coincide. Alongside these aims, the review will also examine the quality of the reporting of the prediction models using key domains listed within the Checklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies (CHARMS) (Moons et al., 2014).

The results of this review were published in the Journal of Clinical Epidemiology (Whittle et al., 2018).

3.1.2 Chapter aims

The specific objectives of this chapter are:

- To review the methodology and reporting within a sample of approximately 30 articles developing prediction (prognostic/diagnostic) models for individualised risk prediction using regression-based approaches, across the different domains of the CHARMS checklist;
- To examine how susceptible to measurement error the predictors used in the final models are;
- To examine if and how authors accounted for any measurement error;
- To review whether the timing of the predictor measurements was clearly stated, and if so, its relation to the intended moment of use of the prediction model.

3.1.3 Chapter outline

This chapter begins with a detailed description of the methods of the review in Section 3.2, defining the search strategy, the inclusion/exclusion criteria, how the articles to be included were selected and the information that was to be extracted. The results are then split into two sections. The first section, Section 3.3, describes the studies that were included in the review, using a flow chart to show how many articles were excluded at each stage and why articles were excluded, followed by a review of the methodology and reporting of the model development in the articles included. The second section, Section 3.4, then discusses the intended moment of using the models developed; whether the predictors were measured at this time or not; the susceptibility to measurement error of the predictors included in the final models, and whether this was adjusted for. The chapter concludes with a discussion of the results found in Section 3.5, presenting key findings and the strengths and limitations of the review.

3.2 Methods

3.2.1 Search strategy

A systematic search was carried out to identify articles reporting the development of a multivariable prediction model for either individualised diagnosis or prognosis risk classifications. A priori, the research team decided that approximately 30 articles would be sufficient. In addition to pragmatic reasons, the team considered that 30 was likely to be sufficient in providing qualitative saturation of the general standards of reporting, and in particular whether measurement error and incorrect timings was a general concern for the prediction model field.

The search was carried out on 27th November 2015 in the Medline database, for full-text articles published in English. After screening the first 500 titles and abstracts, it was estimated that the required 30 papers should be identified by searching 1000 titles; hence the most recent 1000 results were exported for screening. The search strategy used was an adaptation of a published search string for finding prognostic and diagnostic prediction studies in Medline (Geersing et al., 2012), which adapted the Ingui filter (Ingui and Rogers, 2001). The search filter was adapted by changing the term "OR 'Multivariable'" to "AND (Multivariable OR Multivariate)" to refine the search further to studies developing multivariable prediction models for individualised prediction, which would hopefully remove other studies just examining associations between specific diagnostic/prognostic factors and an outcome but not developing a prediction model. The search filter searched for the terms in the title and abstract only (excluding the one mesh term specified, and the term ending in ".ti" which searched the title only). The search filter used in this study is given in Table 3.1 below.

Table 3.1: Search Filter – an adaptation of the search strategy by Geersing et al. (2012) which uses a variation of the Ingui filter (Ingui and Rogers, 2001)

1.	Validat$
2.	Predict$.ti
3.	Rule$
4.	**1 OR 2 OR 3**
5.	Outcome$
6.	Risk$
7.	Model$
8.	**5 OR 6 OR 7**
9.	Predict$
10.	**8 AND 9**
11.	History
12.	Variable$
13.	Criteria
14.	Scor$
15.	Characteristic$
16.	Finding$
17.	Factor$
18.	**11 OR 12 OR 13 OR 14 OR 15 OR 16 OR 17**
19.	Decision$
20.	Identif$
21.	Prognos$
22.	**9 or 7 or 19 or 20 or 21**
23.	**18 and 22**
24.	Clinical$
25.	Logistic Models/
26.	**7 OR 24 OR 25**
27.	**19 AND 26**
28.	**23 OR 7**
29.	Prognostic
30.	**28 AND 29**
31.	Stratification
32.	ROC Curve[Mesh]
33.	Discrimination
34.	Discriminate
35.	C-statistic
36.	C statistic
37.	“Area under the curve”
38.	AUC
39.	Calibration
40.	Indices
41.	Algorithm
42.	**4 OR 10 OR 23 OR 27 OR 30 OR 31 OR 32 OR 33 OR 34 OR 35 OR 36 OR 37 OR 38 OR 39 OR 40 OR 41**

43.	Multivariable
44.	Multivariate
45.	**43 OR 44**
46.	**42 AND 45**

3.2.2 Inclusion/Exclusion criteria

Relevant articles were those that met any of the following inclusion criteria:

- Studies developing a clinical prediction (prognostic/diagnostic) model for individualised prediction in human participants, based on a multivariable regression model;
- Studies updating a previously developed prediction model for individualised prediction, by adding new predictors to a multivariable regression model.

Articles were considered as not relevant due to any of the following exclusion criteria:

- Studies developing a model using non-regression-based techniques;
- Studies validating a previously developed prediction model;
- Studies creating a risk score from an existing prediction model;
- Studies using a multivariable model to examine whether a particular predictor is associated with the outcome when adjusting for other factors (predictor finding/prognostic factor research);
- Studies estimating the prognostic effect (e.g. hazard ratio) of a previously developed score;
- Studies updating a previously developed prediction model without adding any new predictors to the model;
- Studies investigating the optimal cutoff value of a previously developed model.

3.2.3 Selection of articles

The titles and abstracts of the 1000 most recently published articles found using the search string were screened for inclusion. The full article was obtained for any articles which were deemed to be potentially eligible or for any articles in which it was unclear from the title and abstract whether they met the eligibility criteria. These full articles were then screened for suitability and categorised into one of three groups: ‘include’, ‘exclude’ and ‘unsure’. The selection of articles until this stage was undertaken by a single reviewer (RW). Articles in the ‘include’ and the ‘unsure’ groups were sent to two additional reviewers. Both reviewers checked all ‘unsure’ articles, and the ‘include’ articles were split between the two reviewers to check they met the eligibility criteria. Any ‘unsure’ articles on which an agreement could not be reached were checked by a fourth reviewer and the decision to include/exclude was based on the verdict of the fourth reviewer. Although the aim was to identify 30 articles, any articles in excess of this that met the eligibility criteria were also retained for inclusion, to avoid any potential selection bias concerns when choosing which articles to remove.

3.2.4 Extraction of information

Data were extracted from the selected articles by a single reviewer (RW). The following items shown in Table 3.2, where available, were extracted from each article and were based on the CHARMS checklist (Moons et al., 2014), with the addition of information related to the intended moment of using the model and measurement error:

Table 3.2: Items to be extracted from each article

Design and aim	• Prognostic versus diagnostic prediction model • Intended scope of the review − Clinical area − Aim of prediction model (e.g. inform therapeutic decision making, inform referral or withholding from invasive diagnostic testing, inform patients of probability of event) • Source of data (e.g. cohort, case-control, randomised trial or registry data)
Outcomes to be predicted	• Definition and method for measurement of outcome • Type of outcome (e.g. single or combined endpoints; binary or time to event)
Candidate predictors	• Number and type of predictors (e.g. demographics, patient history, physical examination, additional testing, disease characteristics) • Definition and method for measurement of candidate predictors • Timing of predictor measurement • Handling of predictors in the modelling (e.g. continuous, linear, non-linear transformations or categorised)
Sample size	• Number of participants • Number of outcomes/events • Number of outcomes/events in relation to the number of candidate predictors (events per variable)
Missing data	• How much missing data • Handling of missing data (e.g. complete-case analysis, imputation, or other methods)
Model development	• Modelling method (e.g. logistic or survival) • Method for selection of predictors for inclusion in multivariable modelling • Method for selection of predictors during multivariable modelling
Intended moment of using the model & timing of predictor measurements	• Intended moment of use • Timing of the measurement of predictors included in the final model, and whether it matched the intended moment of using the model
Measurement error of predictors	• Susceptibility to measurement error for the predictors included in the final model • Whether measurement error was accounted for and, if so, how
Model performance	• Calibration (e.g. calibration slope, calibration plot, Hosmer-Lemeshow test)

	• Discrimination (e.g. C-statistic, D statistic, log-rank) • Classification measures (e.g. sensitivity, specificity, predictive values, net reclassification improvement)
Model evaluation	• Method used for testing model performance: internal (e.g. random split of data, resampling methods, none) or external (e.g. temporal, geographical, different setting, different investigators) • In case of poor validation, whether the model was adjusted or updated (e.g. intercept recalibrated, predictor effects adjusted, new predictors added)

3.2.5 Categorisation of susceptibility to measurement error

Measurement error is a difference between the measured values of a variable and the true values of the variable, or if the variable is categorical, the classification to an incorrect category. Measurement error of predictors is not something that is usually quantified in a prediction model study, therefore a subjective decision needed to be made about whether it existed. The level of susceptibility to measurement error for each predictor used in the final models of the included articles was classified into two categories:

- Low risk: Unlikely to be measured with error, or possibly/likely to be measured with error but expected to be unimportant.
- High risk: Possible or likely to be measured with error and may be important.

For example, age and sex are both extremely unlikely to be measured with error, and any error in age recorded would be expected to be negligible. Thus, age and sex would be classed as 'low risk' with regards to important measurement error. Whereas, blood pressure could be measured with error, as error in blood pressure measurement commonly occurs because of improper techniques such as talking during measurement or wrong cuff size (Handler, 2009) and blood pressure is also commonly measured with error due to biological variability (Grassi et al., 2012). This error could be large and could be important when developing a prediction model for hypertension, for example, because blood pressure is an important component of the diagnostic evaluation for hypertension. Hence, blood pressure would be classed as ‘high risk’ of important measurement error. Another high-risk example would be BMI, which the extent of the measurement error would depend on the way in which it was measured, but there would be a high chance it would be measured with some error.

To categorise the list of predictors into the two groups of susceptibility to measurement error, first the literature was searched for any potential publications discussing measurement error in any of the predictors of interest. For those where no evidence could be found, the categorisations were made based on the judgement of the reviewer, which was corroborated by a postdoctoral academic General Practitioner.

For the predictors that were classed as at ‘high risk’ of measurement error, a justification for this reasoning was then given linking to the main reasons for measurement error presented in Chapter 1 (see Section 1.6):

- Fluctuations in human samples/biological variability
- Inaccuracy of measurement instruments
- Imperfect recall
- Cost/resource limitations
- Subjective nature of measures
- Laboratory or measurer error
- Timing error

3.2.6 Timing of predictors and intended moment of model use

First, the included articles were searched for any information describing when the predictors were measured and when the model would be intended to be used in practice. If this was not explicitly stated then, wherever possible, information given in the articles on where the predictor information came from and the setting they were measured in were used to establish a likely time of measurement. If the intended moment of use of the model was not stated then, again where possible, information on what the model would be used for and the predictors that would be used within the model were considered to make a decision on the most probable intended moment of use.

3.3 Results Part 1: Description of included studies and quality of reporting

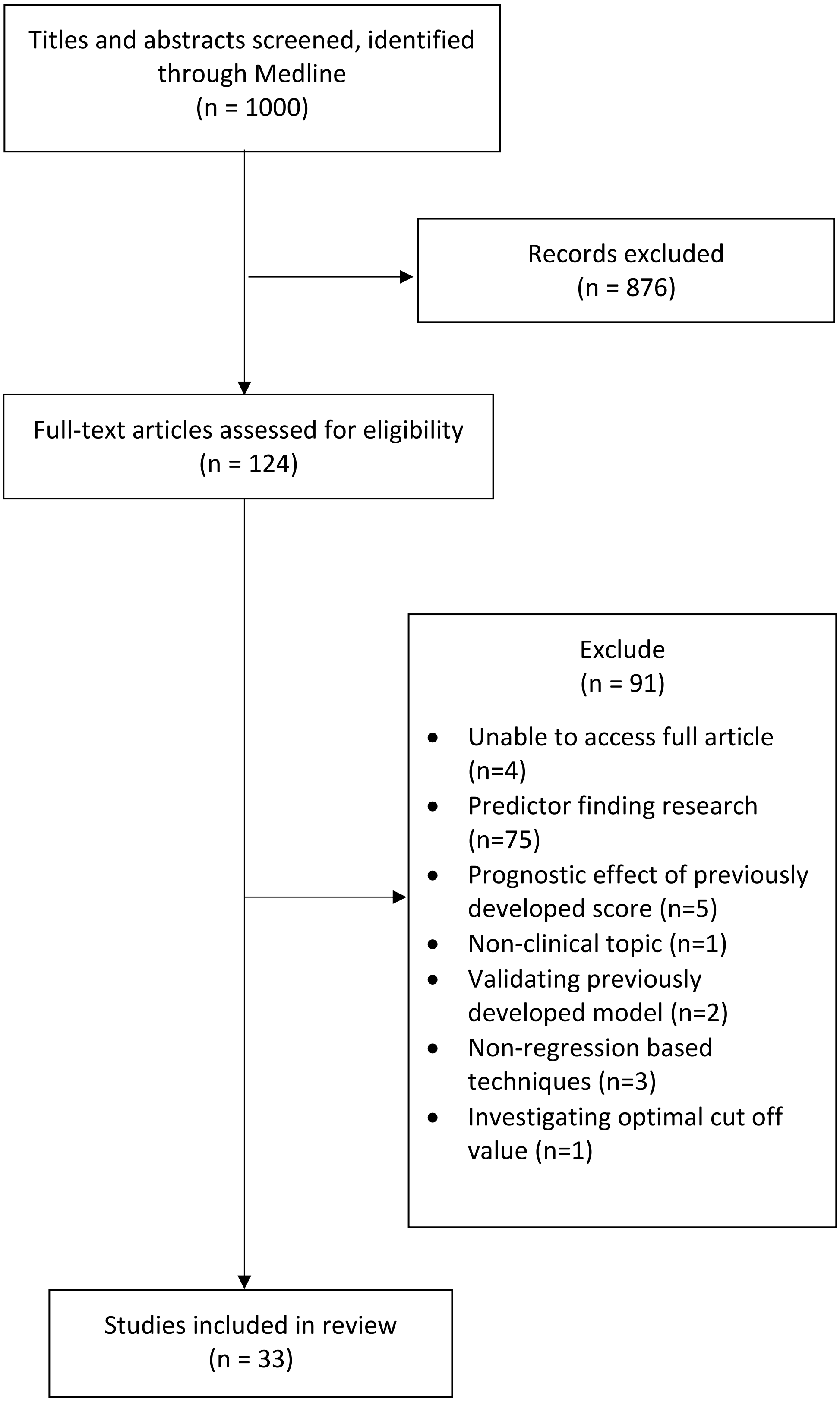
In this first part of the results of the review, a detailed description is given of which articles were included within the review, at what stage articles were excluded, and why they were excluded. This is followed by a summary of the design and aims of the included articles, alongside detailed information about the quality of reporting and characteristics of the development of the models. Part 2 will cover measurement error and timings of predictor measurement.

3.3.1 Literature search and inclusions

A total of 1000 titles and abstracts were extracted from Medline and screened for inclusion. Of these, 876 were excluded based on not meeting the inclusion criteria from screening their title and abstract (Figure – Section 3.3.2). Study eligibility was then assessed for 124 full-text articles by the first reviewer, and 32 were deemed suitable for inclusion, 82 were deemed ineligible and for 10 studies a decision could not be reached. The 32 articles for inclusion were split between the second and third reviewers for double checking, and there was 100% agreement that all these should be included. The 10 unclear articles were sent to both the second and third reviewers and there was agreement between them for six of these articles. The remaining four studies without agreement were reviewed by a fourth reviewer and the decision of this reviewer was taken as final. Of these 10 studies, three were ultimately excluded due to focusing on prognostic factor, rather than prediction model, research.

This left 39 articles for inclusion. However, after beginning the extraction of information from the included articles, six further studies were identified as not meeting the eligibility criteria due to being either prognostic factor research (n=1), not using regression methods (n=3), being unable to access supplementary tables (n=1), or estimating the prognostic effect of previously developed score (n=1). With agreement from the three additional reviewers, these articles were excluded. Therefore, a total of 33 papers, published in 2015, were included in the final review. A separate reference list of included articles is given in Appendix A and will be referred to numerically throughout the remainder of the chapter.

3.3.2 PRISMA flowchart


Titles and abstracts screened, identified through Medline
(n = 1000)
Records excluded
(n = 876)
Full-text articles assessed for eligibility
(n = 124)
Exclude
(n = 91)
• Unable to access full article (n=4)
• Predictor finding research (n=75)
• Prognostic effect of previously developed score (n=5)
• Non-clinical topic (n=1)
• Validating previously developed model (n=2)
• Non-regression based techniques (n=3)
• Investigating optimal cut off value (n=1)
Studies included in review
(n = 33)

3.3.3 Study design and aim

The 33 included articles consisted of 27 prognostic model studies and six diagnostic model studies. The most common clinical condition that a model was developed for was cancer, with 15 (45.5%) of the articles developing a prognostic model for an outcome in cancer patients [1-3, 5, 6, 10, 11, 16, 19, 20, 21, 25, 26, 29, 33] and three (9.1%) developing a diagnostic model for a particular form of cancer [13, 18, 32]. Of the remaining prognostic models developed, one article each developed a prognostic model for an outcome in patients with each of the following clinical conditions: trauma [4], HIV [7], endoscopic retrograde cholangiopancreatography (ERCP) [9], aneurysms [12], pressure ulcers [15], myocardial infarction [17], paediatric cardiac catheterization [22], liver transplant [23], sport concussion [24], atrial fibrillation [27], Crohn's disease [28] and coronary artery disease [30]. The remaining diagnostic models were developed to diagnose a patient with abdominal trauma [14], non-alcoholic fatty liver disease (NAFLD) [8] and acute encephalopathy with biphasic seizures and reduced diffusion (AESD) [31].

In eight (24.2%) of the articles included [1, 2, 7, 10, 16, 23, 24, 26], the prediction model was developed using data from a prospective cohort study. The remaining 25 articles developed the prediction model using data from a retrospective cohort/registry database.

The intended aim of the prediction model in the included articles was most commonly to inform therapeutic decision making (e.g. decisions on surgical treatment or chemotherapy) (n=21, 63.6%) [1, 3, 4, 9, 11-14, 16, 17, 19-21, 23-28, 30, 33]. Other aims of the prediction models were to withhold low risk individuals from invasive diagnostic testing or unnecessary treatment [2, 6, 10, 32], to inform patients of the risk of an event occurring [22, 29], for research purposes [8, 15], to enable early diagnosis of a condition [18, 31], to predict relapse [5] and to identify patients at highest risk for dying to enable targeting of factors (e.g. smoking, depression) that increase mortality [7].

3.3.4 Candidate predictors

In six (18.2%) of the articles included [3, 6, 8, 9, 28, 33], the candidate predictors considered in the development of the prediction model was either not stated or unclear. In the remaining 27 articles, a full list of the candidate predictors could be established.

Twenty-six of these 27 articles had one or more continuous predictors. Of these 26, there were nine (34.6%) which categorised or dichotomised at least one of the continuous candidate predictors, but kept at least one of the continuous predictors as continuous [2, 5, 7, 17, 21, 25, 26, 30, 32]; only four kept all of the continuous predictors as continuous in the model development [4, 12, 19, 27], whilst eight categorised/dichotomised all continuous predictors [10, 11, 16, 18, 20, 22, 29, 31]; in two studies all continuous predictors were kept as continuous within the model but then categorised/dichotomised to simplify for use in practice (e.g. nomogram) [1, 23]; and in three it was unclear how the continuous predictors were handled [13, 15, 24].

In only five (19.2%) of the 26 articles was the linearity assumption of the effect of any of the continuous predictors explicitly considered, or a justification given for why non-linearity was not examined. In these studies, the linearity was considered by using the Box-Tidwell test [17], using restricted cubic splines [13], using fractional polynomials [19], using a log transform and including interactions [28] and by “suitably transforming individual factors where necessary” [1].

3.3.5 Sample size

The number of patients included in the analysis was reported in all 33 of the included articles. The median number of patients included in the model development was 501.5 (IQR 193, 2371). The number of outcome events was clearly reported in 28 out of the 32 (87.5%) studies with a binary outcome. In the studies where it was reported, the median number of outcome events was 88 (IQR 48, 414).

It was not possible to calculate the number of events per candidate predictor for 9 of the 32 (28.1%) studies considering a binary outcome, either because the number of candidate predictors considered was not clear [3, 9, 13, 15, 28], the number of events experienced was not clearly reported [4, 19, 29], or neither the number of events nor candidate predictors were clear [6]. In the 23 studies with a binary outcome that *did* report the number of outcome events and clearly reported the candidate predictors, the median number of events per candidate predictor (also known as events per variable (EPV)) was 5 (IQR 2, 14.6), ranging from 0.75 to 59. Only seven of these 23 studies [2, 5, 8, 10, 12, 16, 22] had an EPV greater than or equal to 10 as is recommended by Peduzzi et al. (1996), and only four had an EPV greater than 20 as recommended by Austin and Steyerberg (2017). Thus, the vast majority of the studies may have too few events relative to the number of predictors considered to produce reasonably precise estimates.

For the one study modelling a continuous outcome [24], the number of participants was 76, and 58 predictors were considered, giving 1.3 participants per each predictor considered, well below the recommendation of 20 (Harrell, 2001, Moons et al., 2014), and even below the recommendation of 2 by Austin and Steyerberg (2015), who found that in linear regression models only 2 subjects were required per predictor for adequate estimation of regression coefficients, standard errors and confidence intervals.

3.3.6 Missing data

Some information on the amount of missing predictor values was given in just 10 (30.3%) of the 33 included studies [4, 5, 9, 11, 14, 16, 18, 21, 26, 28]. However, one of these [18] did not quantify the amount of missing data but stated that two variables (smoking and alcohol status) were not shown because of insufficient data.

In 17 out of the 33 (51.5%) studies, the method of handling missing data was not reported. Of the remaining 16 studies, nine reported the use of complete case analysis [4, 7, 9, 11, 15, 17, 19, 24, 26] with two of these studies [4, 7] using multiple imputation during a sensitivity analysis. Only one study used multiple imputation to deal with missing predictor values [3]. Other approaches to the handling of missing data were case-deletion [16, 28], including a missing category in the analysis [21], excluding variables with lots of missing data [31] and reporting as negative if the result of the test was uninterpretable [30]. One study reported excluding participants in the univariable analysis if they had any missing data but it was not clear how they had handled participants with missing data in the multivariable analysis [5].

3.3.7 Model development

A total of 23 out of the 33 studies (69.7%) used a logistic regression model when developing their prediction model [1-4, 6, 8-10, 12-15, 17, 18, 21-23, 25-27, 31-33], and six used a Cox proportional hazards model [7, 11, 19, 20, 28, 30]. Other methods used to develop models with binary outcomes were joint modelling [5], a competing risk model [16] and a non-mixture cure model [29]. One study developed a prediction model for a continuous outcome, using a linear regression model [24].

In 24 out of the 33 studies (72.7%), it was clearly reported how the predictors were selected for inclusion in the final multivariable model. Of these 24 studies, 15 considered predictors for inclusion in the multivariable model only if they were significant in univariable analysis [1, 8, 9, 11, 14, 16, 18, 20, 21, 23, 25, 29, 31-33], with an additional four studies using a combination of selecting predictors based on a priori hypotheses and their significance in univariable analyses [7, 15, 27, 28].

In 22 out of the 33 studies (66.7%), a stepwise selection method was used to select the variables for the final prediction model based on multivariable analysis. Of these, four used forward selection [9, 10, 26, 33], 11 used backward selection [3, 11, 16, 18, 19, 21-23, 28, 29, 32], one used interactive stepwise selection [7], and in six it was unclear whether forward or backward selection was used [1, 14, 15, 20, 24, 25]. Other methods of selection of the final predictors were to minimise the Akaike's Information Criterion/Bayesian Information Criterion [2, 27], keep statistically significant predictors in the model [5, 12], adaptive least absolute shrinkage and selection operator (LASSO) [8], bootstrap resampling [17], include all predictors significant in the univariable analyses [31] and compare four different models using the AUC [30]. In three studies it was unclear how the final predictors were selected [4, 6, 13].

Two of the 33 studies [18, 28] included a subset of predictors regardless of their significance because they were routinely used in clinical practice or had strong previous evidence of being associated with the outcome.

3.3.8 Model performance

Calibration was assessed in only eight out of the 33 studies (24.2%) included, of which three performed a Hosmer-Lemeshow test [1, 15, 17] and five graphically assessed the calibration [3, 6, 7, 10, 13]. One study presented the gradient of the calibration slope [10].

Conversely, only five out of 33 studies (15.2%) did not evaluate the discrimination of the model [20, 21, 24, 29, 33]. Fourteen (50%) out of the remaining 28 studies presented the receiver operating characteristic (ROC) curve and a value of the C-statistic [1, 3, 4, 6, 8-11, 13, 14, 18, 25-27], 12 (42.9%) reported only the C-statistic [2, 7, 12, 15-17, 19, 22, 23, 28, 30, 32], one study used the Fisher's test to confirm whether the scoring system effectively differentiated between the groups [31] and one study calculated the average accuracy [5]. In addition to the ROC curve and the C-statistic, two studies [17, 27] also reported the integrated discrimination improvement (the difference in predicted probabilities in those who do and do not experience the event of interest), with one of these [27] additionally reporting the net classification improvement (net proportion of events reclassified correctly plus the net proportion of non-events reclassified correctly).

Thirteen out of the 33 studies (39.4%) reported the sensitivity and specificity [1, 3-5, 8, 10, 11, 14, 23, 25, 26, 32, 33], and nine (27.3%) reported the PPV and NPV based on their model predictions according to a chosen cut-point of risk [1, 3, 6, 8, 10, 14, 23, 26, 32].

3.3.9 Model validation

Of the 33 included articles, only three [8, 13, 28] externally validated the model (i.e. evaluating the model in a completely independent dataset to the development data), whereas 17 (51.5%) used some form of internal validation using one of the following methods: randomly split data [1, 8], fivefold cross-validation [2], bootstrap resampling [3, 5, 7, 13, 14, 16, 19, 23, 26, 28], split by date [6, 9], randomly split data and used bootstrapping [4], or 10% validation sample [15]. Of these 17 studies, seven [3, 5, 7, 14, 19, 26, 28] provided an optimism adjusted measure of performance (i.e. C-statistic), but none of the studies used a shrinkage factor to adjust the estimated coefficients of included predictors for potential overfitting (optimism).

3.4 Results Part 2: Measurement error and intended moment of model use

This section assesses whether the predictors included in the final models were susceptible to measurement error, and whether this was adjusted for in the models or was discussed in the articles. The section also details the intended moment of using the models and whether the predictors were measured at this time point or not.

3.4.1 Measurement error

In the 33 articles reviewed, there was a total of 151 different predictors in the final prediction models. Many of the predictors were included in several different models, for example, age and gender were in many of the models (13 models and 5 models, respectively). Of the reported predictors included in the final models, 51 (33.7%) were categorised as high risk of being susceptible to measurement error, and the remaining 100 (66.2%) were categorised as low risk. As there was very little mention of measurement error within any of the studies, the categorisations of susceptibility to measurement error into 'high risk' and 'low risk' had to be made based on the methods reported in Section 3.2.5 rather than from information given in the articles. These categorisations are given in Appendix B for each of the 151 different predictors in all the models examined, which are further grouped into key reasons for likely measurement error in those deemed to be at high risk.

Despite one third of the included predictors being susceptible to measurement error, very few studies acknowledged, or accounted for, measurement error in their included predictors. One study [1] mentioned measurement error as a general limitation due to the study being from a single centre and two studies used repeated measurements of a predictor within the modelling process [2, 5], which may have alleviated the issue to some extent. The first of the studies that used repeated measures [2] used generalised estimating equations (GEE's) to fit models accounting for the correlations among multiple biopsies that were performed on the same patients. The authors state that GEE's yield the same mean predictions as maximum likelihood, but result in inflated standard errors, wider confidence intervals and diminished statistical significance that more accurately reflect the amount of uncertainty in the data. There is no mention of measurement error within the article, and so it assumed that the authors have not made use of the repeated measures in a conscious effort to reduce measurement error, but to take advantage of all the data available (which may consequently potentially minimise measurement error). The second study using repeated measures [5] used joint modelling of longitudinal measures of CA125 with the stated purpose of estimating the time trend of CA125 rather than explicitly accounting for measurement error (although again, this may consequently account for measurement error).

Despite only two of the reviewed articles including repeated measurements in the modelling process, repeated measurements of at least one of the candidate predictors were actually reported to be available in six (18.2%) of the articles. One of the studies [8] had data available from a data registry, and if there was more than one value available for a predictor, recorded within a 12 month period, then they used the average of the values. Multiple pain scores were recorded in one study [9] and the authors explored different ways of using these multiple measures in the prediction modelling, for example, first recorded, highest score, change in score, median score, last reported and area under the pain curve. The area under the curve was used within the model development because it was reported to have the highest correlation with the outcome. Two models were developed within one of the studies reviewed [15], from two different time frames: firstly using data available only at the first encounter and secondly using data available from the whole course of care. The final study with repeated measures recorded [23] had multiple blood test results within the first 24 hours after a liver transplant and used the highest of these values within the modelling. Although multiple measurements were recorded, only the two studies discussed above [2, 5] used methods which could have lessened the impact of measurement error, the other four articles [8, 9, 15, 23] did not address measurement error. Of these 6 articles that repeated measures were available, 4 of these had repeated measures that we categorised as being high risk [5, 8, 9, 23], one of which used the repeated measures within the modelling [5]. In the remaining 28 (84.8%) of the articles, there was no indication that repeated measurements were recorded.

Examples of those predictors that were considered at high risk of error are given in Table 3.3 with a reason and explanation for why they may have been measured with error. One example of a predictor at high risk of error and used in several of the final prediction models is prostate specific antigen (PSA). Roehrborn et al. (1996) conclude that there is significant variability between two serum PSA measurements obtained within a short time interval, which is due to chance alone. Biomarkers such as CA125, creatinine, C-reactive protein, serum albumin and other serological markers are also likely to change if a second sample was assessed, meaning they are measured with error due to biological variability causing discrepancies away from an underlying (mean) value (Braga and Panteghini, 2016). There is also the possibility of laboratory error being present in these biomarkers, as the equipment or methods used to take the measurements within the laboratory may not be accurate.

In another example of a predictor likely measured with error, Ali et al. (2007) found that the depth of myometrial invasion (DMI) was different in 29% of cases when the DMI was reassessed. The area under a patient's pain curve could also be measured with error as it is a subjective measurement that may be affected by various things including how the question is asked, the setting in which the question is asked or when the question is asked. It could also be subject to recall error if the patient is asked about previous days pain levels. Another example is pulse rate, where Kobayashi (2013) found that error occurred when pulse rates were objectively scored for various durations (e.g. 10, 15 or 30 seconds) rather than for a whole minute, so the error in a pulse rate could depend on how long the pulse was taken for. A patient's primary tumour diameter is another example of a predictor susceptible to measurement error. This is because if a histologist determined the diameter under a microscope there would be little deviation from the true value, whereas if a surgeon recorded the diameter using an endoscopy then this could be recorded with error and could have an effect on the therapy chosen to be used (Mori et al., 2015). BMI is also another predictor susceptible to error, and again the amount of error would depend on how it was measured. If measured by a clinician then there is unlikely to be much measurement error, but if measured by the patient and recalled this may be subject to error (Hill and Roberts, 1998).

Table 3.3: Explanation of reasons for measurement error in example predictors at high risk of error

Predictor	**Key Reasons**	**Explanation**
Area under pain curve	Subjective/subject to recall	Requires patient to report pain, which is a subjective measure and could report the same pain differently at a different time/by a different method or if previous scores were not provided (Scott and Huskisson, 1979), and recall incorrectly (Daoust et al., 2017)
Body Mass Index	Inaccuracy of measurement instruments/imperfect recall	Scales may not be calibrated correctly, or patient reported weight may be incorrect (Hill and Roberts, 1998)
CA125	Biological variability/ laboratory error	Assay imprecision can contribute considerably to result variations in a conventional laboratory setting (Tso et al., 2006) and changes can occur due to normal biological variation (Tuxen et al., 1999)
Creatinine on admission	Biological variability/ inaccuracy of measurement method	Bias and imprecision may occur by use of different measurement methods (Peake and Whiting, 2006) and changes can occur due to normal biological variation (Reinhard et al., 2009)
C-reactive protein	Biological variability	Within-individual variability exists, so a second sample may produce different results (Macy et al., 1997)
CRUSADE score	Biological variability/measurer error	May be different if calculated again shortly afterwards as includes measure that vary and may be affected by measurer error such as of blood pressure (Handler, 2009)
Emergency room pulse rate	Biological variability/inaccuracy of measurement method	May change if measured a couple of minutes later and there may be error depending on how long the measurer counted for (Kobayashi, 2013)
History of transactional sex	Imperfect recall	Patient may not be truthful about history (Sawers, 2013)
Glomerular filtration rate	Biological variability	A second sample may produce different results due to biological variation (Delanaye et al., 2012)
Human epididymis protein 4 (HE4)	Biological variability	A second sample may produce different results (Braga et al., 2014)

Ki-67	Biological variability	A second sample may produce different results and differences may be present from different laboratories (Polley et al., 2013)
Myometrial invasion depth	Measurer error/inaccuracy of measurement method	Results may be different when reassessed (Ali et al., 2007)
Prostate specific antigen (PSA)	Biological variability	A second sample may produce different results (Roehrborn et al., 1996)
Serum albumin	Biological variability	A second sample may produce different results (Winkel et al., 1974)
Tumour stage	Subjective/measurer error	May get a different result from different assessors dependent on experience level or areas of speciality
Primary tumour diameter	Laboratory or measurer error	Measuring using an endoscopy could be inaccurate (Mori et al., 2015)

While many of the predictors in the final models could be considered to be susceptible to measurement error, there were also examples of predictors that were considered to be at low risk of important error. For example, one model that aimed to identify trauma patients at high risk of pulmonary embolism included a predictor indicating if the patient arrived at the hospital by helicopter, and it would be unlikely this would be incorrectly classified. Other models included the patient's disease location as a predictor and again, it is unlikely that this would not be recorded correctly.

3.4.2 Intended moment of using the model

Only eight of the articles explicitly stated exactly when the intended moment of using the model would be, or exactly when the predictors used in the final model were measured. However, for the majority of the 33 included articles it was possible to make a reasonable assumption about these details. If these assumptions were indeed correct, then in 30

(90.9%) of the 33 articles the predictor measurements were all either taken at the intended moment of model use or were available prior to this.

For example, one study [20] developed a model to predict survival prognosis after surgery in patients with symptomatic metastatic spinal cord compression from non-small cell lung cancer, with the aim of being able to provide optimal treatment. Although the specific timing of the predictor measurements was not stated, the predictors were specified as preoperative characteristics. The assumption was made that the model would be intended to be used at the point when a treatment decision was being made, as it was reported that those with the most favourable survival prognosis may instead be treated with more radical surgery. Therefore, it was assumed that the preoperative characteristics considered as predictors were either measured prior to or at the point that the model would be intended to be used.

In another example [18], a diagnostic model was developed to predict colorectal cancer in patients selected for colonoscopy in a primary health care setting, with the aim of identifying high risk patients to reduce the time to diagnosis and hence provide more efficient treatment strategies and success. As the model is to be used to help identify high risk patients when being considered for colonoscopy, which would happen during a GP consultation, it was assumed that the model would be intended to be used during a GP consultation when considering referral for colonoscopy. The model used predictors recorded in routine care data, which would all be available at the point of care, and although the article did not state at which time the predictors were recorded, it was assumed that only measurements recorded prior to colonoscopy referral were considered in the model development.

In all six of the articles in which repeated measurements of the predictors were available, each of the repeated measures were recorded either at or prior to the intended moment of using the prediction model. For example, a model was developed [2] to predict disease progression in men with prostate cancer, with the aim of preventing invasive and costly diagnostic testing. The men included in the study were all on active surveillance, so had multiple biopsies undertaken following diagnosis, which were all used within the model development. The article states that the aim was to "identify predictors of the outcome of biopsy on active surveillance, including clinical, biomarker, and pathologic data from previous visits and biopsies" which would suggest that predictors were measured prior to the intended moment of using the model.

In another example in which repeated measures were recorded [5], a model was developed to predict recurrence of ovarian cancer in women who had all reached complete remission after cytoreductive surgery and first-line chemotherapy. Repeated measures of CA125 were used in the model development, and the measurements used were those recorded between the time of diagnosis and the completion of first-line chemotherapy. These measures must all inherently have been recorded before the intended moment of use of the model as recurrence could only occur in those who reached remission after completion of first-line chemotherapy.

In two (6.1%) of the articles [10, 26] it was not possible to make an assumption with regards to when the predictors were measured in relation to when the model was intended to be used. In the first article [10], a prognostic model was developed to predict the specific risk of non-sentinel node metastases in women with breast cancer with the aim of preventing unnecessary axillary lymph node dissections. The model was intended to be used after diagnosis of breast cancer, and as it is to be used to prevent unnecessary axillary lymph node dissections it could be assumed that the model would be intended to be used when deciding whether to perform an intraoperative axillary lymph node dissection. Little information was given on the predictors used in the model meaning the timing of the measurements of the predictors could not be deciphered, hence it was not possible to determine whether the predictors were measured at the intended moment of using the model or not. In the second article [26], a model was developed to predict unfavourable disease in patients with prostate cancer. The aim of the model was to avoid or postpone interventions in subjects with prostate cancer of low biological potential. The article states that the model is intended to be used in patients after radical prostatectomy, but who were eligible for active surveillance. The predictors included were recorded from clinical evaluation, prostatic biopsy and radical prostatectomy specimens, but the timing of the clinical evaluation and prostatic biopsy was unclear and hence it was unknown whether these were before, at, or after the intended moment of using the model.

For one of the included articles [8], a classification algorithm was developed for the diagnosis of non-alcoholic fatty liver disease (NAFLD). The model was not developed to be intended to be used at a specific time but to be used to identify large scale longitudinal cohorts from electronic medical records for use in research studies.

3.5 Discussion

This chapter aimed to provide a review of recently published prediction models by evaluating the quality of the reporting, whether the predictors were measured at the same time that the prediction model is intended to be used, and the potential and adjustment for measurement error within predictors in the development of the models. The key findings, limitations and implications of the review are now summarised.

3.5.1 Key findings

A summary of the key findings is given in Box 3.1.

Box 3.1: Summary of key findings from the review

Measurement error and timing of predictors:

- Many of the final prediction models included predictors likely to be susceptible to measurement error.
- Error in predictors was generally not acknowledged, or accounted for.
- Most of the articles did not explicitly state when the predictors were measured or the intended moment of using the model.
- A reasonable assumption about the timing of predictors and intended moment of using the model could be made for the majority of the articles included and, based on this, there were no articles that obviously recorded a predictor after the time it was intended to be used.

General reporting/development:

- A full list of predictors considered was often not given.
- Continuous predictors were frequently categorised in the model development.
- The amount of missing data and the method of handling missing data was often not reported.
- Predictor selection procedures were commonly not described.
- Calibration was rarely assessed.
- Non-linearity of predictors was rarely assessed.
- Often, the events per candidate predictor could not be calculated due to not enough information being reported on the candidate predictors and/or the number of events.

3.5.1.1 Measurement error

The main finding of the review was that many published clinical prediction models include predictors that are susceptible to potentially important measurement error. However, this was seldom acknowledged. Of 33 articles in this review only two used methods that could potentially account for measurement error by using repeated measurements in the modelling (though, even then, it was not stated that this was why the repeated measurements had been used). Though the impact of ignoring measurement error in the articles reviewed is difficult to establish, it raises an important methodological consideration for future prediction model research to address, particularly as approximately a third of the predictors used in the prediction models were categorised as being at high risk of being susceptible to measurement error. Indeed, measurement error has been found to generally have three main effects if not accounted for in medical research: biased estimates of the parameters, loss of power and masking the features of the data (making it harder to spot relationships via graphical methods) (Carroll et al., 2006). The direction and magnitude of bias from measurement error depends heavily on whether the distribution of errors for one variable depends on the actual value of the variable, the actual values of other variables, or the errors in measuring other variables (Rothman et al., 2008), as well as on the true strength of the association, the prevalence of the predictors (Jurek et al., 2005) and whether the errors are random or systematic. Hence, the direction of bias from predictor measurement error is likely to be difficult to predict. However, we know that failing to adjust for random measurement error can lead to estimates being biased towards the null (Prentice, 1982) (if the error is non-differential and independent (Rothman et al., 2008)), which could subsequently lead to an underestimate of a patients' probability of outcome if measurement error is present in the prediction model used.

Conversely, failing to account for systematic errors may change the results in different directions, which could again lead to incorrect predictions of a patients' probability of future outcome. In fact, measurement error in prediction models has been shown to reduce the C-statistic and increase the Brier score (BS) dramatically (Khudyakov et al., 2015), though this study focused on the gain in prediction performance from using error-free predictors instead of error-prone predictors, rather than the gain in prediction performance from accounting for the measurement error in the model when the true error-free values are not known. The article also only evaluated the scenario where only one error-prone predictor was included in the prediction model. Another article assessed the impact of random and systematic error in self-reported height and weight on the performance of a model used to predict diabetes (Rosella et al., 2012). The authors found that random error reduced the calibration and discrimination, and biased the predicted risk upwards, whereas systematic error reduced the calibration and biased the predicted risk in the direction of the bias, but had no effect on the discrimination.

3.5.1.2 Predictor timings

This review found that over three-quarters of the articles included did not explicitly state the exact timing that the model is intended to be used in clinical practice, or exactly when the predictors used in the modelling development were measured. However, a reasonable assumption could be made for the majority (93.9%) of the articles included and, based on this, there were no articles that obviously recorded a predictor after the time it was intended to be used. Nevertheless, future prediction model research studies must clearly report the timing of their predictors, to make it explicit that they were collected before the intended moment of use (or if not, to justify why).

3.5.1.3 Development and reporting issues

An important consideration when developing a prediction model is the number of participants in the data being used, but more importantly, the number of outcome events occurring in the population used to develop the prediction model. The TRIPOD statement (Collins et al., 2015) states that all predictors used in developing the multivariable prediction model should be clearly defined, including how and when they were assessed. The reporting of a full list of candidate predictors and how they were treated within the development of the model is essential to be able to calculate the EPV, but in many of the articles included in this review either a full list of the predictors considered was not explicitly stated or it was not clear whether continuous predictors were treated continuously within the model development. All of the articles included reported the number of patients used in the analyses, but in nearly a third of the studies the number of events per candidate predictor could not be calculated either because a full list of the candidate predictors was not given or the number of events occurring was not stated. In over two-thirds of those that the EPV could be calculated, the EPV was below the recommended value of 10.

Knowing the extent of missing data and also how participants with missing data were dealt with is important when assessing for any potential bias within a prediction model (Moons et al., 2014), as the amount and type of missing data and the method of dealing with missing data can have a big impact on the accuracy of the predictions from the model

(Gorelick, 2006, Little, 1992). Again, the TRIPOD statement states that a description of how missing data were handled needs to be given and the number of participants with missing data for predictors and outcome should be described (Collins et al., 2015), as bias could arise due to data not being missing completely at random (Moons et al., 2015). Reporting of information on missing data was limited in the included articles, within no mention of the amount of missing data in over two-thirds of the articles and minimal information in the remaining. Similarly, the method of dealing with missing data was not stated in over half of the articles.

The TRIPOD statement also states that all model-building procedures, including predictor selection, should be specified (Collins et al., 2015), in enough detail that a knowledgeable reader could verify the reported results and should understand the reasons for the approaches taken (Moons et al., 2015). Prediction models are often derived from a sequence of data-driven steps, but usually only the best prediction model is reported (Moons et al., 2009b), which is likely to lead to bias and can lead to selecting an over-fitted model. While this was the case for all but two of the articles in which it was clearly stated, almost a third of the articles did not clearly report how the predictors to be considered in the multivariable analysis were selected.

Reporting of model performance is essential for the readers and future model users to be able to judge how well the model will perform in practice. The most important considerations of a model’s performance is the discrimination and calibration, with discrimination being the primary interest in model development studies (as they will be well calibrated by definition) (Moons et al., 2015). Around only a quarter of the articles included reported an assessment of the amount of agreement between the observed outcomes and the predictions (calibration) whereas the majority did report evaluating how well the model identifies those who experience an event from those who do not (discrimination).

As mentioned, the reporting of predictor timings was poor, along with reporting of the predictors considered, the number of outcome events, the amount and handling of missing data, the predictor selection procedures and model performance measures. The reporting was also sub-optimal in regard other things listed in the TRIPOD checklist (Collins et al., 2015). In 12% it was unclear how continuous predictors were handled, only 19% described whether the linearity of continuous predictors was considered, and 13% of the studies with a binary outcome did not clearly state how many participants had the outcome of interest.

3.5.2 Strengths and limitations

A strength of this review was that a clearly defined search strategy which was based on a previously published search (Geersing et al., 2012) was used, so while many of the original articles found may have been irrelevant, relevant papers should not have been missed. Although this review did not include a search of every prediction model published within a certain time period due to the sheer volume of prediction models published (Wessler et al., 2015), a search of a few of the most recently published studies was deemed appropriate to enable a general overview of the current literature and provide qualitative saturation of the general standards of reporting, in particular whether the predictors were likely to be susceptible to measurement error and whether this was considered and also the timing of the predictor measurements in relation to the intended moment of using the model. Only

33 papers were included, but it was judged that little would be gained from reviewing a larger number of articles.

The review may have been limited by only searching one database, and by including articles published in English only, but as the review did not aim to capture all the evidence available this should have had a minimal effect on the results found.

A limitation of the review is that many of the articles did not explicitly state the timing of measurements or when the model is intended to be used. Hence, the reviewer's judgement had to be used and assumptions made. Based on this, all of the papers here did actually measure the predictors at the intended moment of using the model (or before), in those that it was possible to decipher this information. However, it is possible that some of these assumptions made were incorrect.

Another concern within prediction models in relation to predictor timing is the relevant time window, or the length of the induction period, in which the predictor of interest is causally related to the outcome. For some prediction models, certain causal factors may need to be considered from much longer ago than others, i.e. with a longer induction period. For example, if considering asbestos exposure in relation to future lung disease, the association could span back many years, whereas recent asbestos exposure may not be related to the outcome if the induction period is only relatively short, e.g. 1-2 years. On the other hand, when predicting infectious diseases, the current and recent exposure of the patient is likely to be most important, and so a relatively short induction period would be needed. Hence, the duration of follow-up of predictors prior to the intended moment of model use should be clearly specified when developing a prediction model, however we did not assess this within this review.

When developing a prediction model, the calendar year of time in which the measurements were made is important (relative to the calendar time of the intended moment of model use), because the precision of measurements often improves when using newer measurement methods. Using a more recent, up-to-date data set that used more improved measurement techniques to develop a prediction model would potentially provide a more relevant and better performing model than if using an older dataset. While study recruitment dates are generally reported, we did not consider this in relation to when the article was published or would be intended to be used.

Due to many of the included studies not actually stating a complete list of all of the candidate predictors considered in the model development, only the predictors included in the final models were assessed for their susceptibility to measurement error. However, measurement error in the candidate predictors could lead to the exclusion of these predictors in the model development stage and so measurement error in these predictors could be as equally as important as measurement error in the predictors in the final models.

Again, little information was given within the included articles about any measurement error that may be present in the predictors. Without the availability of previous research on the amount of error in certain predictors, a subjective decision on whether measurement error was likely or important had to be made by the reviewer. Although an academic general practitioner also reviewed the list of predictors and gave their opinion on whether they would judge the predictor to be susceptible to measurement error when using in practice, it is possible that the way in which measurements were obtained in the research studies differs to the methods used in clinical practice. One difficulty with making a decision on whether the predictor is likely to be susceptible to measurement error was that for many of the predictors it would depend on exactly how the predictor was measured, but often this level of detail is missing from the article. Despite this subjective approach to categorising measurement error, there were several predictors included in the final models that had corresponding published research suggesting they are likely to be measured with error, and this was not considered within the development of the models.

Since the completion of this review, the Prediction model Risk Of Bias ASsessment Tool (PROBAST) has been published (Wolff et al., 2019). PROBAST is a tool that enables the risk of bias and the applicability of prediction model studies that are extracted as part of a systematic review to be evaluated. However, assessing the risk of bias and the applicability of the included studies in this review would not be relevant as the aim of the review was to get an overview of how prediction model studies are being reported, regardless of their potential bias or applicability to a particular field.

3.6 Recommendations

Based on the findings of this review, a list of recommendations for improving the reporting of prediction models in relation to measurement error, timing of predictors and model development are made in Box 3.2.

Box 3.2: Recommendations for improvements in the reporting of multivariable prediction models, based on findings of the review.

Measurement error and timing of predictors:

- State exactly when and how predictors were measured.
- Clearly state when the prediction model is intended to be used.
- If predictors were measured after the intended moment of use of the model, justify why.
- Discuss any susceptibility to measurement error in predictors used in the modelling process, even if it is thought not to be an issue.
  - Describe any potential error or why error is unlikely to be a problem
  - If the measurement error has been accounted for, state how.
  - If the measurement error has not been accounted for, state why.

General reporting:

- Define all predictors being used in the modelling process.
- Describe how predictors were handled within the modelling (e.g. categorisations, considerations of linearity).
- Specify the number of participants with and without the outcome.
- Justify why the sample size is deemed sufficient.
- Describe how many participants had any missing data and how missing data was handled in the analyses.
- Outline model selection procedures, describing how predictors were selected for inclusion in the final model.
- Report measures of model performance, including calibration and discrimination.

3.7 Conclusions and rationale for next chapter

Although there were no clear examples within this review of a prediction model being developed using a predictor that was measured after the intended moment of using the model, it is common in prognosis studies of recurrent and long-term conditions presenting to primary care for information on predictors (e.g. pain intensity) to be ascertained by mailed self-complete questionnaires, or personal interview and examination in research clinics several days after their index consultation (Diehm et al., 2011, Hermsen et al., 2011, Licht-Strunk et al., 2009, Radanov et al., 1991, Scheele et al., 2011, Von Korff et al., 1993, Wardenaar et al., 2014). It was found in this review that the timing of the measurements and the intended moment of using the model is often not explicitly stated which could mean that future users of the model unknowingly estimate misleading probabilities of a patients' outcome if they are using predictors measured at a different time than those used in the model development in relation to the timing of the model use. Hence, the impact of measuring predictors after the intended moment of use needs to be evaluated. To address this, the effect that measuring a time-varying predictor after the intended moment of using a prediction model has on the predictor-outcome associations and model performance will be assessed and illustrated using a real example in Chapter 4.

It is possible that many published prediction models include predictors that are measured with error, and this is often not accounted for or even considered. Additionally, even if the authors considered the predictors to be measured without error, either because of the way they were measured, or for some other reason, this was still not stated within the articles. This suggests a need to assess whether ignoring measurement error in prediction models is a concern and whether accounting for the error will improve the predictions made and the model performance. However, researchers should be considering how susceptible to measurement error their predictors may be when developing a model and the impact this may have on subsequent performance in new data (in particular, calibration of prediction).

This review has also found that while guidelines have been published providing a checklist for the reporting of prediction models (Collins et al., 2015), many researchers are still omitting vital information when publishing their work. Prediction models are developed to help guide clinicians in practice, and the majority of the models developed in the articles included here were intended to be used to assist clinicians in therapeutic decision making. Poor reporting will have an impact on researchers and practitioners who are planning to use a prediction model already developed when assessing whether it is applicable to their situation. Journal reviewers and editors also need to be able to assess the generalisability of the model and the accuracy of the results, which may be difficult if it is not clearly reported within the article. Articles poorly reporting the development of prediction models may not be implemented in practice or may provide poor predictions if used.

Although the review was undertaken in 2015 (and published in 2018), there is no known updated evidence to suggest that reporting of these issues has improved since then. The TRIPOD guidelines were published around the same time as this review was conducted and therefore it would be of benefit to update this review and investigate whether reporting standards have improved since. However, there have been several recent reviews of prediction models for specific health conditions that have found poor adherence to the TRIPOD guidelines (recurrent stroke in patients with transient ischaemic attack (TIA) and minor stroke (Abdulaziz et al., 2022), melanoma (Kaiser et al., 2022) and idiopathic pulmonary fibrosis (Di et al., 2022)).

Many of the recommendations provided here relate to general reporting guidelines that are already listed within TRIPOD, as is the issue of the timing of predictors. More research may be needed into the impact of measurement error in predictors used in prediction models and how this may be alleviated, before recommendations to include measurement error items in the TRIPOD guidelines are made.

4 The effect of measuring time-varying predictors at a different time point to that of the intended moment of use: an illustrative example

4.1 Introduction

The literature review in Chapter 3 found that the timing of predictor measurements and the intended moment of using a prediction model are often not explicitly stated. This chapter uses an applied example to illustrate the effect of using information on time-varying predictors that was measured after the intended moment of use of the prediction model on the estimates of predictor-outcome associations (predictor effect / prognostic factor effect) and on the prediction model performance. The results of this chapter were published in Diagnostic and Prognostic Research (Whittle et al., 2017).

4.1.1 Background

Many studies are published each year which examine potential predictors (prognostic factors) of outcome risk (Riley et al., 2013), and/or develop a prognostic model containing multiple predictors for individualised risk prediction (Steyerberg, 2010). Prognostic models are intended "to assist clinicians with their prediction of a patient's future outcome and to enhance informed decision making with the patient" (Steyerberg et al., 2013). Hence, predictions from these models should have optimal performance at the time that they are practically implemented – the "intended moment of using the model" (Moons et al., 2014).

The TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis) statement recommends to clearly define when the predictors used in the development of the model were measured (Collins et al., 2015), and states that "all predictors should be measured before or at the study time origin and known at the intended moment the model is intended to be used" (Moons et al., 2015). In the context of primary care, this will typically be at the point of care – the primary care consultation. For a range of practical and ethical reasons, researchers may design prognosis studies that collect predictor information *after* the intended moment of use. For example, one approach commonly used in prognosis studies of recurrent and long-term conditions presenting to primary care is for information on predictors (such as pain intensity) to be ascertained by mailed self-complete questionnaires, or personal interview and examination in research clinics several days *after* their index consultation (e.g. (Diehm et al., 2011, Hermsen et al., 2011, Licht-Strunk et al., 2009, Radanov et al., 1991, Scheele et al., 2011, Von Korff et al., 1993, Wardenaar et al., 2014)). This approach offers several advantages: it permits a wider range of predictor information to be collected than would be possible within a time-constrained primary care consultation, it allows for greater standardisation of data collection procedures, and it provides a 'cooling off period' between the patient being informed about the study at the point of care and consenting to provide information on potential predictors that would not be considered part of routine care. However, this practice also carries potential limitations when the measured values of the predictors included in these studies are time-dependent and particularly when they may additionally be sensitive to the choice of measurement, mode of administration and other contextual influences on participants' responses (Bowling, 2005, Podsakoff et al., 2003, Streiner et al., 2014). In these circumstances, estimates of predictor-outcome associations and prognostic model performance obtained from the study may be systematically different (biased) from those that would have been observed had those predictors been measured at the point of care. This problem is what is referred to as indirectness in the GRADE guidelines (Guyatt et al., 2011), the effect of which could be assessed in a particular prognostic model if external validation was performed in a setting and timeframe the same as when the model would be used in practice, as recommended in the REMARK guidelines (Altman et al., 2012).

4.1.2 Chapter objectives

The aim of this chapter was to illustrate this concern using a real example, showing how using a measure recorded shortly after a patient's index consultation to develop a prediction model can provide misleading estimates if used during this index consultation.

The specific objectives were to:

- Compare the direction and magnitude of predictor-outcome associations of a multivariable prognostic model under two scenarios:
    - Using a time-varying predictor of interest, ascertained by the treating physician at the point of care (i.e. the intended moment of use)
    - Using the same predictor, but ascertained by a self-complete questionnaire mailed several days after the point of care
- Compare the differences in the model performance measures under these two scenarios

The remainder of the chapter is structured as follows. Section 4.2 introduces the datasets used in the example, followed by a description of the methods used. Section 4.3 describes the results of the analyses which are then discussed in Section 4.4.

4.2 Methods

4.2.1 Datasets

Secondary analyses of two primary care longitudinal data sets were undertaken: the Prognosis Research (PROG-RES) observation study (Mallen et al., 2006) and the Primary Care Osteoarthritis Screen cluster randomised Trials (POST) (ISRCTN40721988). These datasets were provided by Keele Clinical Trials Unit.

4.2.1.1 PROG-RES

PROG-RES is a prospective observational cohort of five general practices in North Staffordshire, England. Patients aged above 50 years who were consulting for non-inflammatory musculoskeletal pain were recruited between September 2006 and April 2007. Data were collected by the GP during the initial consultation then by means of a self-complete questionnaire shortly after the consultation (median time between point of care and return of post-consultation questionnaire: 17 days (IQR 13, 27)), and at 3 months, 6 months, 12 months, 24 months and 36 months post-consultation.

4.2.1.2 POST

POST is a cluster randomised controlled trial of an intervention for ultra-brief screening questions for anxiety and depression, and pain intensity measurement against a control of just screening for pain intensity. Patients aged 50+ who consulted for suspected or diagnosed peripheral joint arthritis were recruited between September 2011 and November 2012. Data were collected by the GP during the initial consultation then by means of a self-complete questionnaire shortly after the consultation (median time between point of care and return of post-consultation questionnaire: 21 days (IQR 16, 30)), and at 3 months, 6 months and 12 months post-consultation.

4.2.1.3 Similarities

Both studies included a brief standardised assessment of predictors during the consultation (point of care) by the treating GP which they recorded on the practice computer.

The studies had similar patient populations, recruitment procedures, and measurement of predictors and outcome, thereby allowing an observation of whether similar findings were present within the two comparable studies (Table 4.1).

Table 4.1: Design of the datasets

	PROG-RES	**POST**
Design	Prospective observational cohort	Cluster RCT
Registration	(Protocol; (Mallen et al., 2006))	Current Controlled Trials ISRCTN40721988
Intervention	Usual care	Intervention: Ultra-brief screening questions for anxiety and depression + pain intensity measurement Control: Screen for pain intensity
Setting	5 general practices in North Staffordshire, England	45 general practices in West Midlands, England
Period of recruitment	Sep 2006 - Apr 2007	Sep 2011 - Nov 2012
Inclusion criteria	Consecutive patients aged 50+ years consulting for non-inflammatory musculoskeletal pain	Consecutive patients aged 50+ years consulting for suspected or diagnosed peripheral joint osteoarthritis
Exclusion criteria	Vulnerable patient (e.g. diagnosed with dementia); recent trauma associated with significant injury; inflammatory arthropathy	Vulnerable patient (e.g. diagnosed with a terminal illness); nursing home resident; recent trauma associated with significant injury; inflammatory arthropathy, crystal disease, SpA, PMR
Data collection points*	**In GP consultation (point of care), post-consultation questionnaire**, 3m, **6m**, 12m, 24m, 36m	**In GP consultation (point of care), post-consultation questionnaire**, 3m, **6m**, 12m
Candidate predictor of interest	Current pain intensity (0-10 NRS; (Von Korff et al., 1992))	
Timing of predictor measurement	1. Point of care 2. Post-consultation questionnaire	
Outcome of interest	Patient global rating of change at 6 months (Completely recovered/Much improved/ Improved vs Same/Worse/Much Worse; (van der Windt et al., 1998))	

Abbreviations: GP, general practitioner; IQR, Inter-quartile Range; NRS, numerical rating scale; PMR, Polymyalgia rheumatica; POST, The Primary care Osteoarthritis Screening Trial; PRO-GRES, The Prognostic Research Study; RCT, Randomised Controlled Trial; SD, standard deviation; SpA, Spondyloarthritis.

*Data collection points indicated in bold are the collection points used for this analysis

4.2.2 Predictor measurement

The predictor of interest was current pain intensity in patients presenting to primary care with non-inflammatory musculoskeletal disorders, which has previously been found to be a predictor of unfavourable episode outcomes in several previous primary care studies (Mallen et al., 2007).

The focus was on the predictor-outcome association between an unfavourable outcome at 6 months and current pain intensity (0-10 numerical rating scale (NRS); 0=no pain (Von Korff et al., 1992)). Pain intensity is a time-varying predictor, and we compared its association with unfavourable outcome on two occasions: (i) at the point of care as recorded by the GP, and (ii) recorded in a questionnaire by the patient sent within the week following point of care. Although the questionnaire was mailed within the week after the patients first visit to their GP, in both studies over a quarter of the questionnaires were returned at least a month after their consultation.

In both POST and PROGRESS the post-consultation questionnaires and the instructions to GPs measured current pain intensity in the same standardised format with the same anchors: "How would you rate your pain on a 0-10 scale **at the present time**, that is **right now**, where 0 is "no pain" and 10 is "pain as bad as could be"?"

4.2.3 Outcome measure

A primary concern of patients reporting to a GP with pain is whether their pain will improve in the future, and therefore, the outcome of interest in these analyses was the self-reported patient global rating of change recorded in the 6-month post-consultation questionnaire. For pragmatic reasons, and to enable the models to be easily compared, the categorical responses were dichotomised into having experienced a favourable outcome (completely recovered, much improved or improved) or an unfavourable outcome (same, worse or much worse) (as previously used in van der Windt et al. (1998)).

4.2.4 Statistical analysis: predictor-outcome associations

Participants were eligible for inclusion in the current analyses if they returned their questionnaire, consented to the use of medical records (such that their point of care information was available), and were successfully followed up at 6 months.

Logistic regression models were fitted to estimate the predictor-outcome association between an unfavourable outcome at 6 months and pain intensity rating when recorded (i) at the point of care (i.e. intended point of using the prognostic results), and then (ii) in the post-consultation questionnaire. Separate models were fitted within the PROG-RES and POST datasets. Adjustment factors within all the models were age (as a linear term), gender and general practice, as these were all considered to be established prognostic factors. Pain rating was included within the models as a continuous variable, and its association with the outcome was included as a linear term. Only patients with complete predictor information at the point of care and the questionnaire, with outcome information available at 6 months, were included to ensure all analyses were comparable. Within the POST dataset, the models also included treatment arm as an additional adjustment factor, to account for any differences in outcomes between the treatment and control groups within the study. The adjusted predictor-outcome association estimates (odds ratios (OR)) and 95% confidence intervals (CI) from the point of care model were compared with those from the questionnaire model, for each of PROG-RES and POST datasets separately. The models were transformed back to the absolute risk scale and predicted probabilities were plotted as an illustration. These were calculated from the model coefficients as shown below in equation **(4.1)** and **(4.2)**.

$$\begin{aligned} logit(p_i) = {} & \alpha + \beta_1 Pain_i + \beta_2 Age_i + \beta_3 Gender_i + \beta_4 General\ Practice_{1,i} \\ & + \beta_5 General\ Practice_{2,i} + \cdots \\ & + \beta_{3+S} General\ Practice_{S,i}\ (+\beta_{4+S} Treatment\ Arm_i) \end{aligned} \tag{4.1}$$

$$\mathrm{p_i} = \frac{\exp(y_i)}{1 + \exp(y_i)} \tag{4.2}$$

where p_i is the predicted probability for individual i of having the event and S is the number of general practices in the dataset. The confidence intervals for the predicted probabilities were calculated as

$$95\%\ \mathrm{CI} = logit(p_i) \pm 1.96 \times \text{Standard Error} \tag{4.3}$$

To enable the probabilities to be plotted, specific values for the adjustment factors needed to be chosen. Hence, the plots represented probabilities for a female patient from a randomly selected practice with the mean age in the dataset.

4.2.5 Statistical analysis: prognostic model performance

Next, each of the logistic regression models fitted were considered as prognostic models, such that they were to be (hypothetically) used for predicting individual outcome risk in new individuals. This allowed the focus to be on their overall predictive performance, and in particular enabled the comparison of the performance of the models fitted at the point of care with the models fitted using the questionnaire information. The performance measures examined were the Akaike's Information Criterion (AIC) and discrimination.

The AIC measures the relative goodness of fit of a model, considering both the statistical goodness of fit and the number of parameters used. The formula for the AIC is

$$AIC = 2K - 2\ln(likelihood) \qquad \textbf{(4.4)}$$

where K is the number of parameters in the model and $\ln$ is the natural logarithm. The model with the lowest AIC is the preferred model, but as a rule of thumb, two models are essentially equivalent if the difference in their AICs is less than 3 units (when the sample size is greater than 256) (Hilbe, 2011). The AIC was used here (unlike in Chapter 2), as the AIC is a measure that can be used to compare the fit of multiple models and the primary aim in this chapter was to compare the model using the predictor measured at the point-of-care with the model using the predictor measured a few days later.

The discrimination was measured using the concordance index (C-statistic) (Hanley and McNeil, 1982), which is the ability of the model to differentiate between those who do or do not experience the outcome of interest; in this case, it is the ability of the model to differentiate between those who do or do not experience an unfavourable outcome at 6 months. The C-statistic is the probability that for any randomly selected pair of individuals, one with an unfavourable outcome and one without, the model assigns a higher probability to the individual with the unfavourable outcome. For logistic regression models, as used in this study, the C-statistic is identical to the Area Under the receiver operating characteristic Curve (AUC). A C-statistic of 0.5 indicates that the model is no better than chance and a value of 1 indicates that the model perfectly classifies the individuals.

Calibration tells us the amount of agreement between the observed outcomes and the predictions and is always likely to be high in the dataset the model was developed in. Therefore, comparing the calibration of the two different models in the same dataset does not give a meaningful evaluation of the performance of the two models and hence no measure of calibration was assessed here.

4.2.6 Statistical analysis: sensitivity analyses

Sensitivity analyses were performed to evaluate assumptions made during the main analyses.

4.2.6.1 Non-linearity

To account for the possibility of a non-linear relationship between pain intensity rating and an unfavourable outcome, logistic regression models were fitted allowing for fractional polynomials (Royston et al., 1999) (using the *fp* command in Stata). Fractional polynomials are used in regression models as an alternative to regular polynomials to allow for flexible parameterisation of continuous variables. Fractional powers (-2, -1, -0.5, 0, 0.5, 1, 2, 3) of pain intensity rating were considered, adjusting for age, gender, general practice and treatment arm.

4.2.6.2 Missing data

The main analyses only included patients with complete data at all time points (i.e. point of care, questionnaire and outcome). As there was missing data at both the point of care and the questionnaire, the use of a complete-case analysis could potentially bias the predictor-outcome association and cause a reduction in the statistical power of the analyses. As the aim of this chapter was to compare the associations and performance of the two models when using predictors measured at different time points rather than developing a prediction model to be used in practice, it was deemed not necessary to multiply impute the missing data. However, as a sensitivity analyses and to evaluate the potential impact of the missing data on the reported associations , the main analyses were repeated under two conditions:

- Including all patients in the point of care model with data at that time point, regardless of whether they had missing information in the questionnaire or not
- Including all patients in the questionnaire model with data at that time point, regardless of whether they had missing information at the point of care or not

4.2.6.3 Interaction between pain and treatment arm

The presence of an interaction between pain intensity rating and treatment arm were tested in the point of care and questionnaire models by including an interaction term of pain intensity rating with treatment arm (POST data only), as responders who received treatment may have had a different relationship between their pain ratings and their outcome than those who did not receive treatment.

4.2.6.4 Random effects modelling

Within the main analyses, general practice was accounted for by including practice in the models as a categorical variable. As a sensitivity analysis, the assumption that the relationships between pain intensity ratings (at point of care and questionnaire measurement) and outcome were similar across general practices was assessed in POST. Multilevel logistic regression models, using the *runmlwin* command in Stata (Leckie and Charlton, 2013), were fitted including a random effect coefficient at the practice level. This relationship was not investigated within PROG-RES as five practices was deemed too few to reliably fit a multilevel model.

4.3 Results

4.3.1 Data description

Of 650 potentially eligible patients mailed a questionnaire in PROG-RES, 424 (65.2%) returned it, consented to medical record review and had information at their consultation recorded, and 296 (45.5%) were successfully followed up at 6 months and had completed data at the point of care and baseline. The corresponding figures for POST were 2042, 1230 (60.2%), and 756 (37.0%) (flowcharts provided in Figure 4.1 and Figure 4.2).

Figure 4.1: Participant flow - PROG-RES

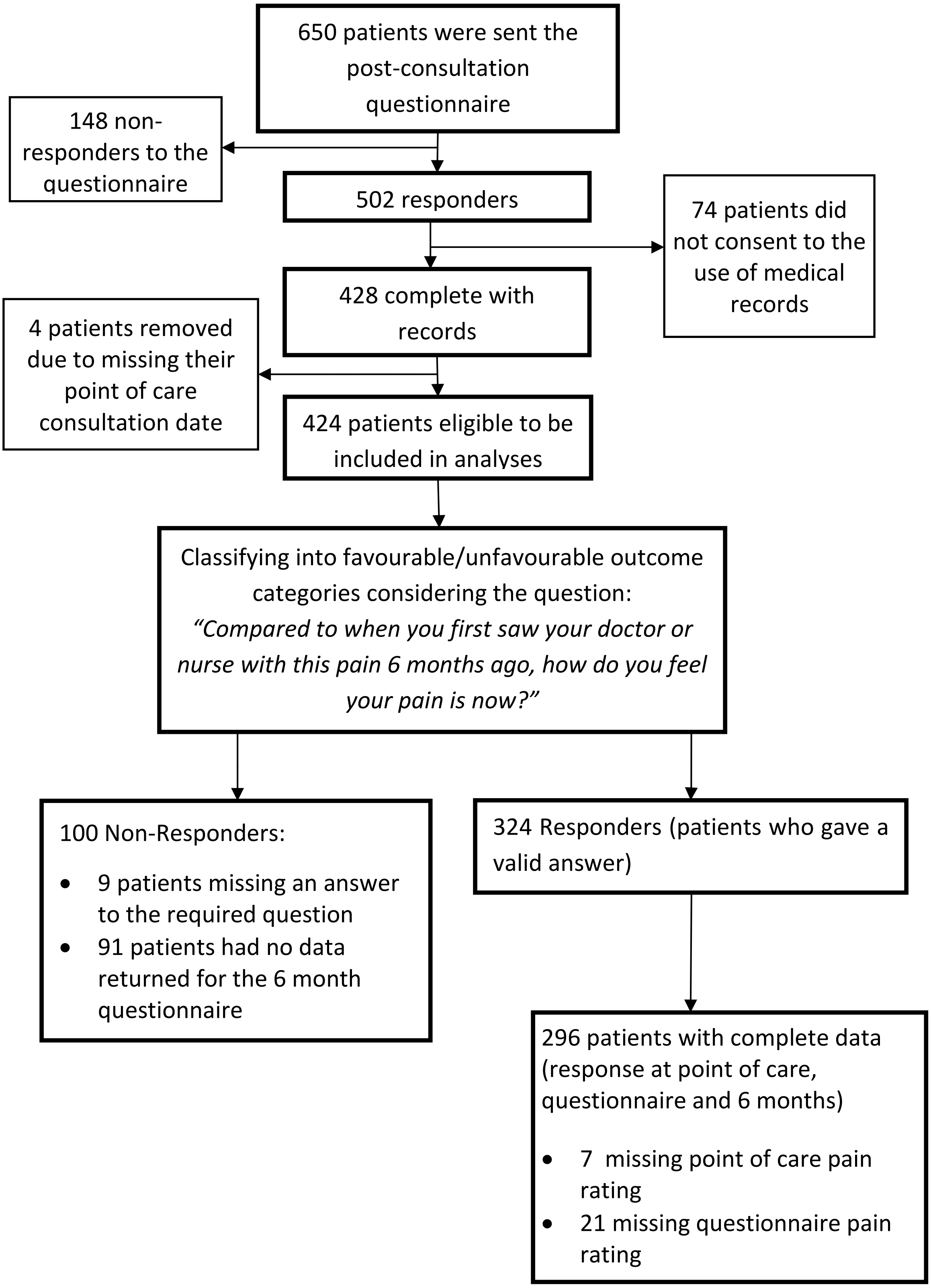
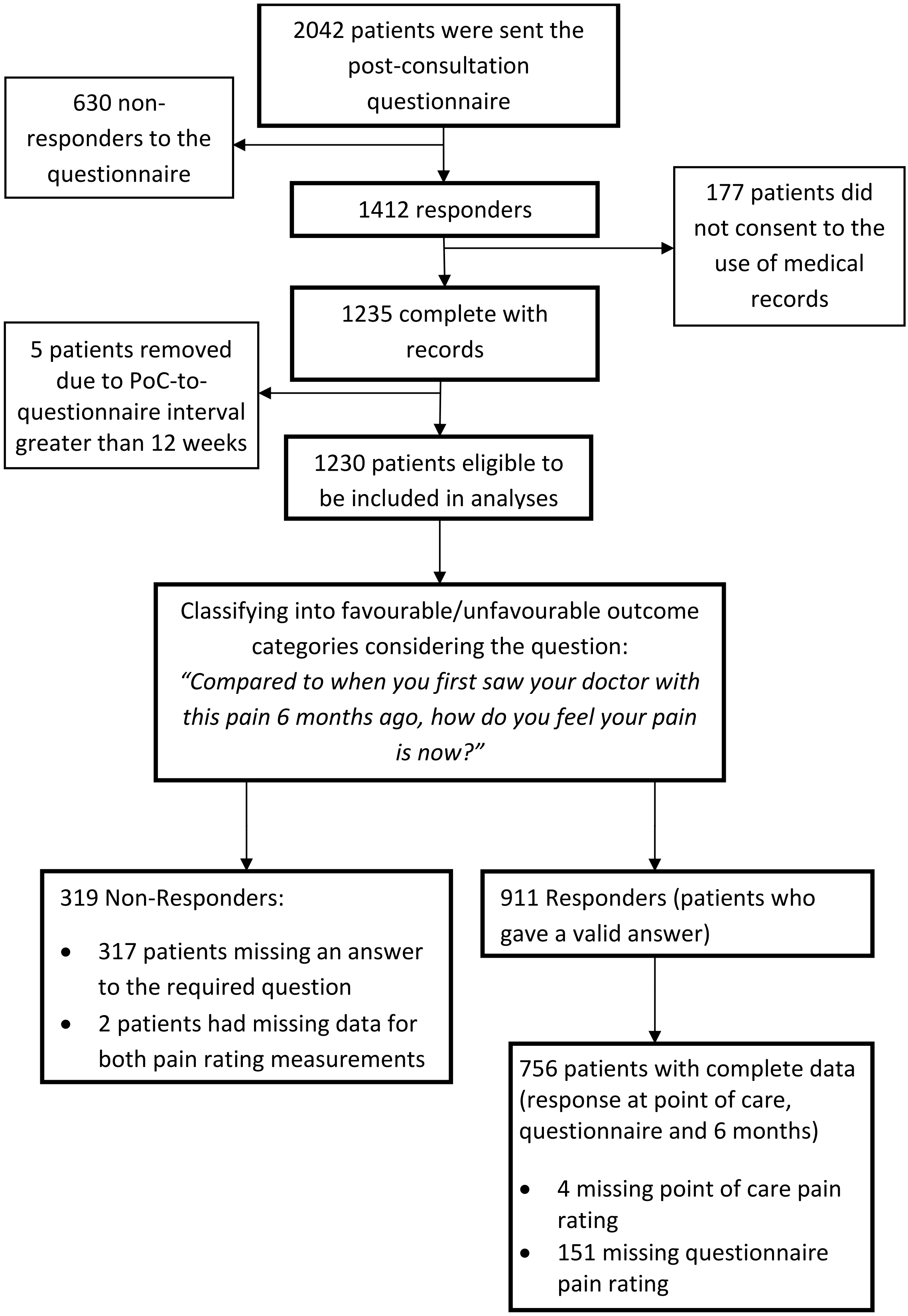

650 patients were sent the post-consultation questionnaire
148 non-responders to the questionnaire
502 responders
74 patients did not consent to the use of medical records
428 complete with records
4 patients removed due to missing their point of care consultation date
424 patients eligible to be included in analyses
Classifying into favourable/unfavourable outcome categories considering the question:
"Compared to when you first saw your doctor or nurse with this pain 6 months ago, how do you feel your pain is now?"
100 Non-Responders:
9 patients missing an answer to the required question
91 patients had no data returned for the 6 month questionnaire
324 Responders (patients who gave a valid answer)
296 patients with complete data (response at point of care, questionnaire and 6 months)
7 missing point of care pain rating
21 missing questionnaire pain rating

Figure 4.2: Participant flow - POST


2042 patients were sent the post-consultation questionnaire
630 non-responders to the questionnaire
1412 responders
177 patients did not consent to the use of medical records
1235 complete with records
5 patients removed due to PoC-to-questionnaire interval greater than 12 weeks
1230 patients eligible to be included in analyses
Classifying into favourable/unfavourable outcome categories considering the question:
"Compared to when you first saw your doctor with this pain 6 months ago, how do you feel your pain is now?"
319 Non-Responders:
• 317 patients missing an answer to the required question
• 2 patients had missing data for both pain rating measurements
911 Responders (patients who gave a valid answer)
756 patients with complete data (response at point of care, questionnaire and 6 months)
• 4 missing point of care pain rating
• 151 missing questionnaire pain rating

Potentially eligible patients lost to follow-up at 6 months did not differ by age or gender but had slightly higher mean pain ratings at the point of care and baseline in both PROG-RES (point of care: responders mean (SD) 6.0 (2.2) vs non-responders 6.4 (2.2); questionnaire: 5.5 (2.6) vs 5.6 (2.5)) and in POST (6.2 (2.1) vs 6.6 (2.0); 5.3 (2.6) vs 5.8 (2.6)).

Table 4.2 shows the characteristics of those with complete data included in the main analyses. The proportion reporting an unfavourable outcome at 6 months was 48.7% in PROG-RES and 54.5% in POST. In both studies, a significant fall in pain intensity ratings between point of care and the post-consultation questionnaire measurement was observed, tested using a paired t-test (PROG-RES: mean (SD) 5.9 (2.2) vs 5.5 (2.6); mean difference (SD): 0.42 (0.17), *P*=0.006; POST: 6.2 (2.1) vs 5.3 (2.6); 0.89 (0.09), *P*<0.001).

Table 4.2: Sample characteristics of the datasets

	PROG-RES	**POST**
Participants eligible for inclusion in main analyses	296	756
Age (years): mean (SD)	64.8 (9.8)	65.8 (9.9)
Male: n (%)	120 (40.5)	339 (44.8)
Current pain intensity at point of care (0-10): mean (SD)	5.9 (2.2)	6.2 (2.1)
Current pain intensity in questionnaire (0-10): mean (SD)	5.5 (2.6)	5.3 (2.6)
Interval between point of care and return of questionnaire (days): median (IQR)	17 (13, 27)	21 (16, 30)
Interval between point of care and return of questionnaire (days): range	6, 75	3, 81
Unfavourable outcome at 6 months: n (%)	144 (48.7)	412 (54.5)

Abbreviations: GP, general practitioner; IQR, Inter-quartile Range; NRS, numerical rating scale; PMR, Polymyalgia rheumatica; POST, The Primary care Osteoarthritis Screening Trial; PRO-GRES, The Prognostic Research Study; RCT, Randomised Controlled Trial; SD, standard deviation; SpA, Spondyloarthritis.

4.3.2 Preliminary analyses

In PROG-RES, a significant mean reduction in pain score overall between point of care and questionnaire was observed in the group who went on to experience a favourable outcome at 6 months (mean reduction (SD): 1.12 (0.24), *P*<0.001) but not in those with an unfavourable outcome (-0.32 (0.21), *P*=0.932). Similar mean reductions were seen in POST (favourable outcome: 1.59 (0.15), *P*<0.001; unfavourable outcome: 0.31 (0.12), *P*=0.004).

4.3.3 Examination of predictor-outcome associations

At the point of care, there was only a weak and non-statistically significant association found between pain intensity and an unfavourable outcome in both PROG-RES (adjusted OR (95% CI): 1.06 (0.95, 1.18)) and POST (1.04 (0.96, 1.12)) (Table 4.3). To translate this to absolute risk, the fitted models were transformed back to the probability scale. As an illustration, Figure 4.3 shows that for a female patient aged 65 (PROG-RES) or 66 (POST) from practice 30 in POST or 4 in PROG-RES, there was little change in the predicted probability of an unfavourable outcome as pain intensity at point of care increased, in both POST and PROG-RES.

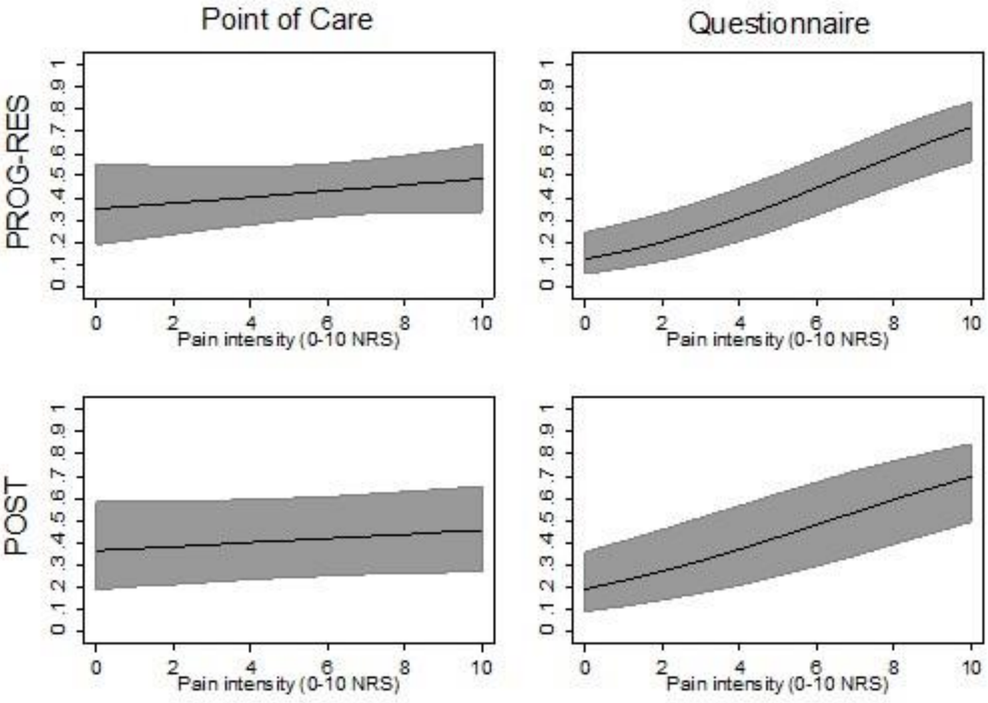
In contrast, the models estimating the independent association between the questionnaire pain rating and outcome found a stronger and statistically significant relationship. In PROG-RES, for each unit increase in pain rating the odds of an unfavourable outcome increased by 34% (adjusted OR (95% CI): 1.34 (1.20, 1.48)) and in POST, for each unit increase in pain, the odds of an unfavourable outcome increased by 26% (1.26 (1.18, 1.34)) (Table 4.3). Transforming the models back to the absolute risk scale, Figure 4.3 shows that for a patient with the same adjustment factors as above, the predicted probability of an unfavourable outcome increased at similar rates as pain intensity at the questionnaire increased, in both datasets. The change in predicted probability is far steeper for the questionnaire models than for the point of care models. For example, in POST, the predicted probability for an individual with a pain score of 8 was 0.59 when using the questionnaire model but 0.44 when using the point of care model.

Table 4.3: Predictor-outcome association between a one unit increase in pain intensity and an unfavourable outcome

	Intended moment of using the prognostic results	**PROG-RES** (n=296)	**POST** (n=756)
		Adjusted OR* (95% CI)	Adjusted OR* (95% CI)
Current pain intensity (0-10 NRS) measured at:			
Point of care	Yes	1.06 (0.95, 1.18)	1.04 (0.96, 1.12)
Post-consultation questionnaire	No	1.34 (1.20, 1.48)	1.26 (1.18, 1.34)

*Adjusted for age, gender, general practice (and treatment allocation - POST only)

Figure 4.3: Predicted probability (95% CI) of an unfavourable outcome at 6 months by pain intensity rating estimated from the point of care and questionnaire models (female patient aged 65 (PROG-RES) or 66 (POST) from practice 30 in POST or 4 in PROG-RES).



4.3.4 Examination of prognostic model performance

Table 4.4 shows the performance measures for the fitted models from Table 4.3. The AIC for the questionnaire models was lower in both datasets than the point of care models, with a difference of 32 units in PROG-RES and 50 units in POST, suggesting that the models fitted using the pain score measured in the questionnaire had a better overall fit than the models using the pain score recorded at the point of care. The C-statistics were higher for the questionnaire models than for the point of care models in both datasets, and thus the discrimination was larger when pain intensity was measured in the questionnaire. This concurs with the larger odds ratio estimates for pain intensity from the questionnaire than the point of care.

Table 4.4: Measures of model performance at the point of care and questionnaire in PROG-RES and POST

	Intended moment of using the prognostic results	**PROG-RES**		**POST**	
		AIC	C-statistic	AIC	C-statistic
Current pain intensity (0-10 NRS) measured at:					
Point of care	Yes	421.8	0.57 (0.51, 0.64)	1066.2	0.66 (0.62, 0.70)
Post-consultation questionnaire	No	389.8	0.69 (0.63, 0.75)	1015.8	0.72 (0.68, 0.76)

4.3.5 Sensitivity analyses

4.3.5.1 Non linearity

When accounting for a potential non-linear relationship between pain ratings and an unfavourable outcome using fractional polynomial models, the best fitting relationships (defined by the deviance) between pain and outcome in PROG-RES were: (i) a cubic relationship when pain was rated at the point of care, and (ii) a log-linear relationship when pain was from the questionnaire. The best fitting relationships for POST were: (i) a log-linear relationship when pain was rated at the point of care, and (ii) a linear relationship when pain was from the questionnaire. Thus, a more complex relationship was identified at the point of care than at the questionnaire for both datasets. Nonetheless, differences in AIC were small between the models that allowed for non-linear associations and the previous models that assumed linear associations.

4.3.5.2 Missing data

In PROG-RES there were only 7 patients missing the point of care rating and 21 patients missing the questionnaire score (Figure 4.1). In POST these figures were 4 and 151 (Figure 4.2). In sensitivity analyses deriving models including the patients missing pain ratings at either the point of care or at questionnaire the strength of associations between pain intensity and outcome did not change from those found in the main analyses. The OR (95% CI) in PROG-RES for the point of care model was 1.06 (0.96, 1.18) which included 303 patients, for the questionnaire model the OR (95% CI) was 1.34 (1.20, 1.48) and included 317 patients. In POST the corresponding figures were 1.04 (0.96, 1.12), n=757 and 1.24 (1.17, 1.31), n=904.

4.3.5.3 Interaction between pain and treatment arm

No strong evidence of an interaction between treatment arm and pain intensity ratings was found with the estimated odds ratios for the interactions being very close to one and their confidence intervals crossing one. The OR (95% CI) for the interaction term between treatment arm and point of care rating was 0.92 (0.78, 1.08) and for the interaction term between treatment arm and questionnaire pain rating was 0.92 (0.81, 1.06), in the POST dataset.

4.3.5.4 Random effects modelling

The odds ratios and C-statistics did not change markedly when modelling general practice using random effects, as can be seen in Table 4.5. Infact, for PROG-RES, the odds ratios and confidence intervals did not change at all. Overall, the same pattern of findings was observed and hence it was decided to model as fixed effects in the main analyses for simplicity.

Table 4.5: Predictor-outcome associations and C-statistics in PROG-RES and POST at the point of care and post-consultation questionnaire when fitting general practice as a random effect

	PROG-RES		**POST**	
Current pain intensity (0-10 NRS) measured at:	OR (95%CI)	C-statistic	OR (95%CI)	C-statistic
Fixed-effect model				
Point of care	1.06 (0.95, 1.18)	0.57	1.04 (0.96, 1.12)	0.66
Post-consultation questionnaire	1.34 (1.20, 1.48)	0.69	1.26 (1.18, 1.34)	0.72
Random-effect model				
Point of care	1.06 (0.95, 1.18)	0.54	1.05 (0.98, 1.13)	0.60
Post-consultation questionnaire	1.34 (1.20, 1.48)	0.68	1.26 (1.19, 1.34)	0.69

4.4 Discussion

4.4.1 Principal findings

The findings in this chapter illustrate how the magnitude of predictor-outcome associations (prognostic factor effects) and prognostic model performance can depend on *when* and/or *how* time-varying predictors are measured. In this example of patients presenting with musculoskeletal pain to general practice, associations between outcome risk and pain intensity recorded at the intended moment of use were lower in magnitude than those associations derived from a self-complete questionnaire mailed to patients up to one week later. The findings were replicated in two datasets with similar measurements, strengthening the belief that similar findings are likely across a range of painful non-inflammatory musculoskeletal disorders. Despite many published studies of musculoskeletal pain in primary care (Mallen et al., 2007), very few report the collection of time-varying predictor information by the GP at the initial point of care (Von Korff, 2013). When a later time is used, and/or with a different measurement method, the study's predictor-outcome associations and prognostic model performance may be misleading, and thus it could signal that the study is at high risk of bias and not applicable for its intended purpose.

4.4.2 Explanation for the findings

Several phenomena may contribute to the observed discrepancy in predictor-outcome associations at the point of care and at a later time-point. Firstly, the timing of predictor measurement may be critical. For example, most musculoskeletal disorders follow an episodic course and therefore, as would be expected, patients in POST and PROG-RES were likely to consult when their pain was more severe than usual. This creates the conditions for regression to the mean following the point of care (Davis, 1976, Whitney and Von Korff, 1992). An initial reduction in group-average pain intensity rating within the first few days following primary care consultation has been consistently observed for acute, recurrent, and chronic low back pain (Artus et al., 2014, Coste et al., 1994, Costa et al., 2012, Roland and Morris, 1983). A similar pattern is likely across other non-inflammatory regional musculoskeletal pains. Although regression to the mean was evident within this study, the whole group mean was lower at the post-consultation questionnaire than at the point of care and so regression to the mean does not, therefore, provide a full explanation for the findings.

The differences found in the strengths of the predictor-outcome associations could also relate to differences in measurement methods. At the point of care, pain intensity measurement was verbally administered and recorded by the physician in a face-to-face consultation. Although in both studies physicians were given guidance on how to gather this information, we cannot know the extent to which physicians recorded their judgements of patients' pain. Physician ratings tend to systematically under-estimate patients' own ratings of pain (Mantyselka et al., 2001, Staton et al., 2007). Assuming that patients' pain ratings were elicited and faithfully recorded at the point of care, it is nevertheless possible that a form of end-aversion bias (Streiner et al., 2014) may operate in the clinical encounter, i.e. patients avoid reporting pain at either end of the severity scale in fear of being judged undeserving or exaggerating (although evidence from this study suggests this may be true of the lower end of the scale but not of the upper end of the scale).

The literature review in Chapter 3 found that many published prediction models included predictors that are likely to be susceptible to measurement error. The example provided in this chapter is no exception, such that – even if the setting and method of measurement were consistent – the predictor-outcome associations may not agree simply by chance variation. Further, if the measurement error was largest at the point-of-care, then the observed predictor-outcome association may be more biased at the point of care, than observed when measured at a later time-point. If measurement error was present, it is likely that in this situation it would be differential measurement error, and the impact of differential measurement could either exaggerate or underestimate the effect. Indeed, the predictor-outcome associations estimated in this study at the point of care and at questionnaire are both likely to be biased as no adjustment was made for measurement error due to insufficient information. Nevertheless, this is unlikely to account for the entire difference in magnitude of the estimated associations at point of care and questionnaire.

4.4.3 Limitations and future research

This chapter focused on predictor-outcome associations intended to be used at the point of care but derived using data collected after the point of care. It may be that a review appointment 2-3 weeks after the first consultation may be a better 'intended moment of use' for prognostic models in this field. Either way, it is clear from this example that the developed prognostic model needs to use data for time-varying predictors measured at the time of its intended use, as otherwise discrepant associations may be included. It may be considered that the model using the score at the later time point should be used as this is performing better, but this model would be misleading if used during the consultation. For example, if we look at the example prediction plots in Figure 4.3, if a patient visited their GP and reported a pain intensity score of 8, using the model developed with the score from the questionnaire would give this patient a predicted probability of experiencing an unfavourable outcome of 0.65. If the model developed using the point of care score was used, their predicted probability would be approximately 0.5.

While the problem highlighted is likely to extend to other commonly investigated predictors whose values are sensitive to the timing and mode of collection, this problem has only been demonstrated for one predictor and thus this remains to be evaluated more widely. Further research should assess whether similar findings are found with other time-varying predictors, and indeed in other clinical conditions and settings.

The impact of the timing of the predictor measurement will depend on exactly what the predictor is, and if the predictor is a test result, what type of test it is. The impact may depend on how long this test result is 'relevant' for, or simply how long it takes to receive the results of the test. For example, it may only take a few days to receive the results of a blood test, however it may take several weeks to receive the results of a MRI scan, and therefore it would not be possible to use the results of these tests at the initial meeting with the health care professional. Care should be taken when planning a study, or using results from a previously conducted study, around exactly when tests are requested, completed, and results received in relation to when they will be used in practice and how long these results are likely to be 'relevant' for.

Dependent error is also likely within this example, as a reduction in pain after the consultation (measured in the post-consultation questionnaire) is intrinsically going to be part of the patient's judgement at 6 months about whether or not they have improved, particularly because these were measured by the same method, and this bias will likely be greater the closer in time the post consultation questionnaire measurement is to the measurement of the outcome. This is a limitation of this particular example and the bias created by this limitation may be less likely to be encountered in other prognostic models.

The models in this chapter were not internally validated, and therefore no adjustment for optimism was made to the c-statistics provided, meaning the magnitude of the c-statistics are likely to be overstated. However, the concern here in this chapter was not with the absolute magnitude of these but the relatively poorer fit and discrimination of a model using a prognostic factor measurement obtained at the point of care (the intended moment of use for the model) than when obtained later.

A future study in which the same mode of data collection is used at the point of care and at post-consultation questionnaire (e.g. patient self-administered questionnaire) is needed to better understand the relative contribution of timing and mode of collection and therefore determine whether and how improved prediction is achievable at the point of care.

4.5 Recommendations

Based on the findings of this chapter, a list of recommendations for developing and using prediction models in relation to the timing of the predictors and the intended moment of use of the model are made in Box 4.1.

Box 4.1: Recommendations for improvements in the reporting of multivariable prediction models, based on findings of the review.

- Whether developing/validating a prediction model, or using a previously developed model in practice, consider exactly when the model is intended to be used in practice.
- When developing a prediction model, ensure any predictors to be used in the development of the model are measured at (or before) the intended moment of using the model.
- If using a previously developed model in practice, ensure the predictor values to be inputted into the model are measured at the same time as those used to originally develop the model.
- When considering previously published studies reporting predictor-outcome associations and prognostic models using time-varying predictors, assess the risk of bias and applicability based on when the predictors were measured in relation to the intended moment of using the model.

4.6 Conclusions

Irrespective of the underlying causes, the findings in this chapter imply the need for caution when applying predictor-outcome associations or existing prognostic models derived from prognosis research studies that record time-varying predictors at a different time and/or measurement method than is intended upon clinical application. This argument reinforces the need for clearly reporting the intended moment of use in prognostic research, and when the predictors were measured (Collins et al., 2015). Displacing the collection of time-varying predictor information from the intended moment (and mode) of use can result in differences in the magnitude of predictor-outcome associations, and the subsequent accuracy of prognostic model performance. In particular, predictors and models that appear to discriminate well in research studies, may fail to live up to those expectations when applied or externally validated at the intended moment of use. This concern is likely to be particularly justified when the outcome in some way incorporates the prognostic factor, when the interval between later measurement and outcome is short, and when the same mode of assessment is used to collect predictor and outcome information (Lash and Fink, 2003). Unless shown otherwise in validation studies using predictors measured at the clinically relevant time, previously developed prediction models that include time-varying predictors measured after the intended moment of use may overestimate individual risk of experiencing the outcome of interest, which also reinforces the need for external validation using data that reflects the intended moment of use, and clear reporting of differences between validation and development data (Moons et al., 2014).

4.7 Direction of the remainder of the thesis

The focus of the thesis so far has been on prognosis research in single studies. However, there is a growing demand for meta-analyses that utilise IPD from multiple prognosis research studies, as this may offer novel opportunities for the development and validation of clinical prediction models, or for prognostic factor research, that may not be possible with the individual studies alone (Riley et al., 2021b). The following two chapters will focus on the use of IPD meta-analyses for prognostic research. In Chapter 5, IPD will be used to validate existing prediction models that have been developed across several population groups for predicting stillbirth. In Chapter 6, methodology work will be undertaken to develop a method of calculating the power of IPD meta-analysis projects which have the aim of synthesising the IPD to examine prognostic factor effects.

5 External validation of prediction models for stillbirth using individual participant data (IPD) meta-analysis: the IPPIC study

5.1 Introduction

Previous chapters in this thesis have focused on prognosis research using a single study to either investigate potential prognostic factors or to develop a prediction model. A single study may be sufficient for prognosis research if the data are large enough, the outcome of interest is not rare and the prognostic factors (predictors) of interest are prevalent enough. However, this is often not the case, and so there has been an increasing interest in utilising IPD from multiple existing studies to increase the quantity and quality of data available, which can in turn improve the ability and power to examine the prognostic effects of a factor, or the development and validation of clinical prediction models. This chapter provides an applied example of using IPD to validate existing prediction models which have been developed to predict stillbirth in pregnancy.

The results of this chapter were published in *Ultrasound in Obstetrics & Gynaecology* (Allotey et al., 2022), and I am the joint first author, having undertaken all the statistical analyses.

5.1.1 Background

Stillbirth, defined by the World Health Organization (WHO) as fetal death after 28 weeks of gestation, accounts for around 3 million deaths worldwide annually (Organization, 2006).

In the UK, where stillbirth is defined as fetal death after 24 weeks, there were 3286 stillbirths in 2013 (4.2 per 1000 births) (Manktelow et al., 2015). This remains one of the highest rates in Europe, with little improvement in decades (Flenady et al., 2011). Stillbirth has now become a top priority on UK political and medical agendas. Many professional and parent advocacy groups have called for focused research to understand and prevent stillbirth, leading to the launch in November 2015 of a Government drive to halve the number of stillbirths and neonatal deaths in the UK. The Royal College of Obstetricians and Gynaecologists has launched its 'Each Baby Counts' campaign, with similar aims. The Every Newborn Action Plan aims to reduce the national stillbirth rate to 2 or fewer stillbirths per 1000 births by 2030.

Until recently, as many as two thirds of stillbirths occurred in 'low risk' women and were considered 'unexplained' and therefore unavoidable. In fact, rigorous investigation and classification can identify a likely cause for stillbirth in up to 85% of cases (Gardosi et al., 2005). When attempting to classify stillbirth according to the relevant condition at death, 43% of stillborn babies were found to be small for gestational age (SGA) and 9% had proven placental insufficiency (Gardosi et al., 2005). Screening for placental insufficiency alone is unlikely to adequately detect all women at risk of stillbirth, and further associations must be explored. Ongoing research into the causes of stillbirth has identified a number of factors known to be associated with the risk of fetal demise, including maternal age, parity, medical co-morbidities, ultrasound findings such as the uterine artery Doppler and biochemical markers, e.g. Pregnancy-associated plasma protein A (PAPP-A), alpha-Fetoprotein (AFP), soluble fms-like tyrosine kinase-1 (sFlt-1) and prothrombotic mutations. Notably, only few studies have reported on the combinations of these markers and absolute risk for individual pregnancies.

Importantly, a third of stillbirths occur in babies over 36 weeks' gestation and up to two thirds after 34 weeks. At these gestations, delivery could be considered if the risk of stillbirth outweighed the risks of premature delivery. At present, induction of labour is offered as a preventative intervention to women considered to be at high risk of stillbirth where that balance of risk and benefit is reached – for example, post term pregnancies between 41-42 weeks or women with pre-existing or gestational diabetes at 37-40 weeks. This represents basic screening and treatment to prevent stillbirth but clearly, with many women thought to be low risk still suffering devastating pregnancy loss, the performance of the de facto screening tests currently applied for stillbirth is unacceptable. Better prediction of stillbirth and individualisation of risk is a key priority for stillbirth research identified by the UK Stillbirth Priority Setting Partnership (Heazell et al., 2015).

Previous systematic reviews using aggregate data meta-analysis have evaluated various risk factors separately or in combination for prediction of stillbirth and perinatal mortality (Conde-Agudelo et al., 2015). These reviews represent the best available evidence relating to prediction of stillbirth and yet are limited by the heterogeneity of reporting and data in the primary, largely observational, studies. IPD meta-analysis offers several advantages compared to standard methodology that are of relevance to the prediction of stillbirth.

There is significant international variation in the classification and definition of stillbirth, as WHO defines stillbirth as fetal loss after 28 weeks, but most developed countries use a gestational cut off from 20-24 weeks. There are some countries where perinatal deaths are counted together. It follows that the outcome of stillbirth reported in studies, which would seem to be a binary outcome, is highly heterogenous and context dependent. When added to heterogeneity in the population characteristics, timing of tests and cut-offs this can present serious limitations to aggregate meta-analysis. With access to all the patient data, heterogeneity can be reduced by IPD through uniform cut-offs and definitions being applied across all included patients.

Prediction models must be externally validated in datasets not used for model development before they can be considered for use in clinical practice, and the lack of external validation is a major factor in the lack of prediction models available for clinical use. Using raw individual participant data in multiple datasets allows existing models to be externally validated.

A large scale IPD meta-analysis is particularly appropriate for the validation of prediction models for stillbirth, as it is an uncommon outcome with significant health implications for families and economic effects for health systems, and combining IPD from multiple sources allows greater power and ability to examine prediction model performance across multiple settings.

5.1.2 Chapter objectives

The aim of this chapter is to validate previously identified existing prediction models across several population groups for predicting stillbirth, using IPD from the IPPIC (International Prediction of Pregnancy Complications Network) collaborative data.

The specific objectives are to:

- Identify cohorts available within IPPIC for use in the validation of the prediction models

- Summarise the distributions of the linear predictors and predicted probabilities for each identified prediction model in the individual cohorts available
- Assess and compare the predictive performance of the models using discrimination and calibration statistics
- Use decision curve analysis to assess and compare the clinical utility of the prediction models.
- Pool and summarise the model performance across datasets using meta-analysis

5.2 Methods

5.2.1 Data

The International Prediction of Pregnancy Complications Collaborative Network (IPPIC) is a collaborative network of investigators from global research groups that have undertaken studies on clinical characteristics and biochemical and ultrasound markers in complications in pregnancy. The network includes 125 researchers from 25 countries who have contributed IPD for over two million pregnancies.

Any prospective or retrospective cohort study, cohorts nested in randomised controlled trials and birth and population-based cohorts were considered for inclusion in the IPD dataset if they provided information on clinical, biochemical or ultrasound variables with information on perinatal mortality outcomes. Only singleton pregnancies were included in the analyses.

5.2.2 Models to validate

A systematic review of primary studies reporting on prediction models for stillbirth was conducted prior to my involvement in the study, to identify models for external validation in the IPPIC cohorts by researchers in the IPPIC network (Allotey et al., 2022). The systematic review included a comprehensive search of MEDLINE, Embase, DH-DATA (database of the Department of Health's Library and Information Services) and AMED (Allied and Complementary Medicine Database) databases from inception to December 2020 to identify all studies that developed or updated prognostic models for stillbirth for use at any time during pregnancy. Reference lists of relevant articles and systematic reviews were also hand searched to identify potentially eligible studies. The search included terms for stillbirth, intrauterine fetal death and perinatal mortality. Study selection was completed independently by two researchers. The complete search strategy is provided in Allotey et al. (2022) along with more details on the search process and data extraction.

Each model was validated using IPD from studies that contain all predictors in the model and the relevant outcome (stillbirth and gestational age), with at least two outcomes occurring in the cohort. Ideally, as noted in the previous chapter, the time of measurement of the predictors and outcomes should match the setting in which the model was developed; however, time of predictor and outcome measurement was not always available or may have differed slightly. This is discussed further in the Discussion in Section 5.4.

5.2.3 Imputation of missing predictor data

Only the IPPIC cohorts that recorded all the predictors included in each prediction model were used for validation of that particular prediction model. If a predictor from a prediction model was missing for all the participants in the cohort, it was considered systematically missing, and therefore this cohort was not used to validate that particular model. Second trimester measurements of BMI and weight were used for model validation if first trimester values were missing in the cohort, and vice versa. Any partially missing predictors or outcome values missing for <95% of individuals in a cohort were multiply imputed under the missing at random (MAR) assumption using multiple imputation by chained equations (MICE) (Jolani et al., 2015, Resche-Rigon and White, 2016). Linear regression was used to impute for approximately normally distributed continuous variables, logistic regression for binary variables, and multinomial logistic regression for categorical variables. Non-normally distributed continuous variables were transformed before imputation. Multiple imputation was carried out for each individual cohort separately rather than all IPD cohorts combined. Fifty imputed datasets were generated for each cohort. It was recognised that more imputations were needed than the 10 used in Chapter 2, and ideally more than 50 imputed datasets would have been generated, equal to the percentage of missing observations as suggested by White et al. (2011). However, due to time constraints and the amount of computational power available, a practical approach of generating 50 imputed datasets was taken.

All relevant predictors, for all prediction models to be validated using that particular cohort, were identified and imputed at the same time to avoid imputing values for each different prediction model separately. This was to ensure a coherent set of imputed datasets, to be used consistently in all analyses, regardless of the prediction model being validated.

Other predictors that were available within the cohort were also included in the imputation models as auxiliary variables. Auxiliary variables are believed to be correlated with the variables with missing data, or are associated with missingness, and can improve the imputed estimates and make the MAR assumption more plausible, regardless of whether they include missing data themselves (Enders, 2008). This has been shown to be particularly important when there is a high proportion of missing data (Johnson and Young, 2011), as there is here. The more correlated with the missing data the auxiliary variables are, the more they can reduce the bias and standard errors, but including them can do no harm irrespective of how correlated they are (Johnson and Young, 2011).

Missing outcomes were imputed in the same way and at the same time as missing predictor values, using as many available variables as possible in the imputation model. Imputation checks were completed by looking at histograms, summary statistics and tables of values across imputations, as well as checking trace plots for convergence issues.

5.2.4 External validation of models

Each model was validated separately by applying the model equation to each participant in the cohort to calculate the linear predictor for that participant (LP_i, the estimated logit risk from the model for individual i), which is the value of the linear combination of predictors in the model equation for individual i. Also, the predicted probability of stillbirth (inverse logit transformation of LP_i) was calculated for each individual. For each prediction model, the distribution of LP_i values were summarised for each cohort. The predictive performance of the models was assessed using discrimination and calibration statistics (Altman et al., 2009). Performance statistics were calculated in each imputed dataset and then averaged across imputations using Rubin's rules to obtain one estimate and standard error (SE) for each performance statistic in each cohort (Rubin, 1987), with the exception of the c-statistic, as described in the following section.

5.2.4.1 C-statistic (discrimination)

The discrimination, as measured by the concordance statistic (C-statistic), gives the probability of a randomly selected woman who had the outcome (stillbirth) having a higher predicted probability than a randomly selected woman without the outcome. The C-statistic is equivalent to the area under the receiver operating characteristic (ROC) curve, and was calculated (along with its SE) using non-parametric ROC analysis in Stata using the *roctab* command. It is likely that the distribution of the C-statistic was not normal since it is a proportion and therefore bounded by the value 1. Hence, the logit scale was used to combine C-statistics across imputations (Snell et al., 2018). The SE for the logit(C-statistic) were calculated using the following formula (Debray et al., 2017) (the delta method):

$$\mathrm{SE}\big(\mathrm{logit}(C)\big) = \frac{\mathrm{SE}(C)}{C(1-C)}$$

5.2.4.2 Calibration-in-the-large

The calibration-in-the-large measures the extent to which the model predictions are systematically too low or too high across the cohort, with an ideal value of 0. The estimate of the calibration-in-the large and its SE were calculated by fitting the calibration model

$$\mathrm{logit}(p_i) = \alpha + \beta(LP_i)$$

where α is the estimate of the calibration-in-the-large, when $\beta = 1$ (fitted using an offset term) and i refers to a participant.

5.2.4.3 Calibration slope

The calibration slope is the slope of the regression line fitted between predicted and observed risk probabilities on the logit scale. It indicates whether there is agreement between observed outcomes and predictions across the range of predicted risks. The calibration model was fitted,

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

were β is the estimated calibration slope. Ideally, the calibration slope would be equal or very close to 1 for good calibration. However, a slope < 1 indicates overfitting of the model, whereas a slope > 1 indicates underfitting.

5.2.4.4 Calibration plots

Model calibration was also visually assessed using calibration plots in each dataset separately, showing the observed (O) versus expected (E) probabilities for groups of participants. Average predicted probabilities were obtained for individuals by pooling their linear predictor values across imputed datasets using Rubin's rules, and then transforming to the probability scale. Participants were grouped into tenths of this average predicted probability using deciles of the predicted values, and O versus E was plotted for each of the ten groups. A lowess smoother curve was applied to show calibration across the entire range of predicted probabilities at the individual level (i.e. without categorisation). Calibration plots are presented for datasets with at least 100 events as it was decided in collaboration with the research team that plots for datasets with too few events would not prove meaningful and it has been suggested that a minimum of 100 events (and 100 nonevents) are required for external validation samples (Vergouwe et al., 2005). Due to the very low probability of an event occurring, the calibration plots were presented for the lower range of the observed and expected probabilities (0 to 0.02) to enable the detail of the plots to be viewed more clearly. Calibration plots were plotted using the *pmcalplot* command in Stata (Ensor et al., 2020).

5.2.4.5 Pooled model performance

Performance measures of prediction models that were validated in more than two independent cohorts were summarised using a random effects meta-analysis to calculate a summary estimate for the model's discrimination and calibration performance. Random-effects is preferred because it was expected that there would be significant heterogeneity between cohorts due to the differences in the background populations and selection procedures. Random-effects meta-analysis allows us to quantify and assess the heterogeneity in performance across cohorts, settings, and clinically relevant subgroups (e.g. defined by treatment and populations) and predict model performance in other similar settings using approximate 95% prediction intervals (Riley et al., 2011). The random-effects model for a performance measure can be written as

$$Y_k \sim Normal(\mu_k, \sigma_k^2)$$

$$\mu_k \sim Normal(\mu, \tau^2)$$

where k refers to the dataset. The model assumes normality of the within-study and between-study performance statistic. Based on the results of a simulation study (Snell et al., 2018), the C-statistic was pooled on the logit scale, as the simulation study suggested this to be a more appropriate scale for pooling C-statistics in a meta-analysis. The calibration slope and calibration-in-the large were pooled on their original scale. Model performance was summarised for each statistic using the average estimate and corresponding 95% confidence interval calculated using the Hartung-Knapp-Sidik-Jonkman approach to account for uncertainty in variance estimates (Hartung and Knapp, 2001, Langan et al., 2019). REML estimation was used to fit the random-effects model.

Between-study heterogeneity (τ^2) and the proportion of variability due to between-study heterogeneity (I^2) (Higgins et al., 2003) were summarised. The approximate 95% prediction intervals, for potential predictive performance in a new study, were calculated using the approach of Higgins et al. (2009).

Model performance across cohorts was also shown graphically using forest plots for each performance statistic.

5.2.5 Decision curve analysis

In addition to comparing models by discrimination and calibration performance, decision curve analysis (DCA) was performed to assess the clinical value of the models, on cohorts with at least 100 events. Decision curves allow the net benefit (i.e. benefit versus harm) of the models to be determined across a range of clinically plausible threshold probabilities. These included any values up to 0.1, given the generally very low risk of stillbirth, and the range was guided by the clinical collaborators on the project. The decision curves were compared to either simply classifying all women as having an intervention, or no women as having an intervention (Vickers and Elkin, 2006). The strategy with the highest net benefit at a particular threshold has the highest clinical value (Vickers et al., 2016).

Decision curves were plotted in Stata using the *dca* command (Vickers et al., 2008). The net benefit is represented as a function of the decision threshold in decision curve plots. For a probability threshold (p_t), the net benefit is calculated as:

$$\frac{True\ positives}{N} - \left(\frac{False\ positives}{N} \times \frac{p_t}{1 - p_t}\right)$$

where 'true positives' and 'false positives' represent the numbers of individuals with a predicted probability ≥ p_t that do and do not have the outcome of interest respectively, and N is the total sample size (Vickers and Elkin, 2006, Vickers et al., 2016). Clinical collaborators agreed that probability thresholds above 0.05% may be clinically meaningful to make decisions on interventions such as early induction of labour, as this would be a threshold they would roughly implement in practice when considering how high risk a mother and baby would be.

5.3 Results

5.3.1 Prediction models to be included in validation

From 5055 citations, 17 articles were identified describing the development of 40 stillbirth prediction models published between 2007 and 2020. Of these, 11 articles did not publish the final model (Akolekar et al., 2016a, Akolekar et al., 2016b, Akolekar et al., 2011, Åmark et al., 2018, Aupont et al., 2016, Cantarutti et al., 2018, Familiari et al., 2016, Goyal et al., 2015, Mastrodima et al., 2016, Malacova et al., 2020, Reddy et al., 2010)), hence could not be validated, and three included predictors that were not recorded within any of the datasets in IPPIC (Kayode et al., 2016, Payne et al., 2015, Vellamkondu et al., 2017). Within the remaining eligible three articles (Smith et al., 2007, Trudell et al., 2017, Yerlikaya et al., 2016), four prediction models were identified.

The characteristics of included studies and models are described in Table 5.1. All four models were developed using binary logistic regression in unselected populations of pregnant women, and the definition of stillbirth varied between the studies. Two models included only maternal clinical characteristics as predictors (Trudell et al., 2017, Yerlikaya et al., 2016), while the other two models additionally included ultrasound markers (Smith et al., 2007). Only one study had at least 10 events per predictor for model development (Yerlikaya et al., 2016), the others did not justify why their sample size were sufficient. Using the PROBAST tool, the overall risk of bias for all four models was high, with all models assessed as being at high risk of bias in the analysis domain. Whilst the risk of bias was not assessed as part of this thesis, and so is not included here, more details can be found in Allotey et al. (2022).

Three models were developed using UK populations (Smith et al., 2007, Yerlikaya et al., 2016) and the fourth was developed in a population of women in the US (Trudell et al., 2017). Whilst none of the studies indicated that the models did not perform well, none suggested that the models were useful (i.e. good internal validity). However, Smith et al. (2007) suggest that the models they developed are likely to be generalisable to other populations as the association between abnormal uterine artery doppler and risk of stillbirth did not differ between the different centres analysed in the study. Yerlikaya et al. (2016) indicated that the performance of their model may be overestimated as they used the same dataset to derive and test the model, and Trudell et al. (2017) acknowledged the low AUC of their model, however go on to state that "it has been demonstrated that any model developed to predict a rare outcome using covariates that have individual risk that are relatively low for the outcome would be expected to have low discriminative accuracy”. They then recommend that external validation is the next step.

There were no concerns regarding the timing of predictor measurement in any of the identified models.

Table 5.1: Stillbirth prediction model equations externally validated in the IPPIC datasets

Model no.	Author, year; AUC (95% CI)	Outcome; Gestation at stillbirth	Predictor category	Prediction model equation for linear predictor(LP)*
1a	Smith et al. (2007) 0.89 (0.84, 0.95)	24-32 weeks	Clinical characteristics and ultrasound markers	LP = - 9.996 + 1.896(mean pulsatility index) + 1.593(if bilateral notch present) +1.066(if African-American ethnicity) + 1.517(if previous pregnancy loss)

1b	(Smith et al., 2007)
	0.70 (0.64, 0.77)
		≥33 weeks	Clinical characteristics and ultrasounds markers	LP = - 7.806 + 0.867(mean pulsatility index)
				+ 0.768(if BMI 25-29.9) + 0.768(if BMI≥30) + 0.624(if African-American ethnicity)

2	Yerlikaya et al. (2016) 0.608	≥24 weeks	Clinical characteristics	LP = - 6.02615 + 0.01037(weight(kg) – 69) + 0.70027(if Afro-Caribbean ethnicity) + 0.57994(if assisted conception) + 0.53367(if smoke cigarettes) + 0.96253(if chronic hypertension) + 1.28416(if APS or SLE) + 0.93628(if diabetic) + 1.57086(if parous with previous stillbirth)
3	Trudell et al. (2017) 0.66 (0.60, 0.72)	≥32 weeks	Clinical characteristics	LP = - 6.8772 – 0.8707(if maternal age < 18) + 0.2094(if maternal age 35-39) + 0.4377(if maternal age > 40) + 0.8536(if black race) + 0.3423(if nulliparous) – 0.0219(if BMI 25-29.9) + 0.5607(if BMI 30-34.9) – 0.5948(if BMI 35-39.9) + 0.1593(if BMI>40) + 0.2770(if current smoker) + 0.6255(if chronic hypertension) + 0.9863(if pre-gestational diabetes)

BMI=body mass index; APS=antiphospholipid syndrome; SLE=systemic lupus erythematosus.
* For logistic regression, logit(p)=LP where the linear predictor (LP) = $\alpha + \beta_1 * x_1 + \beta_2 * x_2 + \ldots$, and absolute predicted probabilities (p) can be obtained using the transformation $p=\frac{e^{LP}}{1+e^{LP}}$.

5.3.2 Characteristics of the IPPIC validation cohorts

Of the 78 cohorts in the IPPIC repository, 19 cohorts (24%) contained relevant data that could be used to externally validate at least one of the prediction models identified. Only women with singleton pregnancies in the cohorts were used for external validation. Seventy-four percent (14/19) of the cohorts had an overall low risk of bias, 21% (4/19) had a high risk and one had an unclear risk, as assessed by PROBAST. The PROBAST risk assessment was conducted prior to my involvement in this study, however the full results of the risk of bias assessment can be found in Supplementary Table 2 in Allotey et al. (2022).

For model 1a, described in Smith et al. (2007), only six outcomes were identified within the IPPIC cohorts (1, 2 and 3 between 3 datasets), and hence the decision was made that this model could not be validated due to too few events available.

Summary maternal characteristics and outcomes of women in the validation cohort are provided in Table 5.2 and Table 5.3 respectively. The mean age was consistent across most cohorts, being higher in Goetzinger and lower in WHO and NICHD LR. The median BMI was similar across most cohorts, ranging between 20 and 25, apart from NICHD HR, POUCH and Van Oostwaard 2014 in which it was slightly higher. Ethnicity varied across the different cohorts, with 11 (58%) cohorts predominantly consisting of white women, 7 (37%) including a mix of ethnicities and 1 (5%) including only Asian women. All or most (>92%) of the women in 7 (37%) of the studies were nulliparous and 2 (11%) of the studies only included women who were not nulliparous. The proportion of nulliparous women in the other cohorts ranged from 20% to 56%.

A summary of missing data for each predictor and outcome in each cohort is provided in Table 5.4a and Table 5.4b.

Table 5.2: Maternal characteristics and outcomes of IPPIC individual participant datasets used for validation

Dataset	**Country**	**N**	**Maternal age: mean (SD); range**	**BMI: median [IQR], range**	**Ethnicity, n (%)**						**Nulliparous, n (%)**
					White	Black	Asian	Hispanic	Mixed	Other	
Stork	England	54635	30.5 (5.6); 13, 54	23.5 [21.3, 26.8]; 13, 54	33257 (62)	7820 (15)	10388 (19)	5 (<1)	1528 (3)	555 (1)	29313 (54)
Test	Ireland	557	32.0 (4.8); 18, 43	24 [21.6, 27.1]; 17.4, 45.2	539 (97)	2 (<1)	10 (2)	0 (0)	6 (1)	0 (0)	557 (100)
POP	England	4212	29.9 (5.1); 16, 48	24.1 [21.8, 27.3]; 14.7, 54.7	3,900 (93)	25 (<1)	91 (2)	0 (0)	1 (<1)	195 (5)	4212 (100)
Allen	England	1045	29.9 (5.1); 15, 48	23.6 [21.0, 26.8]; 14.8, 51.1	398 (38)	108 (10)	495 (497)	0 (0)	12 (1)	30 (3)	584 (56)
Goetzinger	US	4035	34.8 (4.4); 16, 52	24.4 [21.8, 28.8]; 15.4, 62.4	3282 (83)	397 (10)	1120 (3)	65 (2)	0 (0)	116 (3)	751 (20)
Chie	Japan	379390	32.2 (5.4); 10, 59	20.5 [19.0, 22.6]; 10.5, 69.8	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	379390 (100)	195983 (52)
StorkG	Oslo	812	29.8 (4.8); 19, 45	25.1 [22.3, 28.4]; 16.2, 49.8	375 (46)	61 (8)	198 (24)	12 (1)	0 (0)	166 (20)	377 (46)
Scope	NZ, Aus, UK, RoI	5628	28.7 (5.5); 14, 45	24.2 [21.9, 27.5]; 15.4, 58.5	5061 (90)	65 (1)	304 (5)	24 (<1)	0 (0)	174 (3)	5628 (100)
ALSPAC	England	15038	27.7 (4.9); 13, 46	21.5 [19.7, 23.7]; 11.7, 61.3	11769 (97)	127 (1)	113 (1)	0 (0)	0 (0)	76 (<1)	5704 (45)
Antsaklis	Greece	3328	30.9 (4.8); 14, 47	22.7 [20.6, 25.7]; 14.5, 50.1	3229 (97)	49 (1)	32 (1)	0 (0)	0 (0)	11 (<1)	3328 (100)

Dataset	Country	N	Maternal age: mean (SD); range	BMI: median [IQR], range	Ethnicity, n (%)						Nulliparous, n (%)
					White	Black	Asian	Hispanic	Mixed	Other	
WHO	Multiple	7273	22.5 (5.8); 11, 51	23.1 [21.0, 26.1]; 13.5, 54.8	2222 (31)	756 (10)	1443 (20)	0 (0)	0 (0)	2846 (39)	6710 (92)
Andersen	Denmark	2120	30.2 (4.5); 17, 45	23.4 [21.2, 26.2]; 14.9, 49.9	1765 (97)	3 (<1)	31 (2)	5 (<1)	0 (0)	24 (1)	1193 (56)
NICHD HR	Netherlands	1848	27.1 (6.3); 15, 43	28.4 [23.5, 35.0]; 13.4, 68.5	612 (33)	1079 (58)	2 (<1)	148 (8)	0 (0)	7 (<1)	430 (23)
NICHD LR	US	3097	20.6 (4.4); 15, 39	22.7 [20.4, 25.7]; 13.4, 51.2	548 (18)	1515 (49)	2 (<1)	1010 (33)	0 (0)	22 (<1)	3097 (100)
POUCH	US	3019	26.4 (5.8); 15, 47	27.7 [24.3, 32.9]; 15.1, 66.3	2018 (67)	743 (25)	57 (2)	160 (5)	0 (0)	41 (1)	1293 (43)
Rumbold	NZ and Aus	1877	26.4 (5.7); 13, 44	24.1 [21.5, 27.6]; 13.7, 57.6	1777 (95)	3 (<1)	1 (<1)	1 (<1)	4 (<1)	87 (5)	1877 (100)
Indonesian cohort	Indonesia	2223	28.6 (5.9); 10, 59	22.9 [20.1, 26.3]; 13.3, 67.6	0 (0)	0 (0)	2223 (100)	0 (0)	0 (0)	0 (0)	664 (43)
Van Oostwaard 2012	Netherlands	425	32.0 (4.1); 23, 42	24.3 [21.5, 27.9]; 16.2, 41.8	288 (84)	46 (13)	4 (1)	0 (0)	3 (1)	2 (1)	0 (0)
Van Oostwaard 2014	Netherlands	639	32.1 (4.4); 21, 43	25.9 [22.5, 31.2]; 17.7, 56.5	360 (72)	119 (24)	17 (4)	0 (0)	3 (1)	2 (<1)	0 (0)

Table 5.3: Number of outcomes in each IPPIC individual participant dataset used for validation

Dataset	N	Outcome, n (%)		
		≥33weeks	*≥24weeks*	*≥32weeks*
Stork	54635	148 (0.27)	233 (0.43)	160 (0.29)
Test	557	4 (0.73)	5 (0.92)	4 (0.73)
POP	4212	8 (0.19)	11 (0.26)	8 (0.19)
Allen	1045	3 (0.29)	3 (0.29)	3 (0.29)
Goetzinger	4035	15 (0.37)	15 (0.37)	15 (0.37)
Chie	379390	801 (0.21)	1792 (0.47)	895 (0.24)
StorkG	812	4 (0.49)	6 (0.74)	5 (0.62)
Scope	5628	8 (0.14)	17 (0.30)	9 (0.16)
ALSPAC	15038	26 (0.17)	41 (0.27)	27 (0.18)
Antsaklis	3328	2 (0.06)	2 (0.06)	2 (0.06)
WHO	7273	8 (0.46)	8 (0.46)	8 (0.46)
Andersen	2120	4 (0.19)	6 (0.28)	4 (0.19)
NICHD HR	1848	8 (0.44)	23 (1.26)	8 (0.44)
NICHD LR	3097	6 (0.20)	13 (0.44)	6 (0.20)
POUCH	3019	4 (0.13)	10 (0.33)	4 (0.13)
Rumbold	1877	9 (0.48)	11 (0.59)	9 (0.48)
Indonesian cohort	2223	6 (0.35)	12 (0.70)	6 (0.35)
Van Oostwaard 2012	425	2 (1.05)	2 (1.05)	2 (1.05)
Van Oostwaard 2014	639	3 (0.98)	5 (1.64)	3 (0.98)

Table 5.4a: Number and proportion missing (or not recorded) for each predictor in each dataset used for external validation

Dataset	**N (%) missing or not recorded**						
	Maternal age	**T1 BMI**	**T2 BMI**	**T1 Weight**	**Ethnicity**	**Pulsatility index**	**Assisted conception**
Stork	0 (0)	13286 (24)	25186 (46)	12173 (22)	1082 (2)	28109 (51)	1427 (3)
Test	0 (0)	1 (<1)	557 (100)	1 (<1)	0 (0)	0 (0)	0 (0)
POP	0 (0)	152 (4)	57 (1)	146 (3)	0 (0)	133 (3)	0 (0)
Allen	1 (<1)	5 (<1)	1040 (99)	5 (<1)	2 (<1)	1040 (99)	0 (0)
Goetzinger	72 (2)	606 (15)	4035 (100)	531 (13)	63 (2)	4035 (100)	92 (2)
Chie	1108 (<1)	53711 (14)	379390 (100)	51196 (13)	0 (0)	379390 (100)	0 (0)
StorkG	0 (0)	414 (51)	245 (30)	414 (51)	0 (0)	812 (100)	0 (0)
Scope	0 (0)	5490 (98)	7 (<1)	5490 (98)	0 (0)	5628 (100)	0 (0)
ALSPAC	2047 (14)	3103 (21)	15038 (100)	9884 (66)	2953 (20)	15038 (100)	2861 (19)
Antsaklis	11 (<1)	480 (14)	2966 (89)	122 (4)	7 (<1)	3204 (96)	3328 (100)
WHO	1 (<1)	2585 (36)	1078 (15)	4 (<1)	6 (<1)	7273 (100)	7273 (100)
Andersen	0 (0)	1070 (50)	1506 (71)	650 (31)	292 (14)	2120 (100)	2120 (100)
NICHD HR	9 (<1)	1711 (93)	157 (9)	18 (1)	0 (0)	1848 (100)	1848 (100)
NICHD LR	99 (3)	3024 (98)	177 (6)	95 (3)	0 (0)	3097 (100)	3097 (100)
POUCH	0 (0)	1 (<1)	1 (<1)	0 (0)	0 (0)	3019 (100)	3019 (100)
Rumbold	0 (0)	967 (52)	1128 (60)	171 (9)	4 (<1)	1877 (100)	39 (2)
Indonesian cohort	74 (3)	203 (9)	1601 (72)	935 (42)	0 (0)	2223 (100)	2223 (100)
Van Oostwaard 2012	232 (55)	265 (62)	425 (100)	425 (100)	82 (19)	425 (100)	425 (100)
Van Oostwaard 2014	329 (100)	388 (61)	639 (100)	639 (100)	138 (22)	639 (100)	639 (100)

Table 5.4b: Number and proportion missing (or not recorded) for each predictor in each dataset used for external validation

Dataset	N (%) missing or not recorded						
	Smoker	Hyper-tension	APS/ SLE	Previous stillbirth	Nulliparous	Pre-gestational diabetes	Outcome
Stork	4053 (7)	0 (0)	54635 (100)	54635 (100)	104 (<1)	54635 (100)	0 (0)
Test	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	557 (100)	11 (2)
POP	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	3 (<1)
Allen	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Goetzinger	244 (6)	281 (7)	0 (0)	0 (0)	302 (7)	3406 (84)	11 (<1)
Chie	79224 (21)	0 (0)	0 (0)	0 (0)	1490 (<1)	0 (0)	10 (<1)
StorkG	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (<1)
Scope	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
ALSPAC	2666 (18)	3001 (20)	15038 (100)	2232 (15)	2439 (16)	2813 (19)	77 (<1)
Antsaklis	6 (<1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	3 (<1)
WHO	2 (<1)	1 (<1)	2 (<1)	7273 (100)	0 (0)	1 (<1)	5548 (77)
Andersen	2 (<1)	192 (9)	2120 (100)	2120 (100)	0 (0)	193 (9)	0 (0)
NICHD HR	0 (0)	0 (0)	1848 (100)	1848 (100)	0 (0)	0 (0)	28 (2)
NICHD LR	6 (<1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	159 (5)
POUCH	6 (<1)	1 (<1)	2458 (81)	14 (<1)	1 (<1)	0 (0)	0 (0)
Rumbold	39 (2)	0 (0)	1877 (100)	0 (0)	0 (0)	0 (0)	0 (0)
Indonesian cohort	1249 (56)	18 (1)	2223 (100)	2223 (100)	667 (30)	1 (<1)	499 (22)
Van Oostwaard 2012	242 (57)	273 (64)	278 (65)	0 (0)	0 (0)	278 (65)	235 (55)
Van Oostwaard 2014	350 (55)	404 (63)	407 (64)	0 (0)	0 (0)	406 (64)	334 (53)

5.3.3 Linear predictors

Table 5.5 gives a summary of the linear predictors and predicted probabilities for each model and validation cohort. These linear predictors are also presented graphically in Figure 5.1 by their median and range. It can be seen here that the linear predictors were similar for each cohort within each of the models, and also similar in model 1b and model 2, but overall higher in model 3. These linear predictors correspond to very small, predicted probabilities, which is expected for a rare outcome. The median predicted probability for models 1b and 3 was around 0.001. It was slightly higher for model 2, ranging from 0.009 to 0.03. Similarly, the ranges of predicted probabilities were extremely low, again being highest for model 2 (up to 0.6 for the Goetzinger cohort). This means that no one is predicted a high probability of having the outcome, even if they did have the outcome, which makes it difficult to discriminate between those who do and do not have the outcome. The four cohorts used to validate model 2 were also used to validate model 3, so it can be seen that model 2 is predicting a higher probability of having the outcome even in the same cohorts (therefore the same individuals).

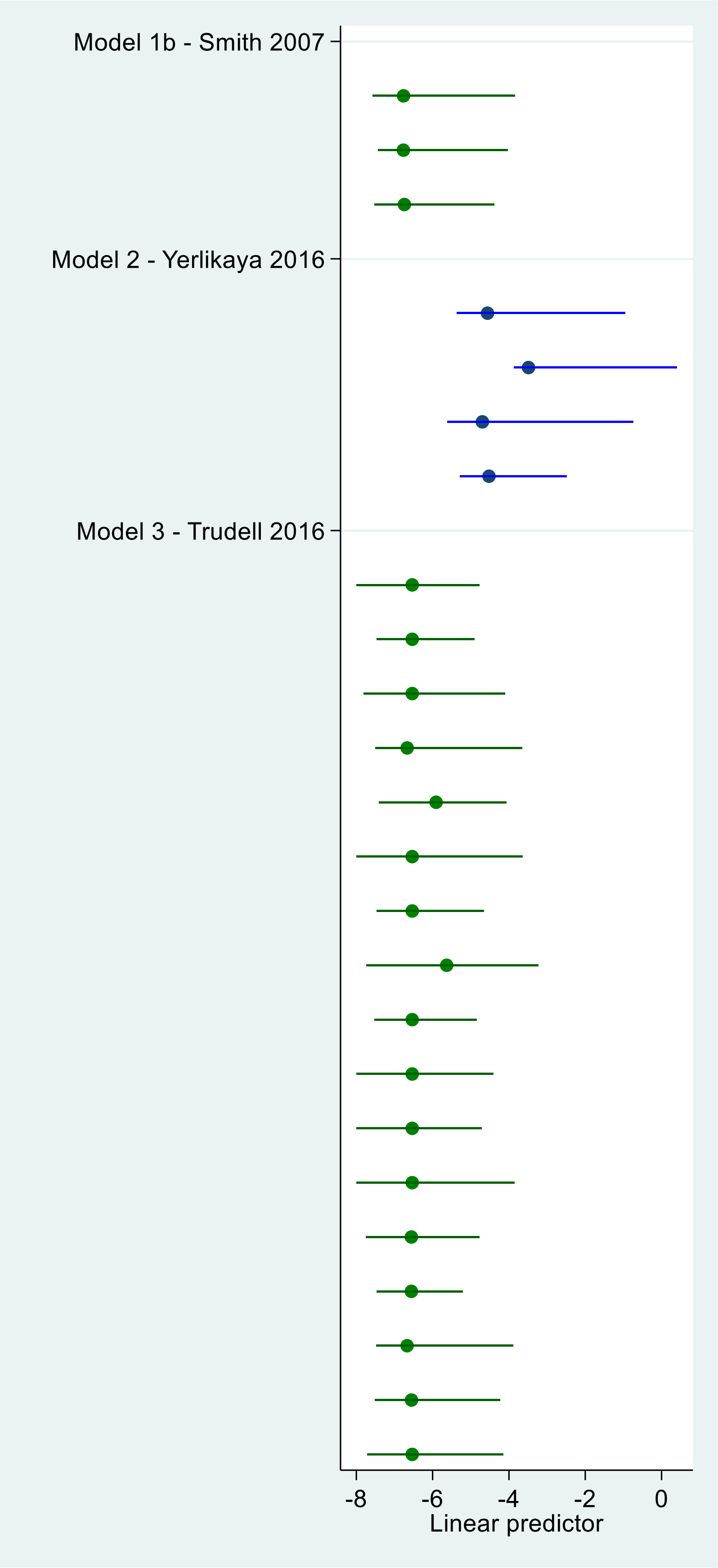
Table 5.5: Summary of linear predictors and predicted probabilities for the models for each study used in the validation

Model no.	**First author (year);** ***Outcome***	**Study**	**N Total**	**No. Events (%)**	**Linear predictor**			**Predicted probability**		
					Median	**Interquartile range**	**Range (min to max)**	**Median**	**Interquartile range**	**Range (min to max)**
1b	Smith 2007 *33+ weeks*	Stork	54635	148 (0.27)	-6.762	-7.112, -6.284	-7.572, -3.845	0.00116	0.00081, 0.00186	0.00052, 0.02095
		TEST	557	4 (0.72)	-6.767	-7.104, -6.327	-7.428, -4.025	0.00115	0.00082, 0.00179	0.00059, 0.01755
		POP	4212	8 (0.18)	-6.741	-7.082, -6.302	-7.524, -4.380	0.00118	0.00083, 0.00183	0.00054, 0.01236
2	Yerlikaya 2016 *24+ weeks*	Allen	1045	3 (0.29)	-4.562	-4.655, -4.427	-5.370, -0.955	0.0103	0.0094, 0.0118	0.0046, 0.2780
		Goetzinger	4035	26 (0.64)	-3.487	-3.600, -3.064	-3.867, 0.406	0.0297	0.0266, 0.0446	0.0205, 0.5992
		Chie	379390	1802 (0.47)	-4.697	-4.759, -4.624	-5.611, 0.738	0.0090	0.0085, 0.0097	0.0036, 0.3235
		StorkG	812	7 (0.86)	-4.522	-4.622, -4.361	-5.283, 2.482	0.0107	0.0097, 0.0126	0.0051, 0.0771
3	Trudell 2016 *32+ weeks*	Scope	5628	9 (0.16)	-6.535	-6.557, -6.326	-8.000, -4.779	0.00145	0.00142, 0.00179	0.00034, 0.00834
		Allen	1045	3 (0.29)	-6.535	-6.877, -6.347	-7.472, -4.911	0.00145	0.00103, 0.00175	0.00057, 0.00731

Model no.	First author (year); *Outcome*	Study	N Total	No. Events (%)	Linear predictor			Predicted probability		
					Median	Interquartile range	Range (min to max)	Median	Interquartile range	Range (min to max)
3	Trudell 2016 *32+ weeks*	ALSPAC	15038	27 (0.18)	-6.535	-6.877, -6.312	-7.812, -4.103	0.00145	0.00103, 0.00181	0.00041, 0.01634
		Goetzinger	4035	24 (0.59)	-6.668	-6.690, -6.324	-7.504, -3.652	0.00127	0.00124, 0.00179	0.00056, 0.02565
		Antsaklis	3328	2 (0.06)	-5.909	-6.258, -5.722	-7.406, -4.070	0.00271	0.00191, 0.00326	0.00061, 0.01680
		WHO	7273	63 (0.87)	-6.535	-6.557, -6.535	-8.000, -3.642	0.00145	0.00142, 0.00145	0.00034, 0.02554
		Andersen	2120	4 (0.19)	-6.535	-6.877, -6.535	-7.472, -4.663	0.00145	0.00103, 0.00145	0.00057, 0.00984
		NICHD HR	1848	8 (0.43)	-5.632	-6.028, -5.143	-7.748, -3.232	0.00357	0.00241, 0.00581	0.00043, 0.03799
		NICHD LR	3097	7 (0.23)	-6.535	-6.552, -5.681	-7.531, -4.844	0.00145	0.00143, 0.00340	0.00055, 0.00782
		POUCH	3019	4 (0.13)	-6.535	-6.718, -6.040	-8.000, -4.406	0.00145	0.00121, 0.00238	0.00034, 0.01206
		Rumbold	1877	9 (0.48)	-6.535	-6.557, -6.280	-8.000, -4.711	0.00145	0.00142, 0.00187	0.00034, 0.00892

Model no.	First author (year); *Outcome*	Study	N Total	No. Events (%)	Linear predictor			Predicted probability		
					Median	Interquartile range	Range (min to max)	Median	Interquartile range	Range (min to max)
3	Trudell 2016 *32+ weeks*	Chie	379390	897 (0.24)	-6.535	-6.877, -6.535	-8.000, -3.858	0.00145	0.00103, 0.00145	0.00034, 0.02072
		Indonesian cohort	2223	11 (0.49)	-6.557	-6.877, -6.535	-7.752, -4.773	0.00142	0.00103, 0.00145	0.00043, 0.00839
		StorkG	812	6 (0.74)	-6.557	-6.877, -6.403	-7.472, -5.211	0.001418	0.00103, 0.00166	0.00057, 0.00544
		Van Oostwaard 2012	425	14 (3.29)	-6.668	-6.877, -6.086	-7.478, -3.887	0.001269	0.00103, 0.00227	0.00057, 0.02042
		Van Oostwaard 2014	639	4 (0.63)	-6.555	-6.877, -6.028	-7.519, -4.234	0.001425	0.00103, 0.00241	0.00055, 0.01522
		POP	4212	8 (0.19)	-6.535	-6.557, -6.326	-7.723, -4.153	0.00145	0.00142, 0.00179	0.00044, 0.01547

Figure 5.1: Median and range of linear predictors (logit outcome probabilities) from the various prediction models



5.3.4 Performance statistics

Table 5.6 gives the cohort-specific performance statistics, i.e. C-statistic, calibration slope and calibration-in-the-large (CITL), for each of the models. This table highlights the huge uncertainty in the results due to such rare events, as can be seen from the generally very large confidence intervals for the C-statistic. Focusing on those with 100+ events in individual studies, the Stork cohort for model 1b has 148 events with a C-statistic of 0.65 (0.60, 0.70). The calibration slope (which can also be seen graphically in Figure 5.11) is only slightly below 1, however the confidence interval for this is quite large. Ideally, the CITL should be close to zero, but here as it is above 0 this indicates that the predictions are systematically too low.

For model 2, there is one cohort (Chie) with more than 100 events, which has 1802 events. The confidence interval for the C-statistic is fairly narrow for this cohort as it is a much larger dataset than the others, however it is only just above 0.5 (C-statistic = 0.54 (95% CI 0.53, 0.56)) which means the model is only slightly better than chance at discriminating between those with and without an event. The calibration slope is also much below 1 here, and the CITL is below 0, indicating that the predictions are systematically too large. This is not surprising after seeing that the predicted probabilities were higher than for the other models (Table 5.5 and Figure 5.1).

For model 3, although there are a lot more cohorts included in the validation of this model, Chie is the only cohort with more than 100 events (897). The C-statistic is 0.53 (95% CI 0.51, 0.55), which is again very poor. The model is also predicting systematically too low in this dataset as the CITL is positive.

Table 5.7 shows the summary estimates of the predictive performance statistics from the meta-analysis across all validation cohorts for each model. The summary performance statistics are also shown graphically for all models in the forest plots in Figure 5.2 to Figure 5.10. The C-statistics are all quite close to 0.5, and seem to be very dominated by the large cohorts used to validate each of the models. The calibration slope for the model 1b is close to 1, however has a large confidence interval, while the calibration slopes for the other two models are much lower, indicating poor levels of agreement between the observed outcomes and the predictions. Although many of the cohorts only had a few events, heterogeneity was estimated as 0 for the C-statistic and calibration slope of model 3. However, there was heterogeneity in the CITL for this model (τ^2=0.552).

Direct comparison of the prediction models is difficult due to different outcomes and different cohorts contributing towards the validation of each prediction model.

Table 5.6: Study specific performance statistics

Model no.	First author (year)	Outcome	Study	N Total	No. Events (%)	Performance statistic (95% CI)		
						C-statistic	Calibration slope	Calibration-in-the-large
1b	Smith 2007	33+ weeks	Stork	54635	148 (0.27)	0.650 (0.600, 0.696)	0.866 (0.574, 1.159)	0.573 (0.411, 0.734)
			TEST	557	4 (0.72)	0.817 (0.520, 0.949)	1.573 (0.157, 2.989)	1.737 (0.751, 2.722)
			POP	4212	8 (0.18)	0.560 (0.357, 0.745)	0.492 (-0.934, 1.918)	0.288 (-0.406, 0.982)
2	Yerlikaya 2016	24+ weeks	Allen	1045	3 (0.29)	0.643 (0.308, 0.879)	0.541 (-1.570, 2.652)	-1.524 (-2.659, -0.389)
			Goetzinger	4035	26 (0.64)	0.629 (0.417, 0.801)	0.660 (-0.099, 1.421)	-1.980 (-2.371, -1.589)
			Chie	379390	1802 (0.47)	0.544 (0.529, 0.559)	0.436 (0.321, 0.552)	-0.742 (-0.789, -0.696)
			StorkG	812	7 (0.86)	0.730 (0.562, 0.851)	1.043 (-0.415, 2.500)	-0.408 (-1.153, 0.337)
3	Trudell 2016	32+ weeks	Scope	5628	9 (0.16)	0.336 (0.200, 0.507)	-1.842 (-3.769, 0.857)	-0.032 (-0.686, 0.622)
			Allen	1045	3 (0.29)	0.469 (0.176, 0.785)	-0.279 (-3.427, 2.869)	0.575 (-0.559, 1.709)
			ALSPAC	15038	27 (0.18)	0.477 (0.331, 0.628)	-0.0449 (-1.773, 1.683)	0.149 (-0.229, 0.526)
			Goetzinger	4035	24 (0.59)	0.542 (0.271, 0.789)	0.523 (-0.700, 1.746)	1.201 (0.783, 1.619)
			Antsaklis	3328	2 (0.06)	0.430 (0.097, 0.842)	-1.078 (-4.723, 2.568)	-1.269 (-2.643, 0.104)
			WHO	7273	63 (0.87)	0.539 (0.404, 0.668)	0.172 (-0.725, 1.070)	1.730 (1.002, 2.459)
			Andersen	2120	4 (0.19)	0.620 (0.279, 0.873)	1.549 (-1.998, 5.096)	0.251 (-0.730, 1.232)
			NICHD HR	1848	8 (0.43)	0.611 (0.387, 0.796)	0.438 (-0.566, 1.442)	-0.026 (-0.721, 0.668)

Model no.	First author (year)	Outcome	Study	N Total	No. Events (%)	Performance statistic (95% CI)		
						C-statistic	Calibration slope	Calibration-in-the-large
3	Trudell 2016	32+ weeks	NICHD LR	3097	7 (0.23)	0.638 (0.347, 0.854)	0.883 (-0.598, 2.363)	0.046 (-0.755, 0.847)
			POUCH	3019	4 (0.13)	0.639 (0.420, 0.812)	0.659 (-1.098, 2.416)	-0.383 (-1.364, 0.597)
			Rumbold	1877	9 (0.48)	0.470 (0.265, 0.686)	-0.676 (-2.636, 1.284)	1.073 (0.418, 1.728)
			Chie	379390	897 (0.24)	0.530 (0.511, 0.548)	0.412 (0.180, 0.645)	0.493 (0.428, 0.559)
			Indonesian cohort	2223	11 (0.49)	0.693 (0.484, 0.845)	1.922 (0.066, 3.777)	1.295 (0.567, 2.024)
			StorkG	812	6 (0.74)	0.432 (0.156, 0.757)	0.287 (-1.794, 2.369)	1.581 (0.774, 2.388)
			Van Oostwaard 2012	425	14 (3.29)	0.644 (0.354, 0.856)	0.650 (-0.745, 2.046)	2.887 (1.711, 4.063)
			Van Oostwaard 2014	639	4 (0.63)	0.594 (0.239, 0.872)	0.378 (-1.484, 2.240)	1.203 (0.032, 2.374)
			POP	4212	8 (0.19)	0.631 (0.399, 0.815)	1.195 (-0.424, 2.814)	0.088 (-0.606, 0.782)

Table 5.7: Summary estimates of performance statistics from meta-analysis

Model No.	**Author (year)**	**Outcome**	**No. of validation cohorts**	**Total events**	**Summary estimate of performance statistic (95% CI), Measures of heterogeneity (I^2, τ^2)**		
					C-statistic	**Calibration slope**	**Calibration-in-the-large**
1b	Smith 2007	≥33 weeks	3	160	0.65 (0.53, 0.75) I^2=0%, τ^2=0	0.88 (0.26, 1.50) I^2=0%, τ^2=0	0.76 (-0.95, 2.48) I^2=76.6%, τ^2=0.292
2	Yerlikaya 2016	≥24 weeks	4	1838	0.61 (0.43, 0.77) I^2=48.6%, τ^2=0.102	0.45 (0.26, 0.63) I^2=0%, τ^2=0	-1.15 (-2.35, 0.05) I^2=91.4%, τ^2=0.462
3	Trudell 2016	≥32 weeks	17	1100	0.53 (0.51, 0.55) I^2=0%, τ^2=0	0.40 (0.19, 0.62) I^2=0%, τ^2=0	0.64 (0.18, 1.11) I^2=89.1%, τ^2=0.552

Figure 5.2: C-statistic – Model 1b

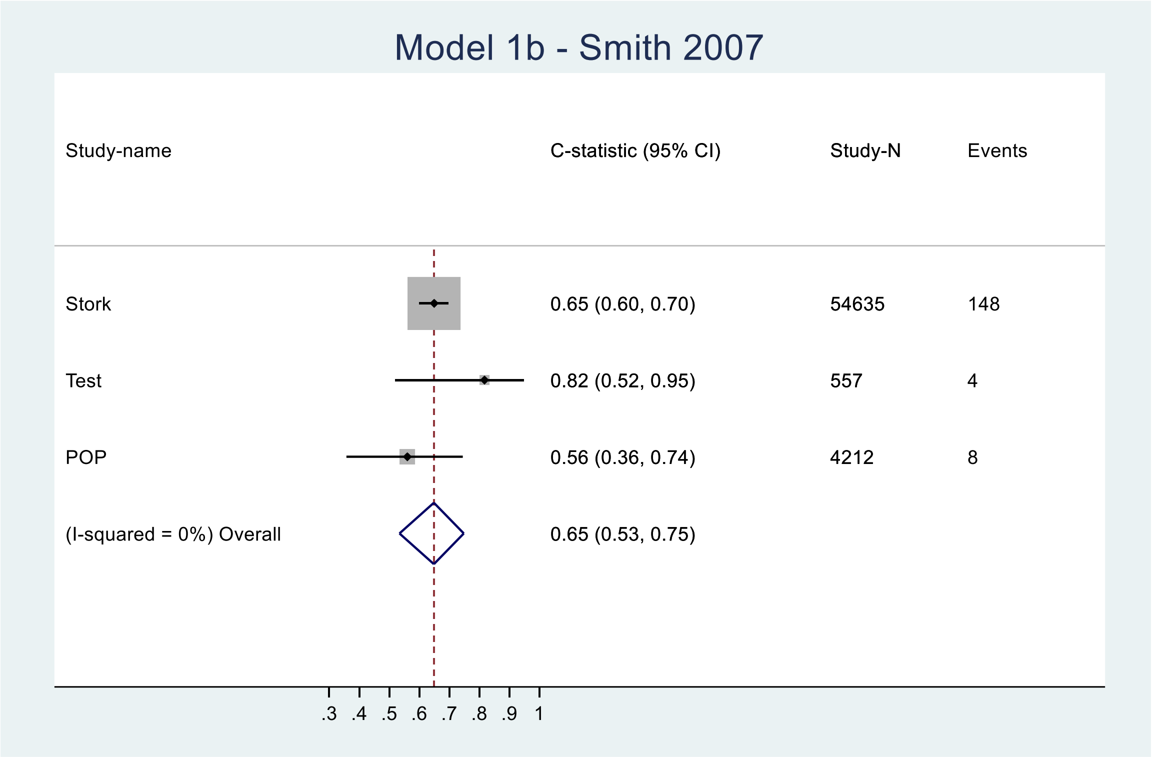


Figure 5.3: Calibration slope – Model 1b

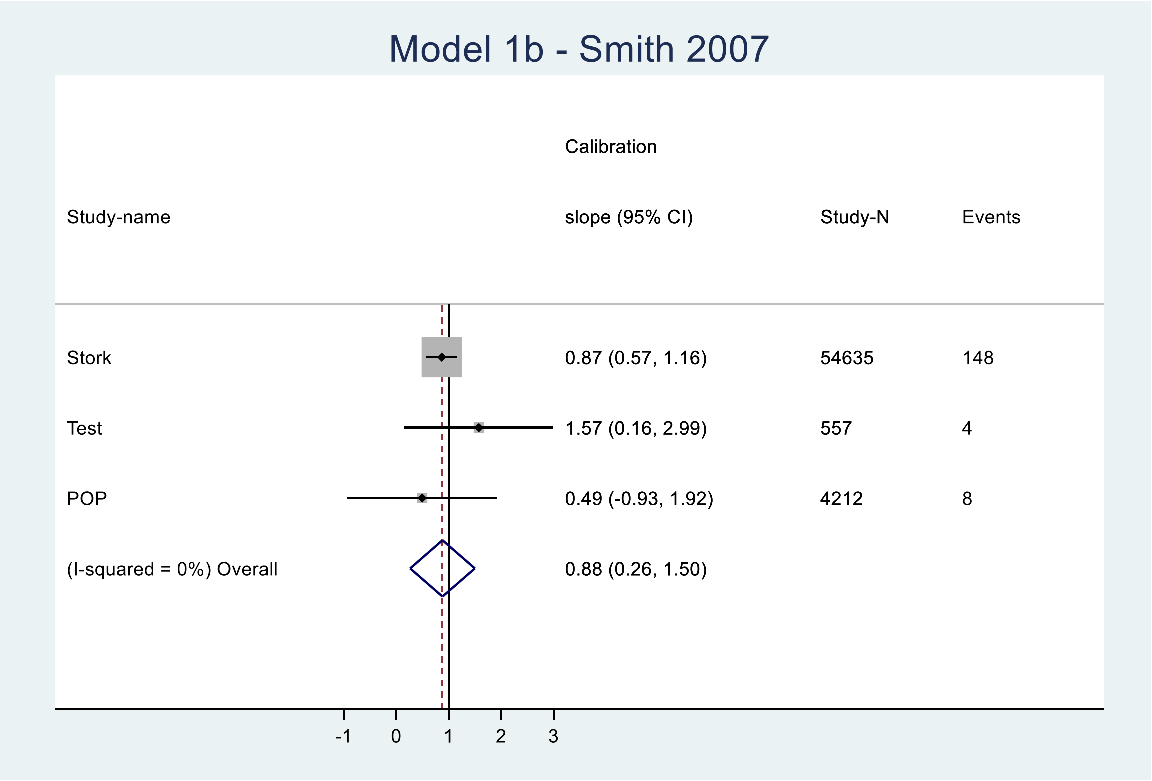


Figure 5.4: Calibration-in-the-large – Model 1b

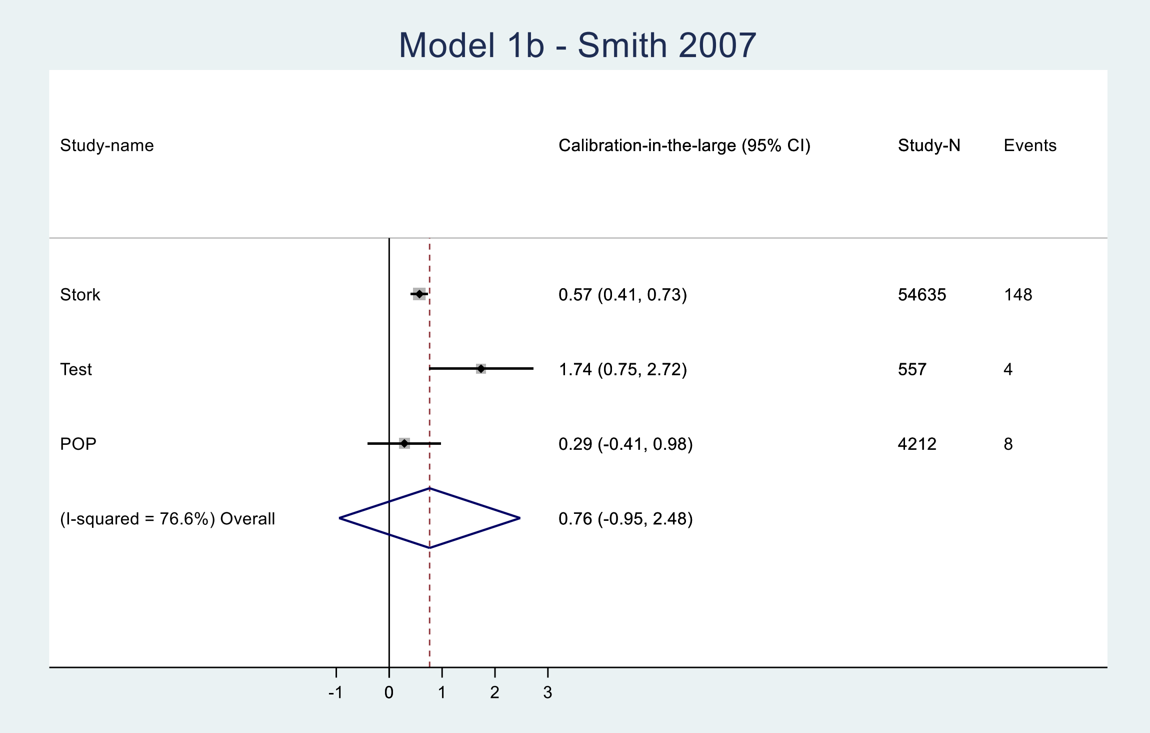


Figure 5.5: C-statistic - Model 2

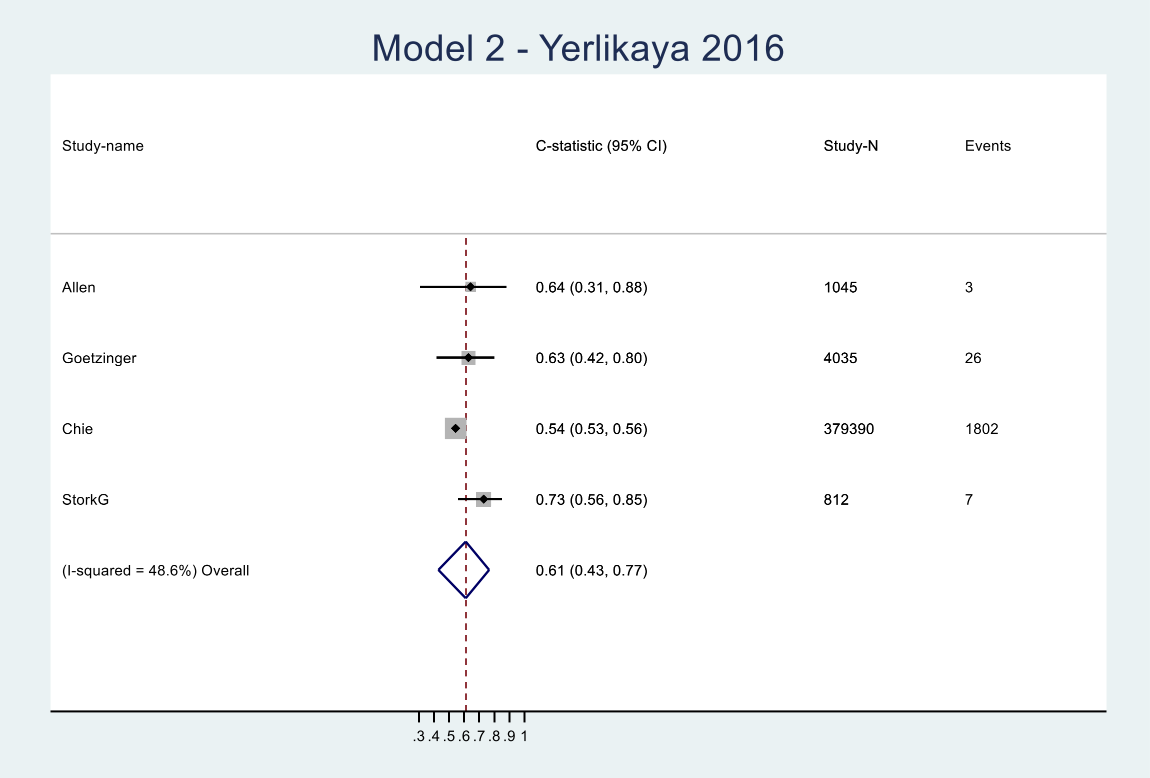
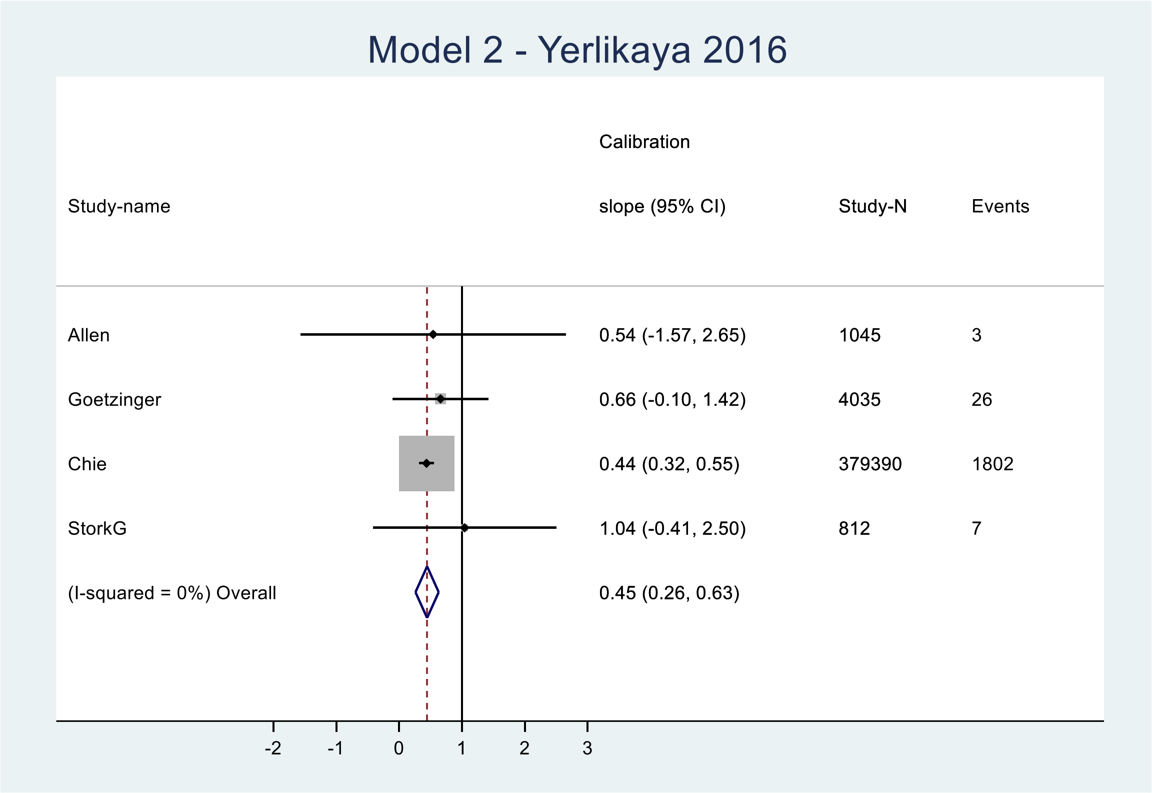
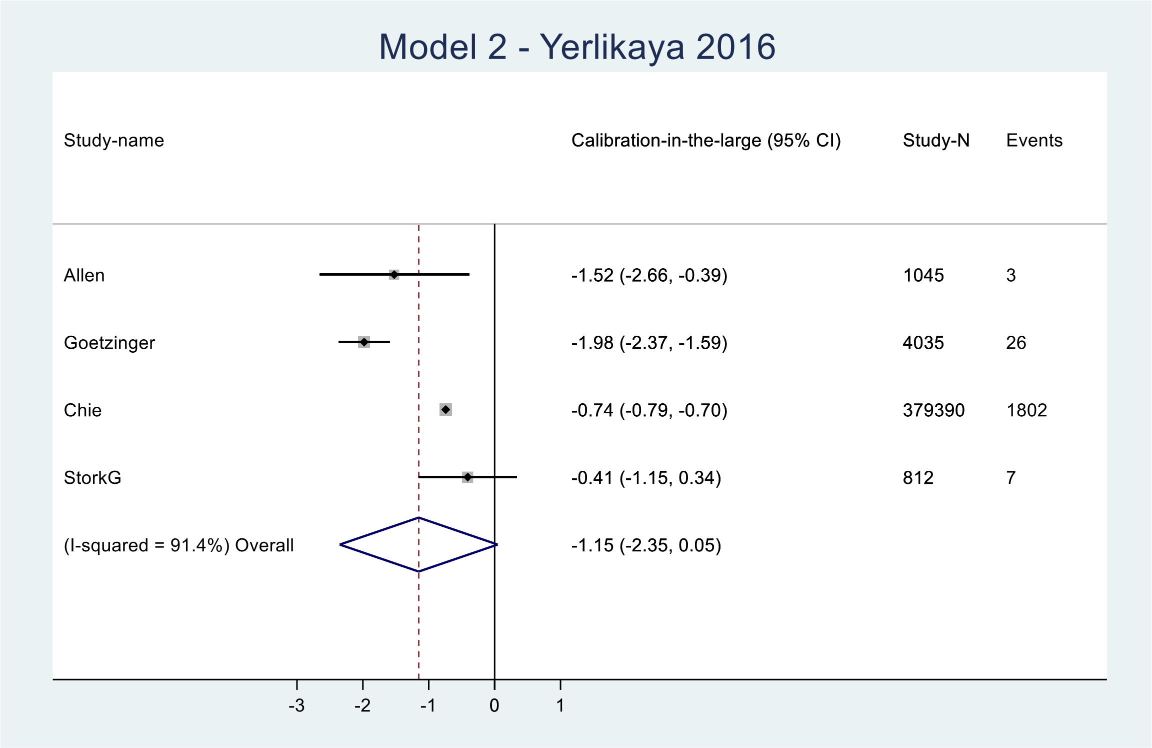


Figure 5.6: Calibration slope – Model 2



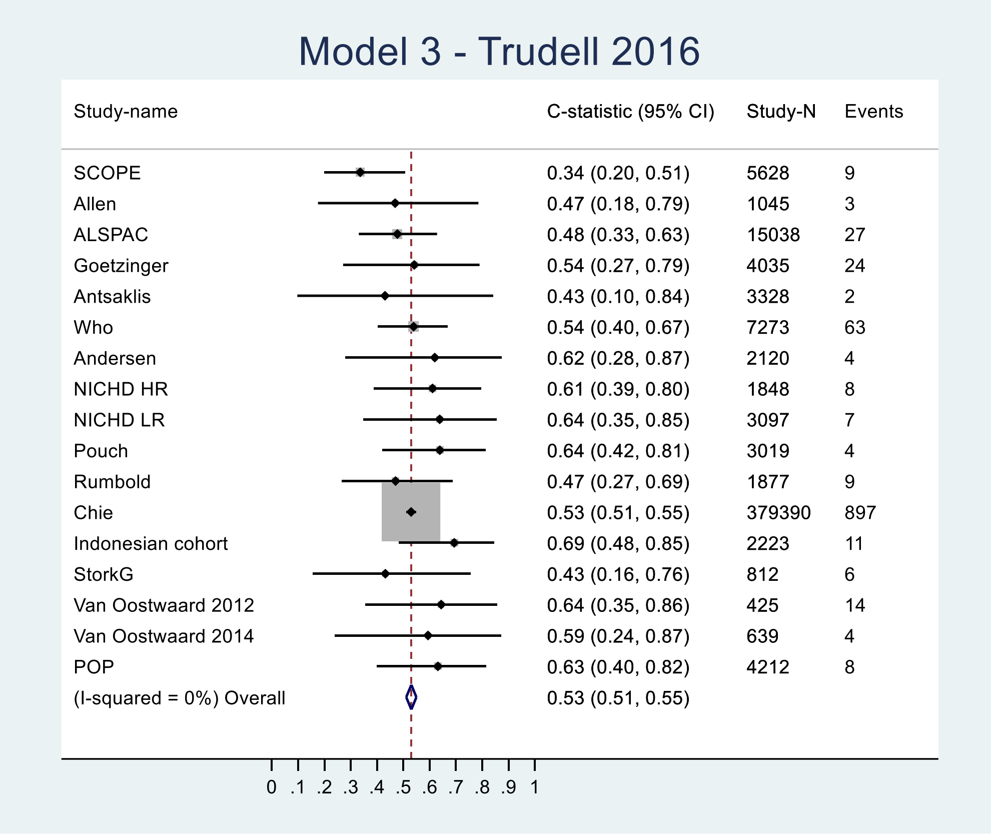
Study-name	Calibration slope (95% CI)	Study-N	Events
Allen	0.54 (-1.57, 2.65)	1045	3
Goetzinger	0.66 (-0.10, 1.42)	4035	26
Chie	0.44 (0.32, 0.55)	379390	1802
StorkG	1.04 (-0.41, 2.50)	812	7
(I-squared = 0%) Overall	0.45 (0.26, 0.63)		

Figure 5.7: Calibration-in-the-large – Model 2



Study-name	Calibration-in-the-large (95% CI)	Study-N	Events
Allen	-1.52 (-2.66, -0.39)	1045	3
Goetzinger	-1.98 (-2.37, -1.59)	4035	26
Chie	-0.74 (-0.79, -0.70)	379390	1802
StorkG	-0.41 (-1.15, 0.34)	812	7
(I-squared = 91.4%) Overall	-1.15 (-2.35, 0.05)		

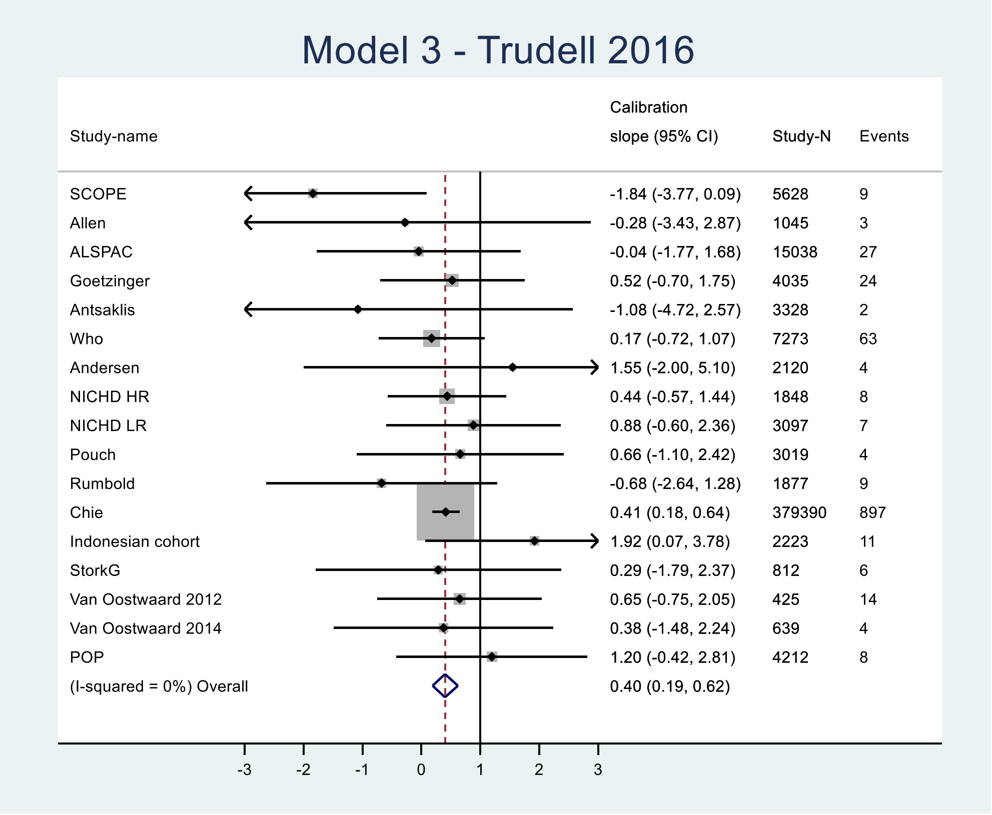
Figure 5.8: C-statistic - Model 3



Model 3 - Trudell 2016

Study-name	C-statistic (95% CI)	Study-N	Events
SCOPE	0.34 (0.20, 0.51)	5628	9
Allen	0.47 (0.18, 0.79)	1045	3
ALSPAC	0.48 (0.33, 0.63)	15038	27
Goetzinger	0.54 (0.27, 0.79)	4035	24
Antsaklis	0.43 (0.10, 0.84)	3328	2
Who	0.54 (0.40, 0.67)	7273	63
Andersen	0.62 (0.28, 0.87)	2120	4
NICHD HR	0.61 (0.39, 0.80)	1848	8
NICHD LR	0.64 (0.35, 0.85)	3097	7
Pouch	0.64 (0.42, 0.81)	3019	4
Rumbold	0.47 (0.27, 0.69)	1877	9
Chie	0.53 (0.51, 0.55)	379390	897
Indonesian cohort	0.69 (0.48, 0.85)	2223	11
StorkG	0.43 (0.16, 0.76)	812	6
Van Oostwaard 2012	0.64 (0.35, 0.86)	425	14
Van Oostwaard 2014	0.59 (0.24, 0.87)	639	4
POP	0.63 (0.40, 0.82)	4212	8
(I-squared = 0%) Overall	0.53 (0.51, 0.55)		

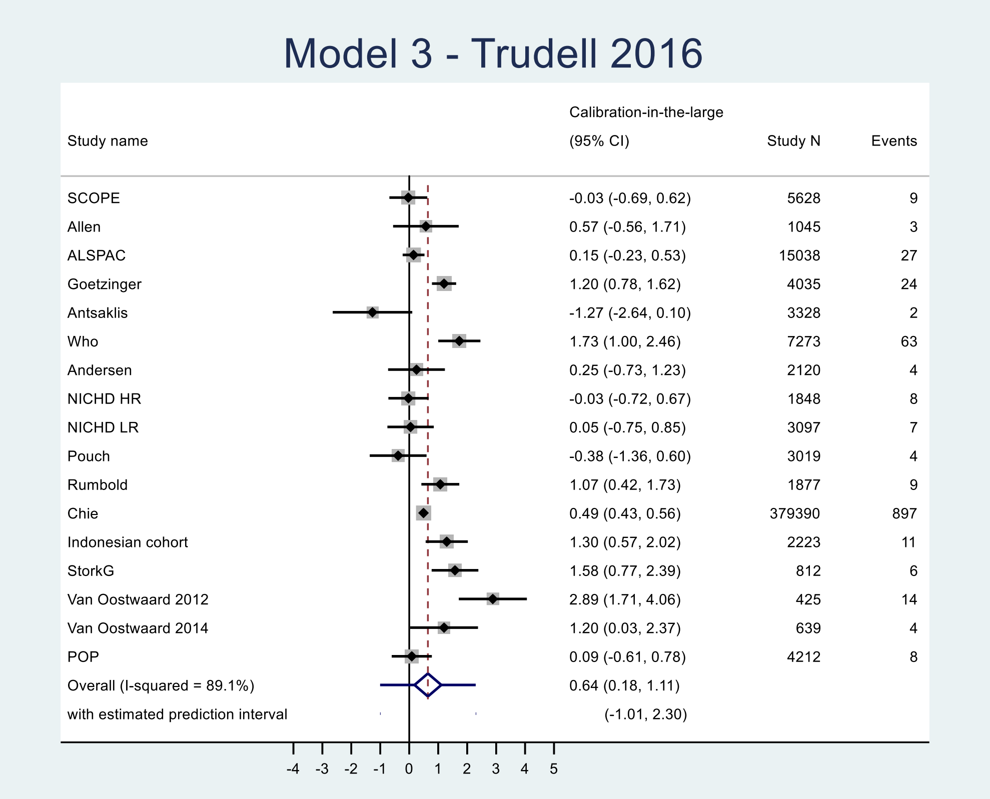
Figure 5.9: Calibration slope – Model 3



Model 3 - Trudell 2016

Study-name	Calibration slope (95% CI)	Study-N	Events
SCOPE	-1.84 (-3.77, 0.09)	5628	9
Allen	-0.28 (-3.43, 2.87)	1045	3
ALSPAC	-0.04 (-1.77, 1.68)	15038	27
Goetzinger	0.52 (-0.70, 1.75)	4035	24
Antsaklis	-1.08 (-4.72, 2.57)	3328	2
Who	0.17 (-0.72, 1.07)	7273	63
Andersen	1.55 (-2.00, 5.10)	2120	4
NICHD HR	0.44 (-0.57, 1.44)	1848	8
NICHD LR	0.88 (-0.60, 2.36)	3097	7
Pouch	0.66 (-1.10, 2.42)	3019	4
Rumbold	-0.68 (-2.64, 1.28)	1877	9
Chie	0.41 (0.18, 0.64)	379390	897
Indonesian cohort	1.92 (0.07, 3.78)	2223	11
StorkG	0.29 (-1.79, 2.37)	812	6
Van Oostwaard 2012	0.65 (-0.75, 2.05)	425	14
Van Oostwaard 2014	0.38 (-1.48, 2.24)	639	4
POP	1.20 (-0.42, 2.81)	4212	8
(I-squared = 0%) Overall	0.40 (0.19, 0.62)		

Figure 5.10: Calibration-in-the-large – Model 3



Study name	Calibration-in-the-large (95% CI)	Study N	Events
SCOPE	-0.03 (-0.69, 0.62)	5628	9
Allen	0.57 (-0.56, 1.71)	1045	3
ALSPAC	0.15 (-0.23, 0.53)	15038	27
Goetzinger	1.20 (0.78, 1.62)	4035	24
Antsaklis	-1.27 (-2.64, 0.10)	3328	2
Who	1.73 (1.00, 2.46)	7273	63
Andersen	0.25 (-0.73, 1.23)	2120	4
NICHD HR	-0.03 (-0.72, 0.67)	1848	8
NICHD LR	0.05 (-0.75, 0.85)	3097	7
Pouch	-0.38 (-1.36, 0.60)	3019	4
Rumbold	1.07 (0.42, 1.73)	1877	9
Chie	0.49 (0.43, 0.56)	379390	897
Indonesian cohort	1.30 (0.57, 2.02)	2223	11
StorkG	1.58 (0.77, 2.39)	812	6
Van Oostwaard 2012	2.89 (1.71, 4.06)	425	14
Van Oostwaard 2014	1.20 (0.03, 2.37)	639	4
POP	0.09 (-0.61, 0.78)	4212	8
Overall (I-squared = 89.1%)	0.64 (0.18, 1.11)		
with estimated prediction interval	(-1.01, 2.30)		

5.3.5 Calibration plots

The calibration plots for each of the cohorts with more than 100 events used for validating the 3 prediction models are given below in Figure 5.11 to Figure 5.13. These clearly show the extent of mis-calibration, highlighting that those with the highest predicted probability of an outcome tend to be individuals who did not have the outcome. However, predicted probabilities were all less than 0.2, therefore absolute risk differences remain small.

Figure 5.11: Calibration plot for model 1b - Stork dataset

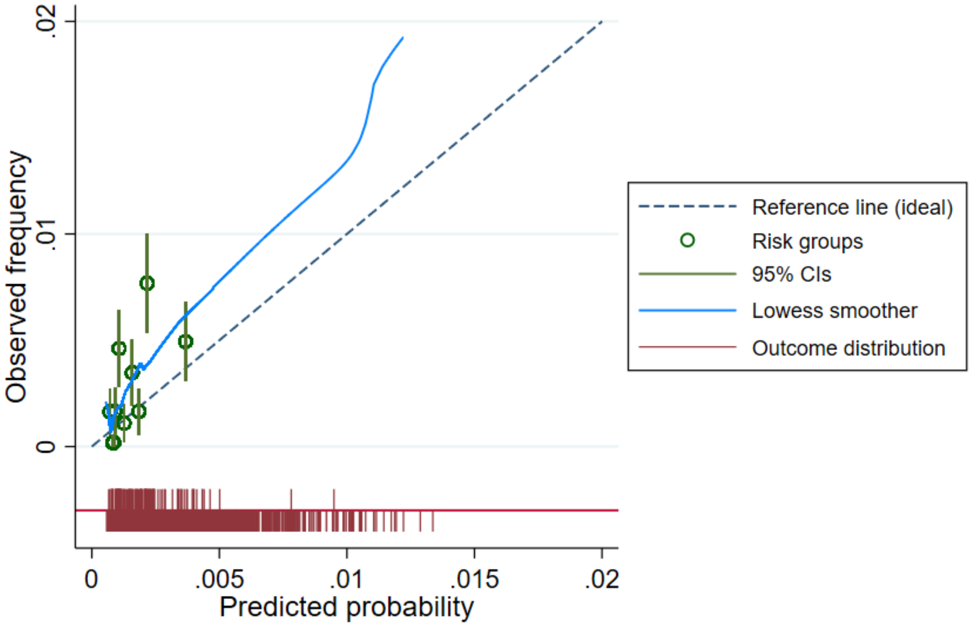


Figure 5.12: Calibration plot for model 2 - Chie dataset

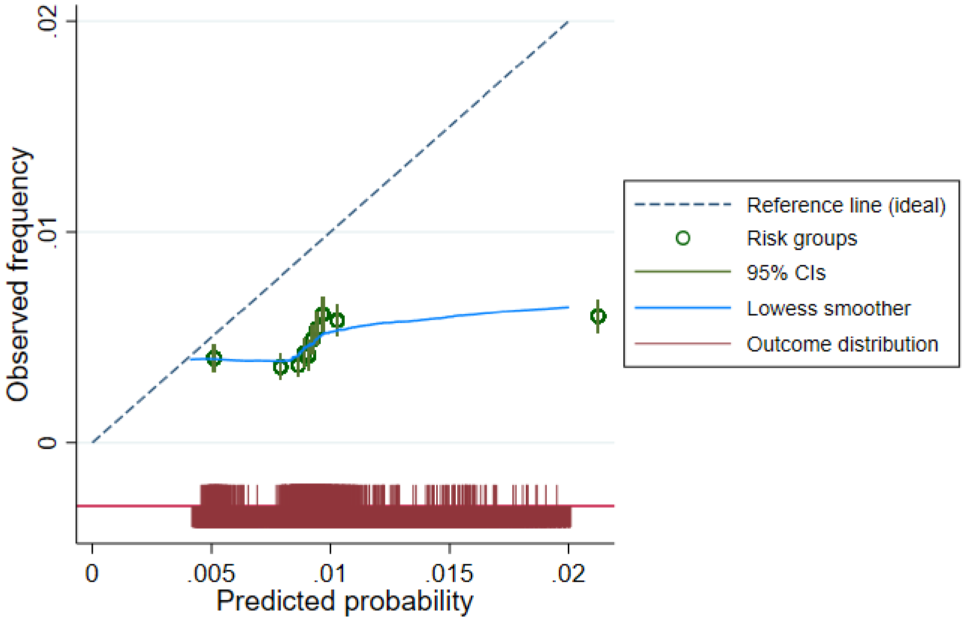
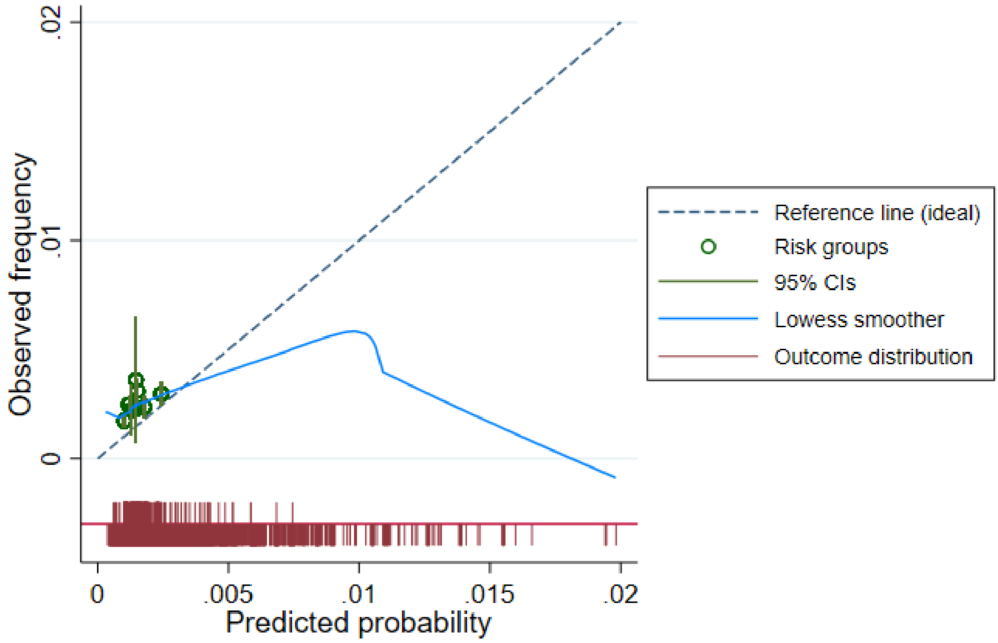


Figure 5.13: Calibration plot for model 3 - Chie dataset



5.3.6 Net benefit

Decision curves for each of the 3 models in the cohorts with more than 100 outcomes are presented below in Figure 5.14 - Figure 5.16. Comparison of models was not possible due to different outcomes being used in each model. It can clearly be seen from these plots that there is no net benefit to using the models over a treat all or treat none strategy in these cohorts. In fact, Figure 5.15 suggests that between the threshold probabilities of 0.005 and 0.1, model 2 shows a **net harm** compared with the treat none strategy.

Figure 5.14: Decision curves for model 1b - Stork dataset

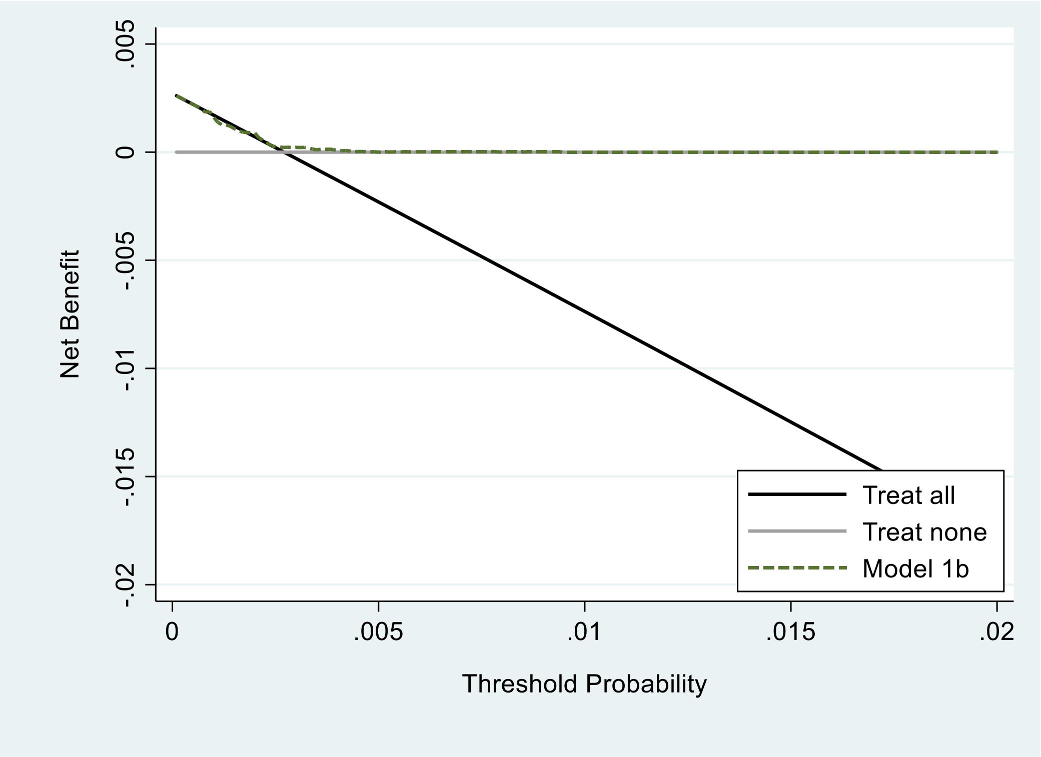


Figure 5.15: Decision curves for model 2 - Chie dataset

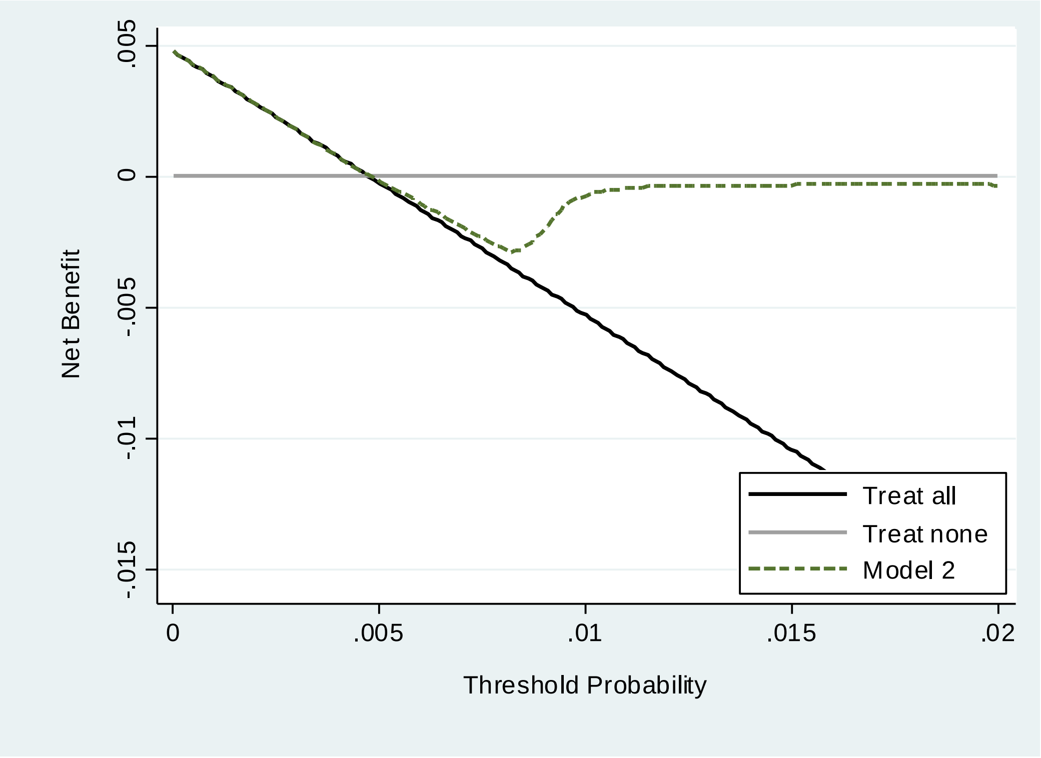
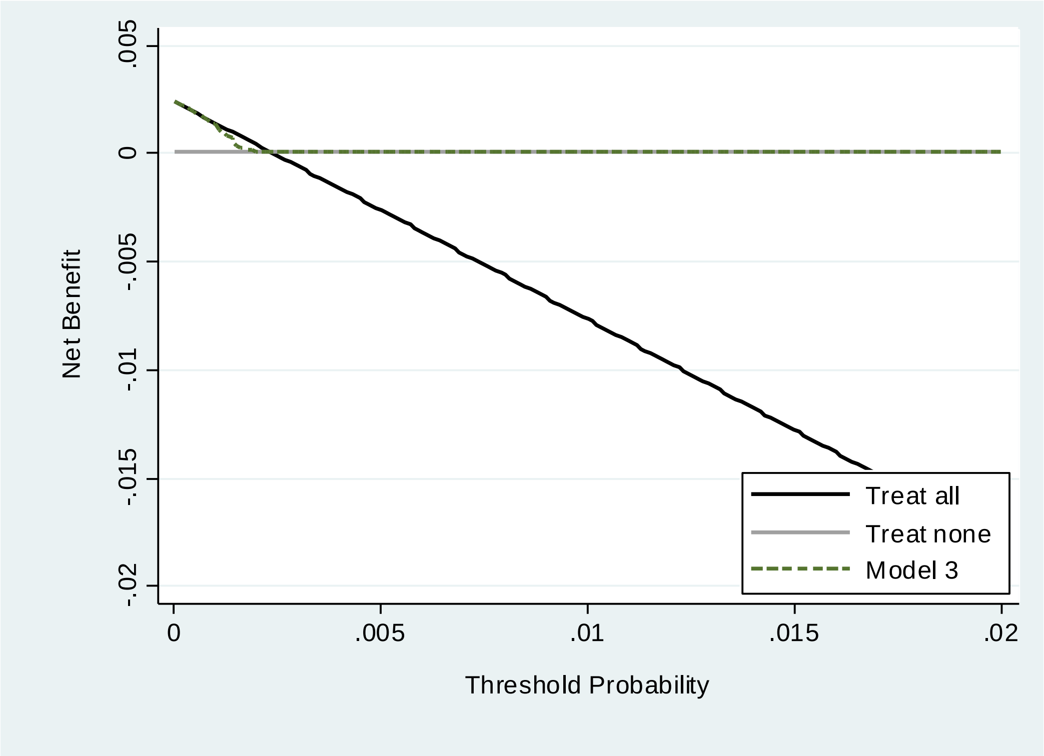


Figure 5.16: Decision curves for model 3 - Chie dataset



5.4 Discussion

5.4.1 Summary of the findings

This IPD meta-analysis has evaluated existing prediction models for stillbirth. Only a fifth of published stillbirth prediction models reported the model equation required for independent external validation. Three models, that were developed in high income countries, could be externally validated using cohorts from the IPPIC data repository. The models were mostly developed using maternal clinical characteristics, but one model additionally included ultrasound markers. PROBAST of the original model development articles suggested risk of bias concerns.

Overall, the findings in this chapter suggest that the models that were validated do not perform well in the external cohorts available. The IPD meta-analysis of model performance showed low discriminatory ability and poor calibration, with calibration slopes mostly <1. However, there was a lot of uncertainty around the results due to such small numbers of events. Although each of the three models could be validated in at least one cohort with over 100 events, confidence intervals of predictive performance were wide for the Smith 2007 model, suggesting further validation is needed for this model. For each of the models, predictions were also systematically too low or too high depending on the cohort used to validate it (calibraton-in-the-large≠0). The models had no clear clinical utility as assessed by DCA and may even have net harm.

5.4.2 Strengths and Limitations

This is the first known IPD meta-analysis to examine the external validation of stillbirth prediction models (Kleinrouweler et al., 2016, Townsend et al., 2021a). The use of IPD from multiple existing studies and cohorts allowed larger sample sizes, allowed for the evaluation of the predictive performance of each model, and enables the overall performance and heterogeneity in performance to be checked across multiple settings. Multiple imputation of predictors and outcomes was performed for each cohort separately, to avoid loss of useful information, and ensure any heterogeneity across cohorts was not masked (Rubin, 1987, White et al., 2011). Although the definition of stillbirth in the validation cohorts were standardised, stillbirth was defined differently in each model, which prevented a head-to-head comparison of model performance.

A significant limitation of this study is that only three of the 40 identified existing models were able to be validated. This was mainly due to the failure of the studies to adhere to reporting standards of publishing the model equation (Collins et al., 2015, Moons et al., 2015), despite only two of the models being published before the release of TRIPOD. Some cohorts used in the external validation had few observed cases of stillbirths, and only two had more than 100 events. Thus, even though the IPD was pooled, numbers of events were often still small, leading to some models with results that had wide confidence intervals. Predicted probabilities in the cohorts only went up to 3%, which makes it difficult for the models to discriminate between women who did and did not have the outcome. This further highlights the primary limitation of stillbirth research, which is the comparative rarity of the outcome.

Prediction models should be externally validated in a similar patient population to that it was developed in to assess the performance of model but can also be validated in a different population to provide information on the generalisability and transportability of the model to that patient population. The three models validated in this chapter were developed in UK or US populations, and many of the cohorts used for validation were also from either the UK, US or western European countries.

To allow as much information as possible to be included in the validation of the prediction models, multiple imputation was performed for data with up to 95% of data missing. Whilst 95% of the data is thought to be a very large proportion to impute, in practice if more data were available to impute from, it is likely that there would be more variation in the data, meaning that the performance of the prediction models validated using this data would be expected to be even poorer than observed.

Although it would have been interesting to assess the relatedness of the validation cohorts to the development studies, as suggested by Debray et al. (2015), it was not possible to do so for multiple reasons including; not having access to the development data, evaluating several models in multiple studies, and having different outcome definitions in some cases. However, IPD did offer the opportunity to assess heterogeneity in performance which could be in part due to case-mix differences.

The literature review in Chapter 3 found that most of the reviewed articles did not explicitly state when the predictors were measured. Chapter 4 then illustrated how the magnitude of the predictor-outcome associations and the prognostic model performance can depend on when the time-varying predictors were measured. This further extends to the timing of predictor measurement in studies used for external validation of previously developed prediction models, as is illustrated in this chapter. Ideally, the time of measurement of the predictors and outcomes should match the setting in which the model was developed. However, for this project, this was not always possible. For example, if first trimester values of BMI and weight were not available within the cohorts, then second trimester values were used instead. When a different time point is used for the collection of predictors than is the intended moment of use of the model in practice, as it was here, the study's predictor-outcome associations, and more importantly when developing a prediction model, the prognostic model performance, may be misleading. This could imply that the models being validated are not applicable for their intended purpose, which may not have been the case if the predictors had been measured at the intended moment of use.

5.4.3 Comparison to existing studies

External validation of prediction models are needed to confirm generalisability and transportability of a model in populations with different characteristics (Moons et al., 2012a). However, independent data with sufficiently large sample sizes of stillbirth and relevant predictors for external validation of models are not readily available. This is a factor on why none of the published models have been recommended for use in clinical practice (Collins et al., 2015). This meta-analysis obtained lower summary estimates for discrimination to that reported in the development datasets, although this might be due to chance as some confidence intervals were wide (e.g. Smith 2007), and so further research is recommended (Smith et al., 2007, Trudell et al., 2017, Yerlikaya et al., 2016). Some published stillbirth models report discrimination of >0.8 (Aupont et al., 2016, Kayode et al., 2016), but these studies either did not report the model equation needed for an independent external validation (Aupont et al., 2016), or did not provide enough information on predictors for external validation (Kayode et al., 2016). In most cases, the performance of a prediction model is often overestimated when only estimated in the dataset used to develop the model due to overfitting, especially when there are few outcomes relative to the number of predictors considered (Riley et al., 2019a, Riley et al., 2020). This study has highlighted several methodological shortcomings in the development of stillbirth prediction models, which is further reflected in the risk of bias assessment of the models.

5.4.4 Relevance to clinical care

The UK government and NHS launched a care initiative in a bid to halve stillbirth rates by 2025, which includes risk assessment as part of a wider care-bundle. The bundle does not include tools to help determine if a woman is at increased risk of stillbirth, instead individual factors have been identified to categorise women as low, moderate, or high risk of fetal growth restriction, the most frequent cause of stillbirth in the UK. An accurate tool to predict which woman is at increased risk of stillbirth would allow for personalised risk stratification in pregnancy, and enable clinicians to make decisions on closer surveillance, or timing of birth to prevent fetal death. It would also empower mothers to make informed decisions on their risk of stillbirth. This would be a more targeted approach than the currently used system of a generalised population level risk factor to identify women at risk of stillbirth. However, none of the models validated in this study had sufficient performance or clinical utility to be recommended for use in practice.

5.4.5 Recommendations for further research

Stillbirth prediction models that can be used in routine care would be especially valuable in low-and-middle-income countries, where the stillbirth burden is disproportionately high. Models that were unable to be externally validated here will need to be independently validated before they can be recommended for use. Apart from improvement in the model development process to reduce overfitting by using larger sample sizes and adjusting for optimism of the predictor effects (e.g. by post-estimation shrinkage or penalising the model coefficients), additional work is needed to identify novel prognostic factors for use in model development, to improve the discriminatory performance of prediction models (Riley et al., 2019b). A closer examination of existing stillbirth risk-factors could potentially allow inaccurate risk predictors to be abandoned which would enable clinical care and research to be focused on the highest value predictors.

Systematic reviews using aggregate data meta-analysis, currently represent the best available evidence on predictors of stillbirth, and have proposed several risk-factors to categorise women as high-risk (Townsend et al., 2021b). However, these studies are limited by heterogeneity in the data reported within the primary studies, such as in the definition of stillbirth (Townsend et al., 2021b). Existing primary studies are often small with imprecise estimates, and inconsistencies in confounding factors adjusted for in their analysis, which sometimes leads to contradictory factor-outcome associations. Large international cohorts are needed to collect richer data on risk-factors to enable development and validation of prediction models. To enable validation of the identified models, future primary studies and cohorts should record all key predictors being proposed in the models.

Whilst this study has explored validation of different stillbirth prediction models, stillbirth is the final endpoint of several heterogeneous antecedent pathways, with varying biological mechanisms involved (for example, those involving fetal growth restriction, and those secondary to diabetes, typically with a large for gestational age infant). It is possible that more than one model will be needed, either for prediction at different gestational ages, or for stillbirths with similar phenotypes.

5.5 Conclusions

This is the first assessment and independent external validation of published stillbirth prognostic models across multiple cohorts. Findings suggest methodological shortcomings in the development including overfitting of models. None of the three previously published stillbirth models validated in this study showed sufficient performance or clinical utility to be recommended for use in practice. The models all considered similar candidate predictors for model development, which suggest additional and better predictors (prognostic factors) of stillbirth may need to be identified. Further research to validate other existing models, and potentially to develop new models, is needed.

5.5.1 Remainder of the thesis

Completing an IPD meta-analysis project is a huge undertaking and commitment often with a vast amount of resources and time needed to complete it. As noted, this IPD meta-analysis still led to some results with wide confidence intervals, despite the pooling of data. If it was known in advance of collecting the IPD what the power of the project would be, it could either allow researchers to reconsider whether to invest in the project or give them reassurance that the project is worth investing in, depending on how large the power is expected to be.

Further, the rarity of the outcome and hence the very low predicted probabilities from the validated models, raises potential questions regards to whether it is appropriate to build a prediction model for stillbirth, particularly as the definition of the outcome is not standardised, or whether finding strong prognostic factors to guide stratification could be a more useful prognostic tool.

Regardless, identifying strong prognostic factors is the starting point for building useful prediction models, and IPD meta-analysis can help us identify these. Hence, the next chapter describes a method to calculate the power of an IPD meta-analysis, in advance of collecting the IPD, for a project which aims to synthesise the IPD to examine prognostic factor effects.

6 Calculating the power to examine prognostic factor effects when planning an individual participant data meta-analysis with a binary outcome

6.1 Introduction

6.1.1 Chapter outline

The previous chapter utilised IPD to validate existing prediction models for stillbirth, however, none of the validated models were found to have clinical utility and the predicted probabilities from the models were all extremely low (less than 3%) due to the rarity of stillbirth. Some results were also found to have wide confidence intervals, despite the use of meta-analysis to pool the IPD. IPD meta-analysis can also be used to identify prognostic factors, which are the starting point of developing clinically useful models, or as discussed in Section 5.5.1, could be useful prognostic tools themselves in the absence of clinically useful prediction models. Hence, this chapter sought to develop a method to calculate the power of a prospective IPD meta-analysis to detect an important prognostic factor.

6.1.2 Background

There is a growing demand for meta-analyses that utilise IPD for prognostic factor studies as the availability of IPD from existing studies can increase the quantity and quality of data (Riley et al., 2021b), which in turn can improve the ability and power to examine the prognostic effects of a covariate, compared to single studies or a traditional meta-analysis of published aggregate data. However, IPD meta-analysis projects can be time consuming and resource intensive, requiring additional costs, time, and expertise than traditional aggregate data meta-analyses. An IPD meta-analysis project can take upwards of two years to obtain, clean, and harmonise, then meta-analyse the IPD. Hence, researchers and funders need reassurance that the project is worth their investment in time and cost and so the researchers involved should consider how many studies are likely to provide IPD and the power of an IPD meta-analysis using this data. Power and sample size calculations are seldom considered in protocols and publications of IPD meta-analysis projects, but if it was known in advance of collecting the data that the project would have high power, it could give reassurance to both the researchers and funders that the project is worth investing in. Conversely, if the planned IPD meta-analysis would have low power to detect a clinically important effect then the researchers may reconsider investing in the project.

Previous work has focused on calculating the power to identify a treatment-covariate interaction using an IPD meta-analysis (prior to IPD collection) for continuous (Ensor et al., 2018) and binary outcomes (Riley et al., 2022). This chapter modifies these methods to estimate the power of a planned IPD meta-analysis project, in advance of IPD collection, where the primary objective is to examine the effect of a (potential) prognostic factor on a binary outcome.

The outline of the chapter is as follows. Section 6.2 provides a foundation for the work presented by describing the two-stage approach to estimating a prognostic effect in an IPD meta-analysis with a binary outcome. Section 6.3 details the method for calculating the variance of a prognostic effect estimate in a single study, first for a binary covariate then for a continuous covariate. Section 6.4 proposes a three-step approach to calculating the power of the planned IPD meta-analysis of prognostic factors. Section 6.5 then provides two examples illustrating the methods described, which is followed by an extension to allow for heterogeneity in Section 6.6 and some discussion in Section 6.7. Stata code is provided in Appendix C.

6.2 A two-stage approach to estimating a prognostic effect in an IPD meta-analysis with a binary outcome

To provide the foundation for the power calculations that follow in subsequent sections, the following section describes the two-stage approach for estimating a prognostic effect parameter from an IPD meta-analysis of S studies with a binary outcome.

It is assumed IPD are available from multiple cohort studies. Cohort studies are generally deemed to be the most suitable for a prognostic factor study, as it would usually be unethical or impossible to randomise patients to different prognostic factors for a trial, and case-control studies may be prone to bias (as discussed in Chapter 1, Section 1.2.2).

In the first stage, the prognostic effect parameters are estimated using the IPD for each study individually. Then in the second stage, the prognostic effect estimates are pooled using a chosen meta-analysis model (Simmonds and Higgins, 2007). By only pooling prognostic effect parameters derived from within-study information (i.e. based at the participant-level), this approach automatically avoids study-level confounding and aggregation bias that may occur in meta-regression based on across-study information (Fisher et al., 2011, Thompson et al., 2010), or in one-stage IPD meta-analysis models that do not separate out within-study and across-study prognostic relationships. The two-stage approach can be implemented using *ipdmetan* in Stata (Fisher, 2015).

6.2.1 First-stage

Consider IPD are available for each of S cohort studies, containing a variable z_{ij} denoting a participant level (potential) prognostic factor of interest (e.g. the sex of participant j in study i), observed for all participants in each study, and a variable y_{ij} denoting a binary outcome of interest (i.e. $y_{ij} = 0$ or 1, where 0 denotes no event and 1 denotes event occurred). To estimate the prognostic factor parameter in each study separately, S logistic regression models could be fitted:

$$\begin{gathered} y_{ij} \sim \text{Bernoulli}(p_{ij}) \\ \ln\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \alpha_i + \gamma_i z_{ij} \end{gathered} \tag{6.1}$$

where p_{ij} is the probability of the outcome event for participant j in study i.

The model is usually estimated using maximum likelihood estimation (MLE), which is the focus here. The prognostic factor parameter for each study is denoted by γ_i, which represents the unadjusted log odds ratio (change in log odds) for a 1-unit increase in z_i. This first stage leads to S estimates of the parameter, one for each of the i = 1 to S studies included in the IPD.

For a continuous prognostic factor, the model assumes the effect of the factor on the outcome is linear. Although in practice non-linear trends should be modelled, for the power calculation that follows it is more pragmatic to assume linear effects. Note also that the model examines unadjusted effects. This is a starting point, and additional prognostic factors will be adjusted for in subsequent sections.

6.2.2 Second stage

The first stage produces S estimates of the prognostic effect parameter ($\hat{\gamma}_i$) and its variance ($var(\hat{\gamma}_i)$). In the second stage, the $\hat{\gamma}_i$ values are combined using either a common-effect model (i.e. the true prognostic effect is assumed the same in all studies, denoted by γ):

$$\hat{\gamma}_i \sim N(\gamma,\mathrm{var}(\hat{\gamma}_i)) \tag{6.2}$$

or a random-effects model (i.e. the true prognostic effects are assumed random across studies, drawn randomly from a normal distribution with a mean of γ and between-study variance of τ^2):

$$\begin{gathered}\hat{\gamma}_i \sim N(\gamma_i,\mathrm{var}(\hat{\gamma}_i)) \\ \gamma_i \sim N(\gamma, \tau^2)\end{gathered} \tag{6.3}$$

Maximum likelihood estimation can be used to fit (6.2), whereas restricted maximum likelihood (REML) is recommended to fit model (6.3). The summary estimate of γ will be a weighted average, and it summarises the difference in the log odds in participants with a one unit increase in z.

For the common-effect model, the variance of the summary prognostic effect parameter is

$$\mathrm{var}(\hat{\gamma}) = \frac{1}{\sum_{i=1}^{S}\left(\mathrm{var}(\hat{\gamma}_i)\right)^{-1}} \tag{6.4}$$

where S is the total number of studies in the IPD meta-analysis.

For the random-effects model, the variance of the summary prognostic effect parameter is

$$\mathrm{var}(\hat{\gamma}) = \frac{1}{\sum_{i=1}^{S}(\mathrm{var}(\hat{\gamma}_i) + \hat{\tau}^2)^{-1}} \tag{6.5}$$

Here, each study's weight depends on the sum of two estimated variances: the variance of the studies prognostic effect parameter ($var(\widehat{\gamma_i})$) and the REML estimated between-study variance of prognostic effects ($\hat{\tau}^2$). The smaller the $var(\widehat{\gamma_i})$ for a study, the more weight it has in the meta-analysis.

To consider the potential power of an IPD meta-analysis project, the expected value of the variance of $\hat{\gamma}$ ($var(\widehat{\gamma_i})$) needs to be determined in advance. Fundamentally, this depends on the study variances ($var(\widehat{\gamma_i})$), and so Section 6.3 describes how these may be ascertained in advance of IPD collection.

6.3 Calculating the variance of an unadjusted prognostic effect estimate for a binary outcome in a single study

In this section, analytical (closed-form) solutions for $var(\hat{\gamma}_i)$ in a single study, based on Fisher's Information matrix, are described. These solutions are challenging to obtain as $var(\widehat{\gamma_i})$ will be correlated with the value of $\hat{\gamma}_i$ itself (unlike for continuous outcomes). For generalised linear models such as the logistic regression model, each participant-level variance is a function of the participant's predicted outcome values from the fitted model. So rather than considering one variance term per study (as for continuous outcomes), a separate variance term (σ_{ij}^2) is required for each participant, conditional on their covariate values. For binary outcomes, a participant's response variance is $p_{ij}(1 - p_{ij})$ and thus depends on their expected outcome probability (p_{ij}), which is conditional on the baseline risk in the study and the prognostic effect of any covariates. The analytic solutions for $var(\widehat{\gamma_i})$ derived by Demidenko et al. (2008) have previously been extended by Riley et al. (2022) to calculate the power of examining treatment-covariate interactions when planning an IPD meta-analysis of randomised trials with a binary outcome. The following sections propose methods to use and amend this previous work to enable the power to be calculated when planning an IPD meta-analysis to examine prognostic effects in studies with a binary outcome. These methods are first derived for a binary prognostic factor, and then for a continuous prognostic factor.

6.3.1 Binary prognostic factor

Let z_{ij} be a binary covariate, such as z_{ij} = 1 for males and z_{ij} = 0 for females. After fitting the logistic regression model in equation (6.1) to the IPD in a single study, the variance of $\hat{\gamma}_i$ is:

$$\mathrm{var}(\hat{\gamma}_i) = \boldsymbol{I}_{\boldsymbol{i}}^{-1}(2{,}2)/n_i \qquad \textbf{(6.6)}$$

where n_i is the total sample size of study i, and $\boldsymbol{I}_{\boldsymbol{i}}^{-1}(2{,}2)$ denotes the 2,2 element of the inverse of Fisher's unit information matrix ($\boldsymbol{I}$). The word 'unit' refers to it being independent of sample size.

Let the design matrix $\boldsymbol{X} = (1, z)'$ (dropping the i and j for notational simplicity), then the 2 by 2 unit information matrix for a particular study can be expressed as:

$$\boldsymbol{I} = E_z(p(1-p)\boldsymbol{XX}') \qquad \textbf{(6.7)}$$

where

$$p = \frac{\exp(\alpha + \gamma z)}{1 + \exp(\alpha + \gamma z)}$$

and

$$\boldsymbol{XX}' = \begin{bmatrix} 1 & z \\ z & z^2 \end{bmatrix}$$

As $z = 0$ or $z = 1$ for a binary prognostic factor, then this can be simplified to:

$$\boldsymbol{XX}' = \begin{bmatrix} 1 & z \\ z & z \end{bmatrix}$$

Hence

$$
\begin{aligned}
\boldsymbol{I} &= E_z\left(\frac{\mathrm{exp}(\alpha+\gamma z)}{1+\mathrm{exp}(\alpha+\gamma z)}\left(1-\frac{\mathrm{exp}(\alpha+\gamma z)}{1+\mathrm{exp}(\alpha+\gamma z)}\right)\boldsymbol{XX'}\right) \\
&= E_z\left(\left(\frac{\mathrm{exp}(\alpha+\gamma z)}{1+\mathrm{exp}(\alpha+\gamma z)}-\frac{\mathrm{exp}(\alpha+\gamma z)^2}{(1+\mathrm{exp}(\alpha+\gamma z))^2}\right)\boldsymbol{XX'}\right) \\
&= E_z\left(\left(\frac{\mathrm{exp}(\alpha+\gamma z)(1+\mathrm{exp}(\alpha+\gamma z))}{(1+\mathrm{exp}(\alpha+\gamma z))^2}-\frac{\mathrm{exp}(\alpha+\gamma z)^2}{(1+\mathrm{exp}(\alpha+\gamma z))^2}\right)\boldsymbol{XX'}\right) \\
&= E_z\left(\frac{exp(\alpha+\gamma z)\big(1+exp(\alpha+\gamma z)\big)-\mathrm{exp}(\alpha+\gamma z)^2}{\big(1+exp(\alpha+\gamma z)\big)^2}\boldsymbol{XX'}\right) \\
&= E_z\left(\frac{(\mathrm{exp}(\alpha+\gamma z)+\mathrm{exp}(\alpha+\gamma z)^2-\mathrm{exp}(\alpha+\gamma z)^2)}{(1+\mathrm{exp}(\alpha+\gamma z))^2}\boldsymbol{XX'}\right) \\
&= E_z\left(\frac{\mathrm{exp}(\alpha+\gamma z)}{(1+\mathrm{exp}(\alpha+\gamma z))^2}\boldsymbol{XX'}\right) \\
&= E_z\left(\frac{\mathrm{exp}(\alpha+\gamma z)}{(1+\mathrm{exp}(\alpha+\gamma z))^2}\begin{bmatrix}1 & z\\ z & z\end{bmatrix}\right) \qquad \textbf{(6.8)}
\end{aligned}
$$

This can be expanded into a closed-form solution of:

$$\boldsymbol{I}=\frac{\mathrm{exp}(\alpha)}{(1+\mathrm{exp}(\alpha))^2}\boldsymbol{M}_1\,\mathrm{Pr}(z=0)+\frac{\mathrm{exp}(\alpha+\gamma)}{(1+\mathrm{exp}(\alpha+\gamma))^2}\boldsymbol{M}_2\mathrm{Pr}\,(z=1) \qquad \textbf{(6.9)}$$

where

$$\boldsymbol{M}_1=\begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix}\quad \boldsymbol{M}_2=\begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}$$

Thus, to derive the unit information matrix after fitting the logistic regression of equation (6.1) to a particular study, the assumed values of parameters α and γ need to be specified along with the probabilities $\Pr(z = Z)$, which are estimated as the proportion of participants in the study classified as $z = 0$ and the proportion classified as $z = 1$.

The asymptotic variance of the prognostic effect estimate can then be derived using equation (6.6).

This will be extended for the IPD meta-analysis setting in Section 6.4.

6.3.2 Continuous prognostic factor

For a continuous covariate, using both equation (6.6) and equation (6.8) again, the Fisher unit information matrix can be written as:

$$\begin{aligned} \boldsymbol{I} &= E_z\left(\frac{\exp(\alpha + \gamma z)}{(1 + \exp(\alpha + \gamma z))^2}\boldsymbol{X}\boldsymbol{X}'\right) \\ &= E_z\left(\frac{\exp(\alpha + \gamma z)}{(1 + \exp(\alpha + \gamma z))^2}\begin{bmatrix} 1 & z \\ z & z^2 \end{bmatrix}\right) \qquad \textbf{(6.10)} \\ &= E_z(\boldsymbol{B}) \end{aligned}$$

where $\boldsymbol{B}$ is a 2 by 2 matrix.

The expected value ($E_z(\boldsymbol{B})$) now depends on the distribution of the continuous covariate and on the values of the logistic regression parameters (α and γ). Hence, it is not possible to modify equation (6.8) into a closed form solution for $\boldsymbol{I}$. One way to derive $E_z(\boldsymbol{B})$ post estimation is to calculate each of the 4 components of $\boldsymbol{B}$ for each participant in the study using the estimated logistic regression parameters and then their means provide the expected values and thus form $\boldsymbol{I}$. Then the asymptotic variance of the prognostic effect parameter can be derived using equation (6.6).

This will be extended to the IPD meta-analysis setting in the next section, to consider how to proceed before IPD are obtained by making distributional assumptions about the prognostic variable of interest.

6.4 Calculating the power of a potential IPD meta-analysis project to estimate a prognostic effect with a binary outcome

The following sections outline a three-step process for calculating the power of an IPD meta-analysis project aiming to estimate a prognostic factor parameter for a binary outcome. The aim is to do this in advance of IPD collection, assuming that studies from which IPD are requested have not reported prognostic effects for the factor of interest (and their variances), and information regarding number of outcome events in each group of the binary prognostic factor is also not available (as otherwise the unadjusted prognostic effect estimate could be derived based on the published aggregate data).

The overall power of the IPD meta-analysis is a function of the estimated variances of the study-specific prognostic factor effects ($\mathrm{var}(\widehat{\gamma_i})$), rather than simply the sum of the power of each study.

Step 1 describes how to derive an estimate of the variance of the prognostic factor (var($\hat{\gamma}$)) for each study using routinely reported aggregate data from study publications, alongside assumptions about the prognostic effect size in each study and (for continuous factors) the distribution of the prognostic factor. Step 2 uses these estimated variances to derive an estimate of the meta-analysis summary result for the prognostic effect parameter. Then step 3 derives the power of the planned IPD meta-analysis using the values obtained in step 1 and step 2. A method to adjust the power calculations for the presence of additional correlated covariates is then described.

6.4.1 Step 1: Estimate the variance of the prognostic factor separately for each study in the planned IPD meta-analysis

6.4.1.1 Binary prognostic factor

The first step is to apply equation (6.9) in each study promising IPD, followed by equation (6.6) to obtain an estimate of $\mathrm{var}(\hat{\gamma}_i)$.

To approximate this before IPD collection, the following aggregate data is needed from each study:

- Total participants in the study (n_i)
- Total number of events (e_i)
- Total participants with $z_{ij} = 1$ ($n_{i,z=1}$)
- Total participants with $z_{ij} = 0$ ($n_{i,z=0}$)

Assumptions also need to be made about the values of parameters α and γ. For the key parameter (γ_i), as is suggested in (Riley et al., 2019b, Riley et al., 2022, Schmoor et al., 2000), it is advised to identify a minimally important value via discussion with clinical experts within the IPD meta-analysis project team. It is also possible to consider a range of potential values of γ_i to assess the impact on the change in power dependent on the assumed value of γ_i. It is simplest to assume γ is common for all studies (i.e. $\gamma_i = \gamma$).

Based on the assumed value of γ, and the aggregate data extracted, α can be estimated using the number of outcome events, the total number of participants and the proportion of z_{ij} = 1.

This requires some algebra. If $p_{i,z=0}$ is defined as the risk (i.e. number of events / total number of participants) of the event occurring in participants with z = 0 in study i, and $p_{i,z=1}$ is the risk of the outcome occurring in patients with z = 1 in study i, then by definition:

$$\begin{aligned}\alpha_{\mathrm{i}} &= \ln\left(\frac{p_{\mathrm{i,z=0}}}{1-p_{\mathrm{i,z=0}}}\right)\\ &= \ln\left(\frac{p_{\mathrm{i,z=1}}}{1-p_{\mathrm{i,z=1}}}\right)-\gamma_{\mathrm{i}}\end{aligned} \qquad \textbf{(6.11)}$$

Note also that a weighted average of $p_{\mathrm{i,z=0}}$ and $p_{\mathrm{i,z=1}}$ can be taken to give an approximation of the overall log odds,

$$\ln\left(\frac{p_i}{1-p_i}\right)=\frac{\left(\ln\left(\frac{p_{\mathrm{i,z=0}}}{1-p_{\mathrm{i,z=0}}}\right)n_0+\ln\left(\frac{p_{\mathrm{i,z=1}}}{1-p_{\mathrm{i,z=1}}}\right)n_{\mathrm{i,z=1}}\right)}{n_i} \qquad \textbf{(6.12)}$$

where p_i is the overall risk in study i, which is assumed to be available alongside n_0 and n_1.

By rearranging equation (6.12), an approximation of the log odds in participants with z = 1 can be derived (dropping the i notation for simplicity):

$$\ln\left(\frac{p_{z=1}}{1-p_{z=1}}\right) = \frac{\left(\ln\left(\frac{p}{1-p}\right)n - \ln\left(\frac{p_{z=0}}{1-p_{z=0}}\right)n_{z=0}\right)}{n_1}$$

$$= \frac{\left(\ln\left(\frac{p}{1-p}\right)n - \alpha n_{z=0}\right)}{n_{z=1}} \qquad \textbf{(6.13)}$$

Where p represents the overall risk, n is the total sample size, and n_0 and n_1 represent the numbers in each group. Equation (6.13) can then be substituted into equation (6.11) to obtain an estimate of α:

$$\alpha = \frac{\left(\ln\left(\frac{p}{1-p}\right)n - \alpha n_0\right)}{n_1} - \gamma$$

$$= \frac{\ln\left(\frac{p}{1-p}\right)n}{n_1} - \alpha\frac{n_0}{n_1} - \gamma$$

$$= \frac{\frac{\ln\left(\frac{p}{1-p}\right)n - n_1\gamma}{n_1}}{1+\frac{n_0}{n_1}}$$

$$= \frac{\frac{\ln\left(\frac{p}{1-p}\right)n - n_1\gamma}{n_1}}{\frac{1}{n_1}(n_1+n_0)}$$

$$= \frac{\ln\left(\frac{p}{1-p}\right)n - n_1\gamma}{n}$$

$$= \ln\left(\frac{p}{1-p}\right) - \frac{n_1}{n}\gamma. \qquad \textbf{(6.14)}$$

Based on the values of α_i (derived from (6.14)) and γ_i (assumed based on clinical discussion), and the necessary aggregate data extracted (i.e. p, n, and n_1) from each study, the equation (6.9) can then be applied followed by equation (6.6) to obtain an estimate of $\mathrm{var}(\hat{\gamma}_i)$.

6.4.1.2 Continuous prognostic factor

The approach to estimate $\mathrm{var}(\hat{\gamma}_i)$ for a continuous covariate is similar, but with the added complexity of having to specify the assumed distribution of the continuous prognostic factor. For simplicity, this might be assumed to be a normal distribution. But it does not need to be. The mean and standard deviation (SD) of key continuous prognostic factors are usually reported in study publications.

The aggregate data required from each study publication are:

- Total participants in the study (n_i)
- Number of outcome events (e_i)
- Characteristics to define the continuous prognostic factor's assumed distribution (e.g. mean and SD)

As with the binary covariate setting, assumptions are needed about the values of parameters α_i and γ_i. Centring the prognostic factor, z_{ij}, by its mean allows α_i to be approximated by the overall log-odds of the outcome in study i, which is simply a transformation of the overall risk which should be available from the study publication. As before, it is advised to identify a minimally important value of γ_i via discussion with clinical experts within the IPD meta-analysis project team or consider a range of values.

An estimate of $\mathrm{var}(\hat{\gamma}_i)$ for each study can then be obtained by estimating Fisher's information matrix as described in Section 6.3.2. To do this, the following is implemented (using Stata code) for each study:

1. Generate a large dataset (e.g. 1 million participants) that mimics the study aggregate data provided in terms of the proportion of patients with the outcome event and the distribution of z (e.g. a normal distribution with a specified mean and standard deviation);
2. Calculate $\boldsymbol{I} = E_z(\boldsymbol{B})$ conditional on the specified α and γ values for that study (equation (6.10))
3. Use equation (6.6) to calculate $\mathrm{var}(\hat{\gamma}_i) = \boldsymbol{I}^{-1}(2,2)/n_i$

6.4.2 Step 2: Estimate the variance of the prognostic factor from the planned IPD meta-analysis

Step 1 produces S estimates of $\mathrm{var}(\hat{\gamma}_i)$, one for each study. The variance of the summary prognostic factor parameter estimate from an IPD meta-analysis of these studies can then be estimated, depending on whether step 1 assumed γ_i was common or random across studies. When assuming γ_i is common ($\gamma_i = \gamma$), equation (6.5) can be used to calculate the anticipated estimate of $\mathrm{var}(\hat{\gamma})$ for the IPD meta-analysis project:

$$\mathrm{var}(\hat{\gamma}) = \frac{1}{\sum_{i=1}^{S}\left(\mathrm{var}(\hat{\gamma}_i)\right)^{-1}} \tag{6.15}$$

6.4.3 Step 3: Calculate the power of the planned IPD meta-analysis

The final step is to calculate the power of the planned IPD meta-analysis project to detect γ. Assuming a common prognostic factor effect for all studies, and based on a Wald-test and a 5% statistical significance level, the power is approximately:

$$\begin{aligned}\text{Power} &= \text{Prob}\left(\frac{\hat{\gamma}}{\sqrt{\text{var}(\hat{\gamma})}} > 1.96\right) + \text{Prob}\left(\frac{\hat{\gamma}}{\sqrt{\text{var}(\hat{\gamma})}} < -1.96\right) \\ &= \Phi\left(-1.96 + \frac{\hat{\gamma}}{\sqrt{\text{var}(\hat{\gamma})}}\right) + \Phi\left(-1.96 - \frac{\hat{\gamma}}{\sqrt{\text{var}(\hat{\gamma})}}\right)\end{aligned} \qquad \textbf{(6.16)}$$

Here, $\Phi(x)$ is the probability of sampling a value < x from a standard normal distribution, $\text{var}(\hat{\gamma})$ is the anticipated variance of the summary prognostic effect estimate (as obtained in step 2), and $\hat{\gamma}$ can be replaced with the assumed true γ (as defined in step 1). This power estimate is usually multiplied by 100 and reported as a percentage.

6.4.4 Adjusting for other covariates

The three-step method proposed in the previous section assumes that the prognostic effect of any other covariate is zero. However, in reality, it is likely that other covariates (e.g. existing prognostic factors) would be included in the model for adjustment and these additional factors may be expected to be correlated with the prognostic factor of primary interest. Hence, the existing power formulae described would not be valid (Schmoor et al., 2000).

Whittemore (1981) has shown that for continuous normal covariates, in the setting where there are multiple covariates, the variance of the prognostic factor of interest ($\text{var}(\gamma)$) can be approximated by inflating the variance of γ obtained in the one parameter model by the variance inflation factor (VIF). The variance inflation factor (VIF), ranging from 1 upwards, is a measure of the amount of correlation between a set of predictors in a model. The VIF is defined as:

$$VIF = \frac{1}{1 - \rho^2}$$

where ρ is the multiple correlation coefficient, the proportion of the variation in the dependent variable that is predictable from the independent variables, and ranges from 0 to 1. Hsieh et al. (1998) has shown that the same VIF also works well for binary covariates.

The VIF measures how much the variances of estimated regression coefficients are inflated when compared to having uncorrelated predictors. Therefore, to gain a more accurate estimate of the power of a planned IPD meta-analysis when there are other prognostic factors to be adjusted for, Step 1 can be completed as described for an unadjusted prognostic effect, but then prior to completing Step 2, each of the S estimates of $\mathrm{var}(\hat{\gamma}_i)$ should be multiplied by the VIF to provide estimates of the inflated variances. These inflated estimates of $\mathrm{var}(\hat{\gamma}_i)$ can then be used in Step 2 to estimate the variance of the summary prognostic factor parameter which can then be used in Step 3 to calculate the power. To allow the estimation of the inflated variances, an assumption needs to be made regarding the value of the correlation coefficient. A pragmatic approach, when other information in unavailable, is to assume a moderate value of ρ such as 0.5.

6.5 Applied examples

The proposed methods are now applied to two examples for illustration.

6.5.1 Example 1: Prognostic effect of age and sex on gastrointestinal bleeding in patients with cirrhosis and oesophageal varices (Poynard)

The first example considers the power of an IPD meta-analysis conducted by Poynard et al.(1991) and is an example considered in previous methodology papers (Kovalchik and Cumberland, 2012, Simmonds and Higgins, 2007). The project aimed to examine the efficacy of beta-adrenergic-antagonist drugs in the prevention of gastrointestinal bleeding for patients with cirrhosis and oesophageal varices. IPD were obtained from four randomised trials involving a total of 286 patients randomised to active treatment and 383 to a control (placebo). However, here we pretend that IPD are not yet available, and the aim is to assess whether collecting IPD would allow sufficient power to examine prognostic factors. The data would essentially be analysed as a cohort study, and the focus is on estimating the prognostic effect of sex and age (individually) on the occurrence of gastrointestinal bleeding (rather than treatment effects as in the original trials).

The aggregate data from the trials are shown in Table 6.1. Aggregate data in the publications were given by treatment/control groups, however, as the data have been combined for the purpose of these examples, a weighted mean age was calculated as:

$$\text{overall mean} = \frac{n_C\mu_C + n_T\mu_T}{\text{total participants}}$$

Where n_C and n_T are the number of participants in the control and treatment groups respectively, and μ_C and μ_T are the mean age reported in the control and treatment groups, respectively. The corresponding standard deviation for age was calculated as (Mathematics Stack Exchange, 2018):

$$\mathrm{SD} = \sqrt{\frac{(n_C - 1)\mathrm{SD}_\mathrm{C}^2 + (n_T - 1)\mathrm{SD}_\mathrm{T}^2}{n_C + n_T - 1} + \frac{n_C n_T (\mu_C - \mu_T)^2}{(n_C + n_T)(n_C + n_T - 1)}}$$

Table 6.1: Aggregate data from 4 randomised trials included in the IPD meta-analysis project of Poynard et al. (1991)

Trial	Total participants	Total Events	Age in years: mean (SD)	Male, %
1	230	49	54 (10)	71
2	174	44	54 (11)	70
3	79	12	54 (8)	72
4	106	26	56 (11)	75

The question of interest here is: if it is assumed that this aggregate data could be obtained in advance (e.g. from study publications or investigators), then what is the estimated power of a planned IPD meta-analysis to estimate the effects of a particular prognostic factor? The factors of interest in this example are sex and age (individually), to provide an example with both a binary and a continuous covariate, and the three-step process described in Section 6.4 is used to undertake the power calculations.

As previously discussed, a value (or values) of γ need to be assumed to be able calculate the power. It will first be assumed that the prognostic effect of any other covariate is zero, then the examples will be repeated making an adjustment for the presence of additional correlated covariates.

For this example, a range of values from $\gamma = \ln{(0.5)}$ to $\gamma = \ln{(3)}$ were used for the sex covariate, which corresponds to an outcome odds that is 50% lower in males compared to females up to an outcome odds that is 3 times higher for males than females. For the age covariate, a range of values from $\gamma = \ln{(0.95)}$ to $\gamma = \ln{(1.05)}$ were used, which correspond to an odds ratio that is 50% lower for every 10-year increase in age, to an odds ratio that 50% higher for every 10-year increase in age. Age is assumed normally distributed in each study, with a mean and SD as given in Table 6.1.

A selection of the results of the power calculations between the range of assumed γ values are shown in Table 6.2, and plots of the calculated powers over the full range of assumed values are given in Figure 6.1 for sex and Figure 6.2 for age. The focus here is on unadjusted results, however, the results are presented following adjustment for other covariates in the following section. There is a power of 91% to detect the assumed prognostic effect of sex with an odds ratio of 0.5, and 99% power to detect an odds ratio of 3. The power decreases significantly the closer the odds ratio gets to 1, with only 8% power to detect an odds ratio of 0.9 and 17% power to detect an odds ratio of 1.25. For age, there is 99% power of detecting an odds ratio of 0.95 or 1.05, which again decreases significantly the closer the assumed odds ratio is to 1, with 19% power of detecting an odds ratio of 0.99 and 55% power of detecting an odds ratio of 1.02.

Table 6.2: Results of the power calculations for the Poynard example for a range of $\hat{\gamma}$ values

Sex			Age		
$\hat{\gamma}$	$\mathrm{var}(\hat{\gamma})$	**Power %**	$\hat{\gamma}$	$\mathrm{var}(\hat{\gamma})$	**Power %**
ln(0.5)	0.04332	91.47	ln(0.95)	0.0001011	99.92
ln(0.6)	0.04425	68.02	ln(0.96)	0.0000967	98.58
ln(0.7)	0.04529	38.83	ln(0.97)	0.0000935	88.30
ln(0.8)	0.04639	17.91	ln(0.98)	0.0000913	56.13
ln(0.9)	0.04752	7.72	ln(0.99)	0.0000900	18.51
ln(1.0)	0.04867	5.00	ln(1.00)	0.0000896	5.00
ln(1.25)	0.05155	16.59	ln(1.01)	0.0000900	18.24
ln(1.5)	0.05443	41.23	ln(1.02)	0.0000912	54.51
ln(2)	0.06003	80.76	ln(1.03)	0.0000933	86.45
ln(2.5)	0.06543	94.76	ln(1.04)	0.0000962	97.93
ln(3)	0.07063	98.51	ln(1.05)	0.0000999	99.83

Figure 6.1: Results of the power calculations for the Poynard example with sex as the prognostic factor for a range of gamma values (presented as odds ratios)

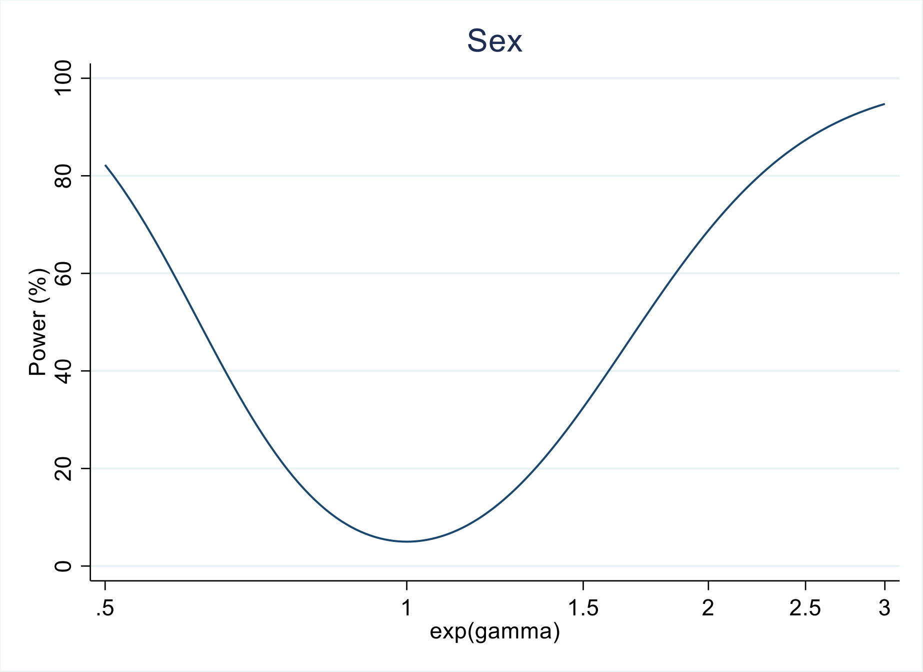
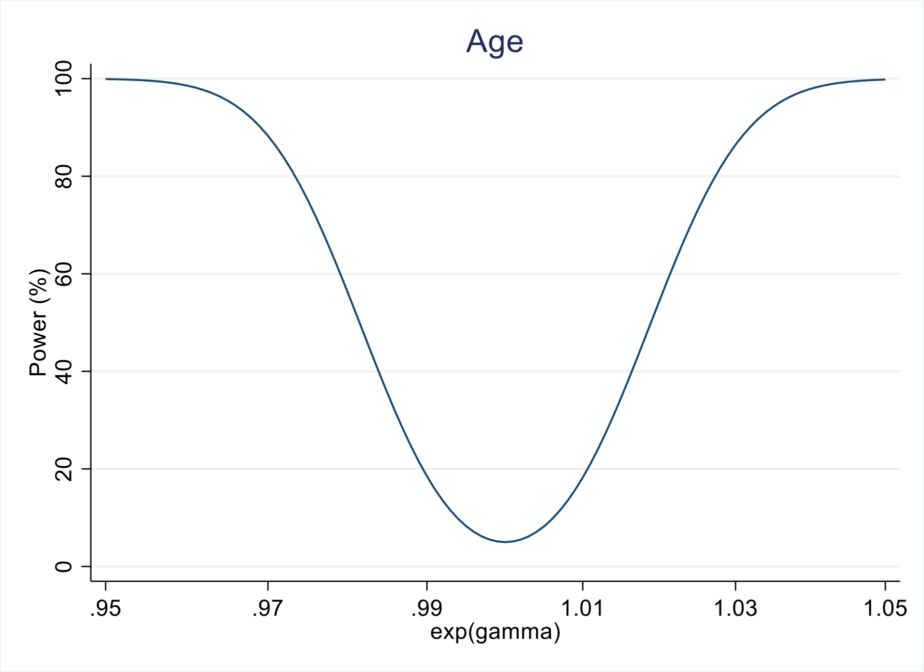


Figure 6.2: Results of the power calculations for the Poynard example with age as the prognostic factor for a range of gamma values (presented as odds ratios)



6.5.1.1 Adjusting for additional covariates

The scenario was then considered for the presence of additional covariates that were correlated with the prognostic factor of interest. The methods used for the unadjusted example were repeated, but the individual variances of γ_i for each study ($\mathrm{var}(\gamma_i)$) were multiplied by a VIF prior to calculating the variance of the summary prognostic factor effect. Three different values of ρ (0.25, 0.5 and 0.75) were used to calculate three VIFs, to assess the impact of varying levels of collinearity between the prognostic factors on the power.

A selection of the results of the power calculations after inflating the variances for the sex covariate, for the values of γ presented in the example with no additional prognostic factors, are given in Table 6.3 and the results for age are given in Table 6.4. Plots of the calculated powers over the range of values for each of the VIFs are given in Figure 6.3 for sex and Figure 6.4 for age.

The results show that inflating the variances by a VIF of 1.0666 (i.e. ρ = 0.25) has only a relatively small impact on the power calculations, lowering it by only 2.8 percentage points for an assumed OR of 0.6 for sex. However, as ρ is increased, the VIF has a greater impact on the power calculations, reducing the power by more than 50% when ρ is 0.75 for certain values of γ. For example, when the OR for sex is assumed to be 1.5, the power reduces from 41.23% when no adjustment for other covariates is made to 20.98% when adjustment is made with an assumed correlation coefficient of 0.75.

In practice, without other information, assuming a moderate correlation of 0.5 may be a pragmatic choice. In this scenario, this IPD meta-analysis project would be unlikely to provide enough power to test the prognostic ability of sex, as for an OR of 0.6, there would only be 56%, and the OR would likely be much closer to one than this in reality (Poynard et al. (1991) found a hazard ratio of 0.89). However, there may be enough power to detect a prognostic effect of age, dependent on the expected size of the effect. It is estimated that there would be 93% power to detect an OR of 1.04, which may be a reasonable OR to expect (however, in practice, this would require clinical input).

Table 6.3: Results of the power calculations for the sex covariate in the Poynard example for a range of $\hat{\gamma}$ values and different VIFs

$\hat{\boldsymbol{\gamma}}$	$\mathbf{var}(\hat{\boldsymbol{\gamma}})$	$\boldsymbol{\rho = 0,}$ **VIF=1**	$\boldsymbol{\rho = 0.25,}$ **VIF=1.0666**	$\boldsymbol{\rho = 0.5,}$ **VIF=1.333**	$\boldsymbol{\rho = 0.75,}$ **VIF=2.286**
ln(0.5)	0.04332	91.47%	89.70%	82.23%	59.60%
ln(0.6)	0.04425	68.02%	65.22%	55.69%	36.19%
ln(0.7)	0.04529	38.83%	36.81%	30.58%	19.83%
ln(0.8)	0.04639	17.91%	17.08%	14.61%	10.53%
ln(0.9)	0.04752	7.72%	7.54%	7.03%	6.18%
ln(1.0)	0.04867	5.00%	5.00%	5.00%	5.00%
ln(1.25)	0.05155	16.59%	15.84%	13.62%	9.96%
ln(1.5)	0.05443	41.23%	39.10%	32.49%	20.98%
ln(2)	0.06003	80.76%	78.21%	68.79%	46.47%
ln(2.5)	0.06543	94.76%	93.43%	87.33%	65.89%
ln(3)	0.07063	98.51%	97.94%	94.74%	78.06%

Table 6.4: Results of the power calculations for the age covariate in the Poynard example for a range of $\hat{\gamma}$ values and different VIFs

$\hat{\boldsymbol{\gamma}}$	$\mathbf{var}(\hat{\boldsymbol{\gamma}})$	$\boldsymbol{\rho = 0,}$ **VIF=1**	$\boldsymbol{\rho = 0.25,}$ **VIF=1.0666**	$\boldsymbol{\rho = 0.5,}$ **VIF=1.333**	$\boldsymbol{\rho = 0.75,}$ **VIF=2.286**
ln(0.95)	0.0001011	99.92%	99.86%	99.30%	92.14%
ln(0.96)	0.0000967	98.58%	98.03%	94.89%	78.39%
ln(0.97)	0.0000935	88.30%	86.22%	77.88%	54.92%
ln(0.98)	0.0000913	56.13%	53.48%	44.88%	28.76%
ln(0.99)	0.0000900	18.51%	17.64%	15.05%	10.78%
ln(1.00)	0.0000896	5.00%	5.00%	5.00%	5.00%
ln(1.01)	0.0000900	18.24%	17.39%	14.85%	10.67%
ln(1.02)	0.0000912	54.51%	51.89%	43.47%	27.85%
ln(1.03)	0.0000933	86.45%	84.22%	75.51%	52.58%
ln(1.04)	0.0000962	97.93%	97.21%	93.37%	75.36%
ln(1.05)	0.0000999	99.83%	99.72%	98.83%	89.77%

Figure 6.3: Results of the power calculations for the Poynard example with sex as the prognostic factor for different values of the VIF for a range of gamma values (presented as odds ratios)

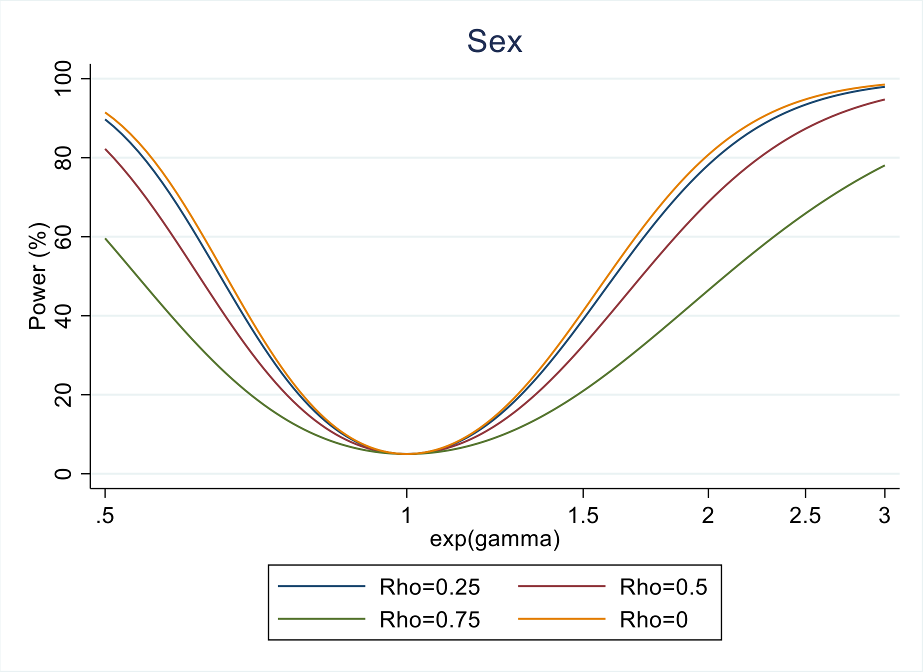
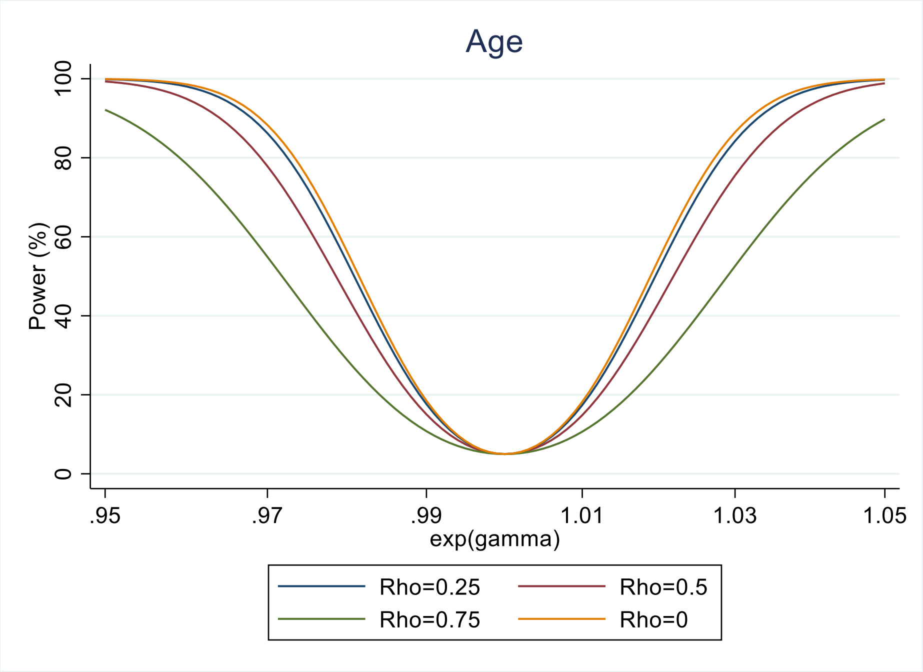


Figure 6.4: Results of the power calculations for the Poynard example with age as the prognostic factor for different values of the VIF for a range of gamma values (presented as odds ratios)



6.5.2 Example 2: Prognostic effect of age and sex on pain improvement in patients with osteoarthritis (STEER-OA)

In the Subgrouping and TargetEd Exercise pRogrammes for OsteoArthritis (STEER OA) project (Holden et al., 2017), IPD were collected from existing randomised trials to examine potential participant-level characteristics that interact with the effect of exercise interventions among people with knee and/or hip OA. Although pain and function outcomes were mostly analysed on a continuous scale, one binary outcome of interest was whether the patient had a reduction of pain (compared to baseline) by 3-6 months. There was a total of 31 trials that provided their IPD, and as in the previous example, the treatment and control arms are combined here to investigate the potential power of using this IPD to examine whether there is a prognostic effect of sex or age. The aggregate data needed to complete the three-step process described above for each trial is provided in Table 6.5. The mean and SD for age were calculated as described in the previous example due to being given for the treatment and control group separately in the trial publication. A range of values for γ were again considered and age was assumed normally distributed with mean and SD as given in the aggregate data. The same range of values of γ were used as in the Poynard example, however different values of γ are given in the table of results to illustrate a greater range in the calculated powers. The power calculation results are shown for the full range of assumed γ values in the plots. Two of the trials did not record sex, and one had only females, hence these three trials were not included in the power calculations for sex.

Table 6.5: Aggregate data from 31 randomised trials included in the IPD meta-analysis project of STEER OA (Holden et al., 2017)

Trial	Total participants	Total Events	Age in years	Male, %
1	210	127	64.96 (11.69)	28.10
2	48	24	66.17 (7.45)	29.17
3	89	48	64.55 (8.30)	51.69
4	199	104	61.74 (5.69)	35.18
5	222	115	63.38 (8.63)	31.08
6	312	111	69.74 (6.53)	37.18
7	126	62	64.07 (8.93)	24.60
8	152	102	70.18 (6.22)	26.32
9	88	44	61.08 (9.67)	35.23
10	39	17	74.23 (6.31)	23.08
11	217	133	68.08 (8.25)	35.02
12	48	17	63.25 (8.22)	18.75
13	71	42	62.45 (8.71)	32.39
14	105	52	65.32 (5.47)	17.14
15	418	264	66.60 (8.40)	29.67
16	218	128	58.68 (10.12)	59.17
17	107	49	64.58 (8.50)	44.86
18	158	84	68.82 (6.33)	NA
19	80	43	57.73 (4.12)	NA
20	87	59	63.86 (2.37)	0
21	28	19	67.68 (6.52)	46.43
22	32	21	72 (6.06)	12.50
23	109	65	57.81 (9.88)	45.87
24	109	44	67.72 (7.17)	31.19
25	40	25	66.62 (7.17)	20
26	34	21	70.18 (5.75)	23.53
27	203	130	65.39 (9.12)	41.38
28	391	188	61.68 (9.57)	36.06
29	55	36	78.91 (7.55)	27.27
30	200	132	68.03 (8.83)	21.50
31	46	11	67.17 (7.50)	56.52

A selection of the results of the power calculations between the range of assumed γ values are given in Table 6.6. Plots of the calculated powers over the range of assumed prognostic effect values are given in Figure 6.5 for sex and Figure 6.6 for age.

There is 100% power to detect a prognostic effect of sex with an odds ratio of less than 0.5 or higher than 1.75. These odds ratios relate to 50% lower odds of a reduction in pain at 3-6months for males compared to females, and 75% higher odds of a reduction in pain for males compared to females, respectively. The power starts to decrease as the odds ratio approaches 1, with a more dramatic decrease in power for odds ratios higher than 0.75, or lower than 1.3. There is 32% power of detecting an odds ratio of 0.9 and 51% power of detecting an odds ratio of 1.15.

For age, there is 99% power of detecting an odds ratio of 0.98 or 1.02, which decreases significantly between these values the closer the assumed odds ratio is to 1. There is 34% power of detecting an odds ratio of 0.994 and 12% power of detecting an odds ratio of 1.003. These correspond to 6% lower odds of reduction in pain at 3-6months for every 10-year increase in age, and 3% higher odds of a reduction in pain for every 10-year increase in age.

Table 6.6: Results of the power calculations for the STEER OA example for a range of γ̂ values

Sex			Age		
$\hat{\gamma}$	$\mathrm{var}(\hat{\gamma})$	Power %	$\hat{\gamma}$	$\mathrm{var}(\hat{\gamma})$	Power %
ln(0.5)	0.00508	100	ln(0.98)	0.0000155	99.92
ln(0.6)	0.00501	99.99	ln(0.99)	0.0000153	72.92
ln(0.8)	0.00497	88.62	ln(0.992)	0.0000153	53.83
ln(0.85)	0.00496	63.56	ln(0.994)	0.0000152	33.81
ln(0.9)	0.00497	32.13	ln(0.996)	0.0000152	17.69
ln(1.0)	0.00498	5.00	ln(1.00)	0.0000152	5.00
ln(1.15)	0.00500	50.63	ln(1.003)	0.0000152	11.98
ln(1.2)	0.00502	73.05	ln(1.01)	0.0000153	72.07
ln(1.25)	0.00503	88.23	ln(1.02)	0.0000155	99.89
ln(1.5)	0.00510	99.99	ln(1.03)	0.0000159	100
ln(1.75)	0.00519	100	ln(1.04)	0.0000164	100

Figure 6.5: Results of the power calculations for the STEER OA example with sex as the prognostic factor for a range of gamma values (presented as odds ratios)

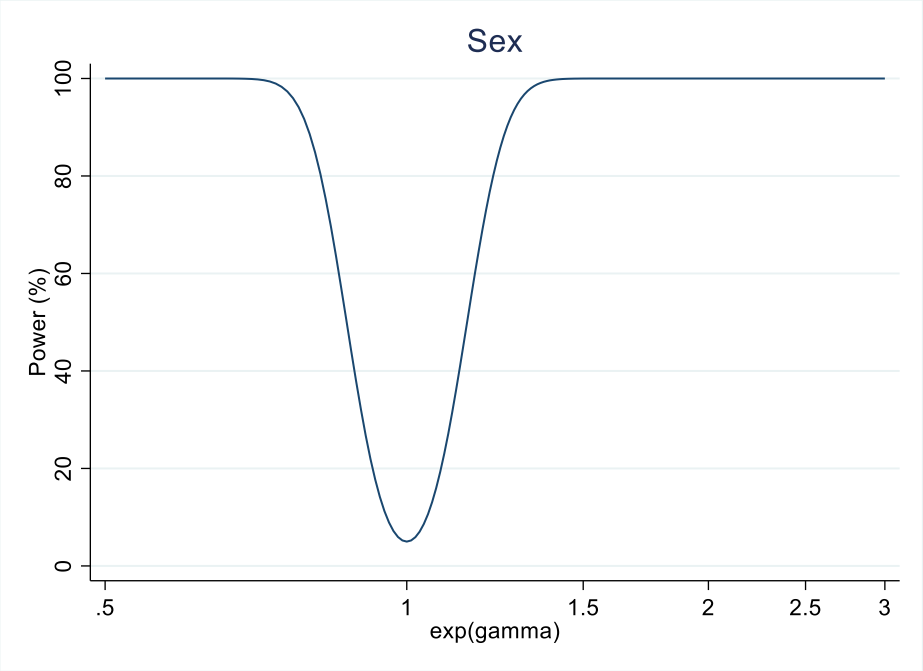
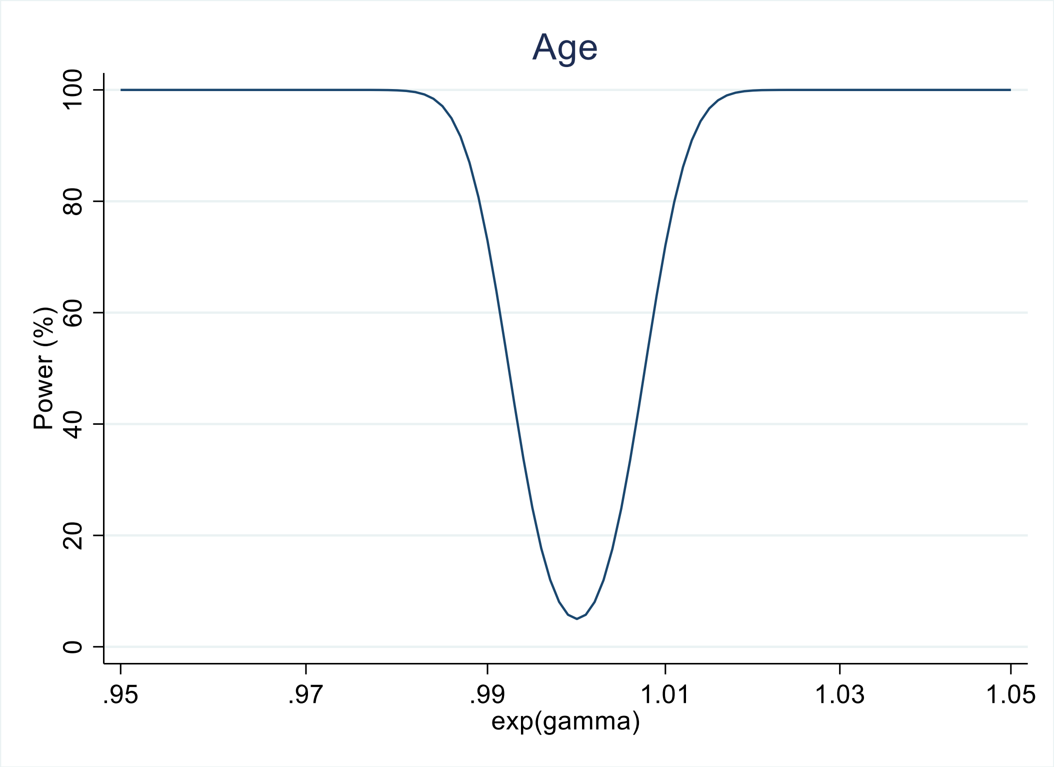


Figure 6.6: Results of the power calculations for the STEER OA example with age as the prognostic factor for a range of gamma values (presented as odds ratios)



6.5.2.1 Adjusting for additional covariates

The analyses conducted in this example were repeated with the supplement of adjusting for the presence of additional correlated covariates, as described in the previous example. A selection of the results of the power calculations after inflating the variances of the sex covariate, for the values of γ presented in the example with no additional covariates, are given in Table 6.7 and the results for age are given in Table 6.8. Plots of the calculated powers over the range of γ values for each of the VIFs are given in Figure 6.7 for sex and Figure 6.8 for age.

As in the Poynard example in Section 6.5.1.1, adjusting the power calculation for a correlation coefficient of ρ = 0.25 has only a modest impact on the power. However, as ρ

is increased, this has a considerable impact on the power available to estimate the prognostic factor of interest.

If we take a pragmatic approach again and assume ρ is 0.5, there would likely be enough power to conduct the IPD meta-analysis project if the prognostic effect of sex would be expected to have an OR above 1.25, which reflects a 25% higher odds of reduction in pain for males compared to females. However, if the relationship between sex and pain improvement is unlikely to be this large, then there would be insufficient power to conduct the project. For age, there would be sufficient power to detect an OR of 1.02 or above (or 0.98 and below), corresponding to a 20% higher (or lower) odds of pain reduction for each 10-year increase in age.

Table 6.7: Results of the power calculations for the sex covariate in the STEER OA example for a range of $\hat{\gamma}$ values and different VIFs

$\hat{\gamma}$	$\mathrm{var}(\hat{\gamma})$	$\rho = 0,$ **VIF=1**	$\rho = 0.25,$ **VIF=1.0666**	$\rho = 0.5,$ **VIF=1.333**	$\rho = 0.75,$ **VIF=2.286**
ln(0.5)	0.00508	100%	100%	100%	100%
ln(0.6)	0.00501	99.99%	100%	100%	100%
ln(0.8)	0.00497	88.62%	86.57%	78.30%	55.36%
ln(0.85)	0.00496	63.56%	60.78%	51.51%	33.23%
ln(0.9)	0.00497	31.13%	30.45%	25.35%	16.74%
ln(1.0)	0.00498	5.00%	5.00%	5.00%	5.00%
ln(1.15)	0.00500	50.63%	48.13%	40.18%	25.73%
ln(1.2)	0.00502	73.05%	70.28%	60.62%	39.86%
ln(1.25)	0.00503	88.23%	86.14%	77.78%	54.83%
ln(1.5)	0.00510	99.99%	99.98%	99.84%	96.36%
ln(1.75)	0.00519	100%	100%	100%	99.93%

Table 6.8: Results of the power calculations for the age covariate in the STEER OA example for a range of γ̂ values and different VIFs

$\hat{\gamma}$	$\text{var}(\hat{\gamma})$	$\rho = 0$, **VIF=1**	$\rho = 0.25$, **VIF=1.0666**	$\rho = 0.5$, **VIF=1.333**	$\rho = 0.75$, **VIF=2.286**
ln(0.98)	0.0000155	99.92%	99.91%	99.53%	93.55%
ln(0.99)	0.0000153	72.92%	72.39%	62.74%	41.50%
ln(0.992)	0.0000153	53.83%	53.34%	44.75%	28.68%
ln(0.994)	0.0000152	33.81%	33.47%	27.82%	18.19%
ln(0.996)	0.0000152	17.69%	17.53%	14.97%	10.73%
ln(1.00)	0.0000152	5.00%	5.00%	5.00%	5.00%
ln(1.003)	0.0000152	11.98%	11.90%	10.49%	8.17%
ln(1.01)	0.0000153	72.07%	71.54%	61.88%	40.83%
ln(1.02)	0.0000155	99.89%	99.88%	99.39%	92.66%
ln(1.03)	0.0000159	100%	100%	100%	99.89%
ln(1.04)	0.0000164	100%	100%	100%	100%

Figure 6.7: Results of the power calculations for the STEER OA example with sex as the prognostic factor for different values of the VIF for a range of gamma values (presented as odds ratios)

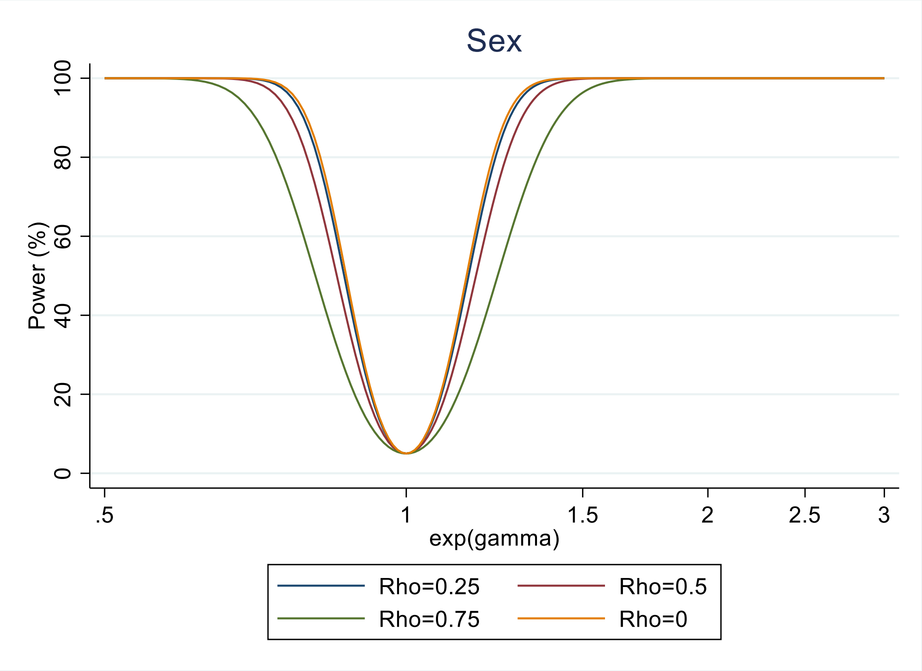
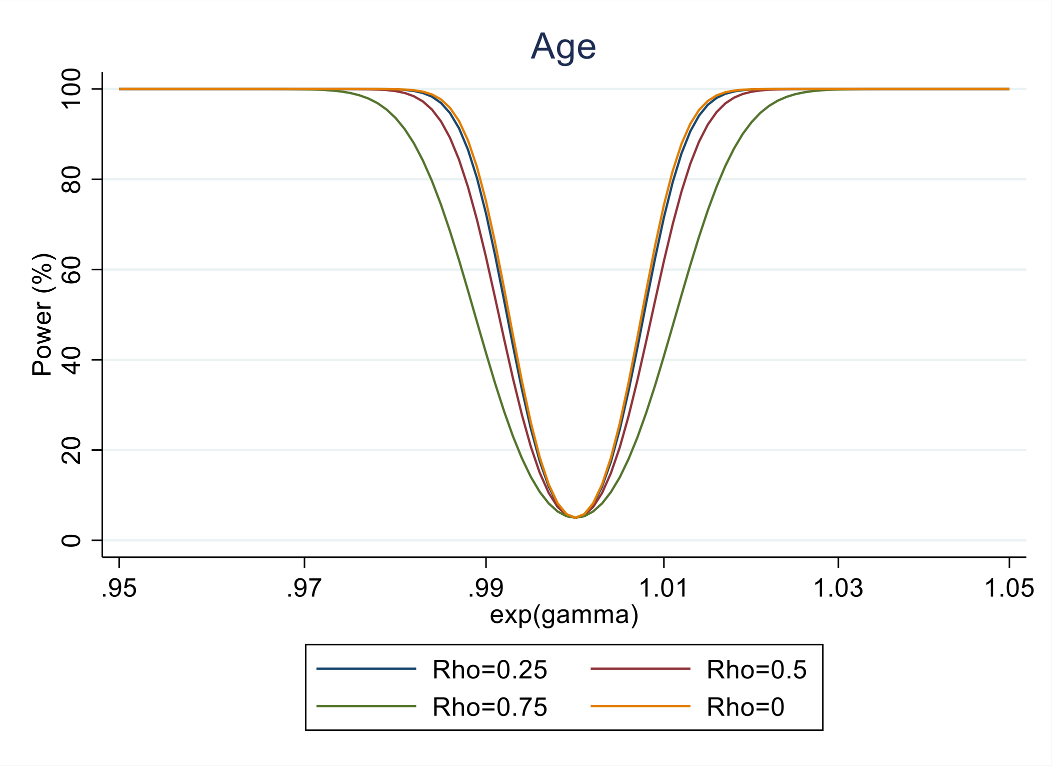


Figure 6.8: Results of the power calculations for the STEER OA example with age as the prognostic factor for different values of the VIF for a range of gamma values (presented as odds ratios)



6.6 Extension: Allowing for heterogeneity

The chapter so far has assumed a common-effect model in the second stage of the meta-analysis, which assumes the true prognostic effect is the same in each study. This section illustrates a method to allow for between-study heterogeneity in the prognostic factor effect, based on Riley et al. (2022) proposed approach for IPD meta-analysis of interactions.

To allow for between-study heterogeneity, a random-effects model must be assumed (equation (6.3)), but to do this, further assumptions must be made about the magnitude of the heterogeneity. The power calculation can be extended to allow for between-study heterogeneity in the prognostic effect:

$$\text{Power} = \mathrm{T}\left(-t_{S-1,0.975} + \frac{\hat{\gamma}}{\sqrt{\mathrm{var}(\hat{\gamma})}}\right) + \mathrm{T}\left(-t_{S-1,0.975} - \frac{\hat{\gamma}}{\sqrt{\mathrm{var}(\hat{\gamma})}}\right) \qquad \textbf{(6.17)}$$

where $\mathrm{T}(x)$ is the probability of sampling a value < x from a t-distribution with a mean of zero and $S-1$ degrees of freedom, and S is the number of studies expected to provide their IPD. The variance of the prognostic factor, $\mathrm{var}(\hat{\gamma})$, now needs to be estimated from equation (6.5), hence, an assumed value of $\hat{\tau}$ (the between-study standard deviation of the prognostic factor effect) must also be given.

As with the Hartung-Knapp-Sidik-Jokman (HKSJ) approach for deriving 95% confidence intervals after fitting a random-effects meta-analysis (Hartung and Knapp, 2001, Sidik and Jonkman, 2002), which uses a t-distribution rather than a normal distribution, a t-distribution is used here to help reflect the extra uncertainty due to $\hat{\tau}$ being estimated rather than already known.

Riley et al. (2022) suggest that equation (6.17) is likely to over-estimate the power as it assumes τ is known, when actually it will be estimated. This will be of greatest concern when there are small numbers of studies providing IPD, and when the true τ is close to zero, as then τ would be poorly estimated and often too high (as the estimate is bounded at zero). A simulation-based approach would be a better reflection of the uncertainty in that situation (Ensor et al., 2018).

Returning to the Poynard example from Section 6.5.1, Table 6.9 below shows the estimated power to detect a prognostic factor effect of γ=ln(1.04) for age, both before and after adjustment for other covariates. For the random-effects model, the power calculation from equation (6.17) was used, deriving the variance using equation (6.5) for a range of assumed

τ values. When adjusting for other covariates, the correlation coefficient (ρ) was assumed to be 0.5. For an assumed τ of 0.005, the power is now estimated to be 72.7% when not adjusting for other covariates, which is considerably lower than when assuming a common-effect model (97.9%). This is even further reduced when adjusting for other covariates, to 57.1% (from 93.4% for the common-effect model). As would be expected, the greater the value of τ, the greater the reduction in the estimated power. The drop off in power seems to be more dramatic for this example when τ increases above 0.005. In practice, if these estimates of power were calculated prior to collection of IPD, allowing for heterogeneity in the calculations would likely change the researchers/funders outlook on whether the IPD meta-analysis project would be worth their investment in both time and money.

Table 6.9: Comparison of the power in the common-effect model and random-effect model for an assumed OR for age of 1.04 in the Poynard example, considering a range of values for τ

τ	**Without** adjustment for other covariates	**With** adjustment for other covariates ($\rho = 0.5$)
Common-effect model		
-	97.93%	93.37%
Random-effect model		
0.001	76.45%	60.38%
0.0025	75.63%	59.63%
0.005	72.71%	57.05%
0.0075	67.91%	53.04%
0.01	61.53%	48.03%
0.015	46.64%	37.19%
0.02	33.56%	27.95%

6.7 Discussion

Evaluation of the power and sample size are an important aspect of planning and funding IPD meta-analysis projects. Previous work by Riley et al. (2022) extended the analytic solutions for the variance of the effect estimate ($var(\hat{\gamma_i})$), originally proposed by Demidenko (2008), to calculate the power of examining treatment-covariate interactions. In this chapter, these analytic solutions have been modified further to remove the treatment and interaction aspect, to be used to calculate the power when designing an IPD meta-analysis project with the primary objective of pooling studies to estimate the effect of a prognostic factor for a binary outcome. Additionally, this chapter has proposed a solution to adjust this power calculation for the presence of additional correlated covariates by using a variance inflation factor. Examples have been presented in this chapter demonstrating the use of these methods, and Stata code has been provided in Appendix C to enable the power calculation to be replicated. Further, the work in this chapter has demonstrated the impact of allowing for between-study heterogeneity in the power calculation, and the dramatic impact the choice of τ can have.

A three-step approach is proposed using an asymptotic solution for calculating variances of prognostic factor effect estimates, which allow the power of the planned IPD meta-analysis project to be calculated in advance of IPD collection, using aggregate data that are frequently reported in study publications. The three-step approach first uses the aggregate data to derive the Fisher's information matrix and an approximate estimate of the variance of each studies prognostic factor effect estimate, which then enables the variance of the summary effect estimate to be calculated from a two-stage IPD meta-analysis and finally the power of the IPD MA project can then be calculated.

If these results from the power calculations are known in advance of IPD collection, this would allow the researchers planning the project and the potential funders to decide whether the project would be worth their investment. It could also provide incentive to pursue IPD from additional studies, if they exist, to increase the power if necessary.

The methods and examples in this chapter give a calculation of the overall power of the planned IPD meta-analysis, including all the individual studies where IPD is expected to be available. However, it is also possible to estimate the individual power contribution of each study by using the estimate of the variance ($\mathrm{var}(\hat{\gamma}_i)$) from each study in the power calculation (equation (6.16)), rather than the estimate of the variance of the prognostic factor for the planned IPD meta-analysis as a whole. This would allow researchers to assess the added contribution of particular studies, which depends not only on the total number of participants and events, but also on the distribution and variance of the potential prognostic factor of interest in that study. This could inform decisions regarding where to focus efforts in terms of IPD collection if it is clear that certain studies would add very little to the power of the meta-analysis.

As with any power calculation, the approach is pragmatic to help gauge potential power under plausible assumptions and the actual power will change depending on various modelling assumptions. For example, the power could change if not all the promised IPD can be obtained. Other reasons that the power could change are if the assumed prognostic factor effects are incorrect, if the assumed distribution of a continuous covariate is wrong, if there is larger heterogeneity in prognostic effects than expected, or if the amount of correlation between the prognostic factor and adjustment covariates is incorrect. Hence, for funding applications it would be wise to display a range of power calculations based on a range of assumptions, as shown in the examples in this chapter.

A key issue when applying the proposed methods is the ability to obtain the necessary aggregate data for each of the potential studies to be included. Basic study information, such as the number of participants and number of outcome events should be available from study publications. However, information about covariate distributions may be more difficult to obtain, particularly for covariates other than the standard covariates such as age and sex, which are likely to be summarised in the baseline characteristics. In this situation, the study investigators can be contacted and asked to provide the summary information needed, which should hopefully be a reasonable request if they have already agreed to provide their IPD.

A further limitation of the proposed approach is the need to approximate α for the binary covariate scenario. It is approximated by using a weighted average of the risks in each group as an approximation for the overall log-odds of the outcome in study i, which can then be rearranged to approximate α. This is believed to be the best available approximation of α from the information available, as α is essentially the risk in group without the prognostic factor of interest, therefore is the overall risk minus the risk in the group with the prognostic factor. However, further work is needed to evaluate how robust the power calculation is to deviations from this approximation of α.

Consideration should also be given to the amount of correlation between the prognostic factor and adjustment variables, as it has been shown in the examples above that this can have a substantial impact on the power of the project, and therefore the presence of additional adjustment covariates should not be ignored when calculating the power of a planned IPD meta-analysis project.

When adjusting for additional covariates, the power calculations provided in this chapter do not consider the number of covariates included in the model, only the correlation between these additional covariates and the prognostic factor of interest. Therefore, when planning an IPD meta-analysis of this kind, consideration should be given to the number of covariates to be included in the model and whether the data to be obtained is large enough to avoid overfitting, or indeed whether all of the covariates are truly needed in the model.

For the main part of this chapter, a common-effect meta-analysis model was assumed, which assumes the true prognostic effect is the same in each study. The approach taken was for practicality, as if a random-effects model is assumed, as was demonstrated in Section 6.6, this would require assumptions to be made about the magnitude of the heterogeneity, which is difficult to ascertain. Section 6.6 provided an illustration of allowing for heterogeneity in one of the previous applied examples, which showed a dramatic change in the estimated power dependent on what value of τ is assumed. This further highlights the difficulty in allowing for heterogeneity in the power calculations. However, it also highlights the potential for drastically overestimating the power of the planned IPD meta-analysis project if heterogeneity is not accounted for.

6.8 Conclusions

This chapter has proposed new methods for calculating the power of a planned IPD meta-analysis project, in advance of IPD collection, which aims to evaluate the effect of a prognostic factor on a binary outcome. The approach has been illustrated using two examples, which highlight the need for choosing realistic assumptions when calculating the power, or indeed calculating the power for a range of assumptions as in these examples, as the calculated power dramatically changes depending on the assumed value of γ. When the focus is on added prognostic value of a factor, this chapter has also highlighted the need to account for additional adjustment covariates in the power calculations, which may be correlated with the prognostic factor of interest, as even a relatively small VIF can considerably reduce the power of the planned project.

7 Discussion

This chapter provides a critical discussion of the thesis. It begins with an overview of the thesis content, including a summary of each of the chapters and the publications that have arisen. A discussion of the contributions made to applied and methodological research is then presented, followed by consideration of the further research needs that now arise. Particular attention is given to the issue of measurement error in prognosis research and the potential implications on prediction models.

7.1 Overview of the thesis

Prognosis research is an important part of medical research as it seeks to understand and improve future outcomes in people with a given disease or health condition (Riley et al., 2013). The work in this thesis has focused on both prognosis research to identify prognostic factors, and on multivariable prediction models to predict a patient's future outcome risk, which can inform clinical decision making and help patients understand their risk.

The overall aims of the thesis were to apply and develop statistical methods for prognosis research. This has been achieved through chapters 2 to 6, in both single study and IPD meta-analysis settings, and led to multiple publications in statistical and clinical journals. Chapters 2 to 4 focused on the use of a single study for prognostic factor and prediction model research, with a particular emphasis on the measurement error that may be present within prognostic factors (predictors) and the impact of measuring a time-varying predictor after the intended moment of using the prediction model. Chapters 5 and 6 focused on the use of IPD from multiple studies for validating prediction models (Chapter 5) and calculating the power of an IPD meta-analysis to examine prognostic factor effects with binary outcomes, based on published study aggregate data (Chapter 6). Although the focus of the thesis was on prognostic factor research and risk prediction modelling, many of the same issues apply to diagnostic models (risk of disease being already present) and risk factor research (factors that increase the risk of disease onset). A short summary of the chapters is given below.

7.1.1 Summary of the chapters

The chapters in the thesis contained a mixture of clinical application and methodological development related to prognosis research. Chapter 2 showcased key statistical approaches for examining potential prognostic factors using an applied example. The clinical aim was to investigate the added prognostic value of potential prognostic factors for the development of complications in MC twin pregnancies, to improve knowledge of complications in MC twin pregnancies. The chapter also illustrated why it is not always sensible to develop a prediction model and demonstrated the instability of developing a prediction model with an insufficient sample size. Measurement error in the prognostic factors was unknown, and so only standard methods were applied to illustrate a typical prognostic factor study. However, some of the variables examined in the chapter were potentially subject to measurement error and the impact of this error on their prognostic value was unclear. This motivated the work in Chapter 3, in which a systematic review of prediction models was performed to ascertain how susceptible to measurement error the predictors used in prediction models are and how often the measurement error was acknowledged or accounted for within the development of the models. The review also examined whether the timing of predictor measurements was clearly stated, and if so, its relation to the intended moment of use of the prediction model. The review found that it is possible that many published prediction models include predictors that are measured with error, but that such error is often not accounted for or even considered. The review also found that the timing of measurements and the intended moment of using the model is often not explicitly stated. Therefore, Chapter 4 used a real example to illustrate the effect that measuring a time-varying predictor after the intended moment of using a prediction model has on the predictor-outcome associations (prognostic factor effects) and on the model performance. The direction and magnitude of predictor-outcome associations of a multivariable prediction model were compared under two scenarios: using a time-varying predictor of interest, ascertained by the treating physician at the point of care (i.e. the intended moment of use) and using the same predictor, but ascertained by a self-complete questionnaire mailed several days after the point of care. The results showed that displacing the collection of time-varying predictor information from the intended moment (and mode) of use can lead to substantial differences in the magnitude of predictor-outcome associations, and the subsequent accuracy of prognostic model performance.

The focus of chapters 2-4 was on prognosis research using a *single* study to either investigate potential prognostic factors or to develop a prediction model. However, there is a growing demand for meta-analyses that utilise IPD from multiple prognosis research studies, as this may offer novel opportunities for the development and validation of clinical prediction models, or for prognostic factor research, that may not be possible with the individual studies alone (Riley et al., 2021b). Therefore, chapters 5 and 6 focused on the use of IPD meta-analyses for prognosis research. Chapter 5 provided an applied example of using IPD from multiple studies to externally validate existing prediction models that have been developed across several population groups for predicting stillbirth. The predictive performance of three previously identified prediction models were assessed and compared using discrimination and calibration statistics. Decision curve analysis was used to assess the clinical utility of the prediction models, and the model performance was pooled and summarised across data sets using a two-stage IPD meta-analysis. The three models did not perform well in the external cohorts available, showing low discrimination and poor calibration and did not show sufficient clinical utility to be recommended for use in practice.

Completing an IPD meta-analysis project is a huge undertaking and commitment often with a vast amount of resources and time needed to complete it. The IPD meta-analysis in Chapter 5 led to some results with wide confidence intervals, despite the pooling of data. However, if it was known in advance of collecting the IPD what the statistical power of the project would be, this would inform decisions about the project's worth. For example, if the power is low then researchers may reconsider whether to invest in the project, but if the power is high it would give them reassurance that the project is worth investing in. Consequently, Chapter 6 described a novel method to calculate the potential power of an IPD meta-analysis, in advance of collecting the IPD, for a project that aims to synthesise IPD to examine prognostic factor effects.

7.1.2 Publications and outputs arising from this thesis

The work in this thesis has led to several publications. The prognostic factor results of Chapter 2 were published in *Diagnostic and Prognostic Research* (Mackie et al., 2019), for which I undertook all statistical analyses and contributed heavily to the write up. The findings of the work in Chapter 2 also contributed to the rationale for a paper showcasing the issues of instability of prognostic models in small datasets, for which I am a co-author (Riley et al., 2021a). I led a manuscript published in *Journal of Clinical Epidemiology* based on the systematic review results in Chapter 3 (Whittle et al., 2018), which shows the current neglect of issues of measurement error and timing of measurement. Another first-author article was published in *Diagnostic and Prognostic Research* describing the work in Chapter 4 (Whittle et al., 2017), highlighting the impact of different timings of predictor measurement. Finally, an article based on the results of Chapter 5 was published in *Ultrasound in Obstetrics and Gynecology*, for which I am joint first author (Allotey et al., 2022), for which I led all the statistical analyses and contributed greatly to the write up.

7.2 Contributions to applied and methodological research

This thesis has contributed to both applied and methodological research as outlined below.

7.2.1.1 Chapter 2

The applied prognostic factor study in Chapter 2 found that the discordance between nuchal translucency and the discordance between crown-rump length in monochorionic diamniotic twins is associated with a fetal adverse outcome, which remains after adjustment for standard prognostic factors. Discordance between babies CRL was also associated with IUFD and antenatally detected growth restriction, and discordance in NT was associated with the development of TTTS. This chapter also highlighted how it may not always be valuable to develop a prognostic model, even in a situation where some of the individual factors are identified as having prognostic ability. An example was provided of developing a prognostic model using the same dataset as that used for the prognostic factor study, where some of the predictors were not very prevalent and the outcome was quite sparse. The results demonstrated a large potential optimism in the model coefficients and the model's predictive performance, with the optimism-adjusted C-statistic much lower than the apparent C-statistic, emphasising the problem of overfitting due to a small sample size. Although adjustment for this overfitting can be done using a uniform shrinkage factor, this was shown to be unreliable, as the shrinkage factor was imprecisely estimated (again due to the small sample size). This finding motivated subsequent methodology work to show the issue of penalisation and shrinkage methods in small sample sizes (Riley et al., 2021a).

7.2.1.2 Chapter 3

A limitation of the work in Chapter 2 was that potential measurement error in the biomarker values was ignored, because such error information (e.g. from repeated biomarker values per individual, of a biomarker assumed to be in a stable state) was not collected and was not available from another source (e.g. a standalone re-test study). Furthermore, the impact and magnitude of this potential error was unclear. Therefore, in Chapter 3, a systematic review of prognostic models was performed to ascertain how susceptible to measurement error the predictors used in the final models were and how often the measurement error was acknowledged or accounted for within the development of the models. The review also investigated whether the timing of predictor measurement and intended moment of model use was clearly reported in articles developing clinical prediction models, and if they coincided.

The review found that many of the final prediction models included predictors that were likely to be susceptible to measurement error and this was often not accounted for or even acknowledged. Additionally, most of the articles did not state when the predictors were measured or the intended moment of using the model, which could mean that future users of the model unknowingly estimate misleading probabilities of a patients' outcome if they are using predictors measured at a different time than those used in the model development in relation to the timing of the model use. This motivated subsequent work in Chapter 4.

The review also found that while guidelines have been published providing a checklist for the reporting of prediction models (Collins et al., 2015), many researchers are still omitting vital information when publishing their work. Prediction models are developed to help guide clinicians in practice, and the majority of the models developed in the articles included here were intended to be used to assist clinicians in therapeutic decision making. Poor reporting will have an impact on researchers and practitioners who are planning to use an existing prediction model when assessing whether it is applicable to their situation. Journal reviewers and editors also need to be able to assess the generalisability of the model and the accuracy of the results, which may be difficult if it is not clearly reported within the article. Hence, articles poorly reporting the development of prediction models may not be implemented in practice or may provide poor predictions if used.

7.2.1.3 Chapter 4

The effect that measuring a time-varying predictor after the intended moment of using a prediction model has on the predictor-outcome associations (prognostic factor effects) and model performance was assessed and illustrated using a real example in Chapter 4. The chapter found that the magnitude of predictor-outcome associations and prognostic model performance can depend on when and/or how time-varying predictors are measured. In the illustrated example of patients presenting with musculoskeletal pain to general practice, associations between outcome risk and pain intensity recorded at the intended moment of use were lower in magnitude than those associations derived from a self-complete questionnaire mailed to patients up to one week later. The findings were replicated in two datasets with similar measurements, strengthening the belief that similar findings are likely across a range of painful non-inflammatory musculoskeletal disorders. Despite many published studies of musculoskeletal pain in primary care (Mallen et al., 2007), very few report the collection of time-varying predictor information by the GP at the initial point of care (Von Korff, 2013). When a later time is used, and/or with a different measurement method, the study's predictor-outcome associations and prognostic model performance may be misleading, and thus it could signal that the study is at high risk of bias and not applicable for its intended purpose. The findings of this chapter imply the need for caution when applying predictor-outcome associations or existing prediction models derived from studies that record time-varying predictors at a different time and/or measurement method than is intended upon clinical application, unless it has been externally validated in this setting. Previously developed prediction models that include time-varying predictors measured after the intended moment of use may overestimate the individual risk of experiencing the outcome of interest, which also reinforces the need for external validation using data that reflects the intended moment of use, and clear reporting of differences between validation and development data (Moons et al., 2015).

7.2.1.4 Chapter 5

Chapter 5 presented an applied example of externally validating existing prediction models which have been developed to predict stillbirth in pregnancy, using IPD collected from multiple studies. Only a fifth of the published stillbirth prediction models that were identified reported the model equation required for independent external validation. Hence, only three models were able to be externally validated. The overall findings from the chapter suggest that the models that were validated do not perform well in the external cohorts available, with none of the models showing sufficient performance or clinical utility to be recommended for use in practice. The IPD meta-analysis of model performance showed low discriminatory ability and poor calibration, with calibration slopes mostly <1. However, there was a lot of uncertainty around the results due to such small number of events. For each of the models, predictions were also systematically too low or too high depending on the cohort used to validate it (calibraton-in-the-large≠0). The models had no clear clinical utility as assessed by decision curve analysis and may even have net harm.

7.2.1.5 Chapter 6

The IPD meta-analysis in Chapter 5 led to some results with wide confidence intervals, despite the pooling of data. Given the huge undertaking and commitment that an IPD meta-analysis often is, in terms of both finance and time, it would be advantageous to know in advance of collecting the IPD what the power of the project would be expected to be. Hence, Chapter 6 described a method to calculate the power of an IPD meta-analysis in advance of collecting the IPD to examine prognostic factor effects. The method was also extended to enable the power to be adjusted for the presence of additional correlated covariates, and to allow for heterogeneity between studies. The approach was illustrated using two examples, which highlighted the need for choosing realistic assumptions when calculating the power, or indeed calculating the power for a range of assumptions, as the calculated power dramatically changed depending on the assumed value of the prognostic effect estimate. The chapter also highlighted the need to account for additional adjustment covariates in the power calculations, which may be correlated with the prognostic factor of interest, as even a relatively small VIF can considerably reduce the power of the planned project.

7.3 Further research needs

This thesis has identified several areas of further work and recommendations for future research, from both the applied and methodological aspects of the thesis, which are discussed below.

7.3.1.1 Chapter 2

Chapter 2 highlighted that although it was not possible to combine the prognostic factors identified to develop a prediction model that would perform well externally, there is some evidence of early physiological changes that may occur which would support the idea that first trimester prognostic factors exist. While some associations were found between the potential prognostic factors and adverse outcomes, as this was an exploratory study the evidence is not yet strong enough to recommend making any changes to practice, and replication would be needed to strengthen these findings. In particular, it would be of interest to collect data longitudinally, recruiting women with MCDA pregnancies in the first trimester, prior to the appearance of any clinical signs of complications, to enable comparison of those who do and do not develop complications. This would help determine whether the differences in the biomarkers are because the biomarker is abnormal earlier in pregnancy, or that it does not increase.

There are currently no established prediction models for predicting adverse outcomes in MCDA twins. Therefore, it would be valuable to collect more data, with the aim of combining the identified prognostic factors with standard prognostic variables to develop a prediction model, which would in turn allow clinicians to identify women who may be at a higher risk of adverse outcomes and provide appropriate treatment pathways. Since the completion of this chapter, Riley et al. (2020) have published guidance on how to calculate the sample size required to develop a clinical prediction model, and therefore this guidance could be used to determine the required sample size to develop a prediction model to predict adverse outcomes in MCDA twins. Whilst MCDA twins are rare, and therefore obtaining the necessary amount of data needed to develop a prediction model may prove difficult, collecting IPD from multiple sources in order to conduct an IPD meta-analysis could help increase sample size, in which would in turn improve the stability of any prediction model developed. However, there was still lots of uncertainty in the performance measures when using IPD to validate prediction models for stillbirth in Chapter 5, highlighting that even after collecting IPD from multiple studies, this may not provide sufficient data to produce precise estimates. It may be that the focus should be on finding strong, reliable, prognostic factors that can be used as indicators of high risk of adverse outcomes.

7.3.1.2 Chapter 3

The systematic review in Chapter 3 found that it is possible that many published prediction models include predictors that are measured with error, and this is often not accounted for or even considered. This suggests a need to assess whether ignoring measurement error in prediction models is a concern and whether accounting for the error will improve the predictions made and the model performance, and how the methods to account for measurement error could be implemented in a simple, easily interpretable way. Even if one (or more) of the estimates of a predictor-outcome association in a prediction model is biased due to measurement error, this may not be an issue if the model as a whole performs well in terms of the absolute risk predictions. However, more investigation is needed to determine the impact of different types of error under various scenarios, for a which a simulation study could be used to explore. The impact of measurement error in general is discussed further in Section 7.4. Researchers should be considering how susceptible to measurement error their predictors may be when developing a model and the impact this may have on subsequent performance in new data (in particular, calibration of prediction). Furthermore, consideration should be taken of potential measurement error when implementing previously published models in practice. However, if a prediction model is being used in clinical practice, it should have first been externally validated, which may reassure users of the model that any potential measurement error does not negatively impact the accuracy of predictions.

A related issue found within the review is that authors often do not specifically state when predictors are being measured or when the model is intended to be used, which is critical for using the model, and may have implications on the accuracy of predictions made (as shown in Chapter 4). Whilst ultimately, better reporting of the timings of predictors measurements and the intended moment of using prediction models is required, researchers and users of previously developed clinical prediction models need to be aware of potential disparities between the moment of time the model is being implemented and when the predictors used to develop the model were recorded, and how this may affect the accuracy of the predictions made.

The review also found poor reporting standards in many of the articles reviewed. Future research studies should be reported following the TRIPOD guidelines (Collins et al., 2015, Moons et al., 2015) to enable a better and more complete presentation of prediction models and their performance. Otherwise poor reporting will have a negative impact on researchers and practitioners who are considering using a prediction model and will hamper assessments of whether models are reliable and applicable to their situation. Additionally, articles poorly reporting the development of prediction models may not be implemented in practice or may provide poor predictions if used. Due to completing this PhD on a part-time basis, and having had two full years leave of absence since enrolling, it has been 7 years since the search for the systematic review was conducted. The search for the review was completed in November 2015, only shortly after the publication of the TRIPOD statement, hence it would be of interest to examine whether reporting standards have now improved 7 years on. However, recent reviews of prediction models for recurrent stroke in patients with transient ischaemic attack (TIA) and minor stroke (Abdulaziz et al., 2022), melanoma (Kaiser et al., 2022) and idiopathic pulmonary fibrosis (Di et al., 2022) have found poor adherence to the TRIPOD guidelines.

Clearly, better reporting of prediction models is needed, and this can be encouraged in several ways. Education is key to ensuring these recommendations are followed in practice, through training courses, conferences, dissemination and social media sites such as Twitter. Another vital approach is collaboration. Working with research teams conducting prediction model studies allows us to influence the methods and reporting conducted to enable them to be of high quality to ultimately develop and report better prediction models. Finally, we must ensure that we, as researchers in the field, produce high quality research using the recommended guidelines, such as TRIPOD, to enable researchers working on projects in the future to follow good examples of reporting and development.

7.3.1.3 Chapter 4

Chapter 4 found that the magnitude of predictor-outcome associations and prognostic model performance can depend on when and/or how time-varying predictors are measured. While the problem highlighted in this chapter is likely to extend to other commonly investigated predictors whose values are sensitive to the timing and mode of collection, this problem has only been demonstrated for one predictor and thus this remains to be evaluated more widely. A future study in which the same mode of data collection is used at the point of care and at post-consultation questionnaire (e.g. patient self-administered questionnaire) is needed to better understand the relative contribution of timing and mode of collection and therefore determine whether and how improved prediction is achievable at the point of care. Previous studies reporting results of predictor-outcome associations measured in this way may also need to be replicated.

Further research should assess whether similar findings are found with other time-varying predictors, and indeed in other clinical conditions and settings, and in predictors that can be measured via multiple methods (i.e., GP assessment or self-reported). Future research could also investigate whether external validation would alleviate the problem, however this would require the external validation to be performed using predictors measured at the same time, and via the same method, as what will be used in practice, to enable reassurance that the predictions being made will be accurate.

The work in this chapter investigates the effect of having one predictor included in the prediction model that is measured at a different time and via a different method to how and when it would be measured in practice, however, it is plausible that there would be multiple predictors to be included in a prediction model that face the same problems. Additional research is needed to evaluate the combined effect of including multiple of these predictors on the performance and whether the effect would be compounded, or whether there would be minimal impact beyond the first predictor included of this kind.

The findings in this chapter further reinforce recommendations made in Chapter 3. Researchers need to consider both the timing and the mode of predictor measurements when developing a prediction model, and whether these will be the same in practice. Additionally, users of clinical prediction models should be cautious when using a model with predictors measured at a different time to that of the intended moment of use of the model, or whether the predictors may have been measured by a different method.

7.3.1.4 Chapter 5

Chapter 5 externally validated existing stillbirth prediction models using IPD from multiple cohort studies. However, only three of the previously published models were able to be validated. The models that were unable to be externally validated here will need to be independently validated before they can be recommended for use, however many of them could not be validated due to not enough information being provided, and the required data were not available for others. This highlights the fact that an IPD meta-analysis is not always the solution, as key predictors for many models may not be available in the IPD. For those models that were able to be validated, there were large amounts of uncertainty around the estimates of the performance measures, again showing that IPD meta-analysis may still not provide enough power to give precise estimates, reinforcing the need to use aggregate data in advance of IPD collection to estimate whether the available data will provide enough power.

For those models that were unable to be validated due to not enough information being provided, authors could be contacted to request further information required, but for those models that the required data were not available, it is unlikely that data would be available elsewhere without collecting new prospectively collected data. Large international cohorts would be needed to collect richer data on potential prognostic factors to enable the development and validation of prediction models. To enable validation of the identified models, future primary studies and cohorts need to record all key factors being proposed in the models.

Additionally, work is needed to identify novel prognostic factors for use in the model development, to improve the discriminatory performance of prediction models (Riley et al., 2019b). A closer examination of existing stillbirth prognostic factors could potentially reveal some that are not prognostic and allow subsequent clinical care and research to be focused on those with the highest prognostic value.

Still birth is a rare, but very important, outcome. The rarity of stillbirth creates barriers to developing clinically useful prediction models, and it raises questions around whether it is appropriate to develop models in data with such few events, particularly as the definition of the outcome is not standardised across countries. It is however a very serious outcome and warrants research to improve outcomes, but more consideration is needed as to whether a prediction model will ever be able to separate out those at high risk when the overall risk for everyone is so low, or whether focussing on finding strong prognostic factors may be a better course of action.

Large amounts of missing data were recorded in the datasets used for validation, and missing data is also likely to be a problem when implementing the models in practice. Therefore, further research could look at pragmatic approaches to dealing with missing data in real time, such as the method proposed by Nijman et al. (2021), as in reality, patients are not always likely to have a measurement of every predictor needed for a model, which would then render the model unusable for that patient.

7.3.1.5 Chapter 6

Chapter 6 provided a method for estimating the power of an IPD meta-analysis with the aim of synthesising data to evaluate prognostic factor effects. If this power was known in advance of IPD collection, this would allow the researchers planning the project and the potential funders to decide whether the project would be worth their investment (Riley et al., 2021b). It could give funders the reassurance needed to fund a study which can potentially provide answers to questions that are unable to be answered without the use of IPD and may identify prognostic factors that could be useful for clinical practice, but that could also be used in future for developing a prediction model. The estimated power might also provide incentive to pursue IPD from additional studies, if they exist, to increase the power if necessary. Hence, it is recommended to estimate the power of the project in advance of data collection using the aggregate data that is usually available in study publications.

The estimates of the power could differ based on several variable assumptions that need to be made, such as if not all of the promised IPD could be obtained, or if the assumed prognostic factor effects are incorrect, if the assumed distribution of a continuous covariate is wrong or if the amount of correlation between the prognostic factor and adjustment covariates is incorrect. Hence it is recommended that, particularly for funding applications, it would be sensible to display a range of power calculations based on a range of assumptions. The calculations provided to estimate the power were also based on making an approximation of alpha, as the overall risk in the group without the prognostic factor of interest, so further work is needed to evaluate how robust the power calculation is to deviations from this approximation.

Guidance on how to calculate the required sample size to develop a prediction model (Riley et al., 2020) and to externally validate a prediction model (Snell et al., 2021) have recently been published, and these methods could be further investigated to consider whether a planned IPD meta-analysis would provide the required level of precision of performance measures for external validation studies and sufficient precision of the overall outcome proportion for studies aiming to develop a prediction model, as well as an adequate shrinkage factor (ideally >=0.9), small optimism and a small mean absolute prediction error (as discussed in Riley et al).

I have recently contributed to related work on this topic to calculate the power to examine treatment-covariate interactions when planning an IPD meta-analysis of randomised trials with a binary outcome (Riley et al., 2022) and further related work I have been involved with, calculating the power of a planned IPD meta-analysis to examine treatment-covariate interactions with survival outcomes, is currently under review. Software packages incorporating each of these calculations are currently under development and will be made available in due course by the research team.

Further work could consider how these methods could be adapted when implementing other types of statistical models or when using machine learning methods rather than traditional statistical approaches, as machine learning methods have been increasing greatly in popularity in recent years.

7.4 Impact of measurement error

The systematic review in Chapter 3 found that many published prediction models may include predictors that are measured with error. This is also likely to be true for the prognostic factor study presented in Chapter 2 and the external validation of prediction models in Chapter 5. However, the error was unable to be accounted for in these studies and the impact of the error was unknown. There are several things that need to be considered with regard to the impact of measurement error:

- The magnitude of the measurement error
- Whether the predictor subject to measurement error is highly correlated with other error-free predictors
- Whether there are several predictors measured with error
- Whether the measurement error is additive or multiplicative

- Whether the measurement error is biased, dependent or differential
- Whether the measurement error model is classical or Berkson
- Whether the variable is continuous or categorical

The direction and magnitude of bias from measurement error depends heavily on whether the distribution of errors for one predictor depends on the actual value of the predictor, the actual values of other predictors, or the errors in measuring other predictors (Rothman et al., 2008). The inclusion of other predictors measured without error has no effect on the measurement error bias if the additional predictors are not correlated with the original predictor ($\boldsymbol{X}$), but if they are correlated, the bias increases. When there is more than one predictor measured with error, the likelihood of producing incorrect conclusions is greater, and the nature of the bias is not as easily estimated without knowing the magnitude of the measurement errors, the correlation between the underlying precisely measured predictors and the correlation between the measurement errors (Gustafson, 2003).

Non-differential misclassification increases the total bias (Drews and Greeland, 1990, Greenland and Robins, 1985), but despite what is commonly assumed, being non-differential alone does not guarantee bias towards the null (no effect) (Chavance et al., 1992, Kristensen, 1992, Walker and Blettner, 1985). Only non-differential AND independent misclassification creates an expected bias towards the null (Rothman et al., 2008). One example of non-differential error that does not produce a bias towards the null is when the probability of a subjects misclassification on one predictor depends on whether the subject was misclassified on a second predictor, so when the errors are dependent. These dependent errors can create a substantial bias away from the null even if the errors are non-differential for both variables (Kristensen, 1992). Non-differential measurement error can also produce a bias away from the null when the variable has more than two levels. The impact of differential measurement error is harder to predict in advance compared to the impact of non-differential error (Gustafson, 2003) as bias caused by differential misclassification can either exaggerate or underestimate an effect (Rothman et al., 2008).

Care must be taken when defining the measurement error model to be classical or Berkson, as defining it incorrectly often leads to incorrect conclusions, i.e. assuming a Berkson error model when in fact the error is classical leads to a hugely optimistic overstatement of power (Carroll et al., 2006).

Misclassification error differs from measurement error in continuous predictors as the surrogate predictors ($\boldsymbol{W}$) cannot be expressed as a sum of the true predictor ($\boldsymbol{X}$) plus a noise (error) variable ($\boldsymbol{U}$). Misclassification must be characterised in terms of the classification probabilities, i.e. given the true classification, how likely is a correct classification (Gustafson, 2003). Measurement error can often plausibly be assumed to be independent of underlying true values, whereas misclassification error is never independent of the underlying value of the variable. Hence, different theory covers the effects of errors in categorical and continuous variables (White et al., 2001).

In general, a binary misclassification is more damaging than measurement error in continuous predictors (Gustafson, 2003), and if misclassification is severe enough, it can eliminate any real association, or even reverse the direction of an association (Rothman et al., 2008).

Measurement error in predictors that are used to develop a prediction model may not be a problem per se. If it is this same error-prone predictor that will be measured in practice when implementing the prediction model, rather than the true predictor, then there is little issue with using the error-prone predictor to develop the prediction model (Carroll et al., 2006) as it will be the prognostic value of the error-prone predictor that is of interest. However, if the prognostic value of a factor measured without error is of interest, either hypothetically or if an instrument is available that can do this, then the error will need to be accounted for. Consideration needs to be taken of whether measurement error in a predictor is actually a problem for the estimand of interest.

7.5 Machine learning

Traditional statistical modelling techniques have been the focus of this thesis, however in recent years there has been a rapidly growing interest in using machine learning methods to develop clinical prediction models (Kourou et al., 2015). Machine learning can be described as data analytical methods that learn from data without being explicitly programmed (Collins et al., 2021), using models that directly and automatically learn from data (Mitchell, 1997). Whereas regression based models are based on theory and assumptions, and benefit from human intervention and subject knowledge for model specification (Christodoulou et al., 2019).

Due to the growing popularity of machine learning methods, it is inevitable that similar issues to those explored in this thesis will arise in machine learning in the future. In fact, recent reviews have found that reporting of prediction models developed using machine learning methods is poor (Andaur Navarro et al., 2022, Dhiman et al., 2021) and that sample sizes used in many published studies are too small (Shillan et al., 2019). There are currently no recommendations for the minimum sample sizes appropriate for developing prediction models using machine learning methods, however, for there to be clear advantages in the use of machine learning over traditional statistical methods it may be required to have tens or hundreds of thousands of patients (Beam and Kohane, 2018, van der Ploeg et al., 2014). Yet in a recent review of prediction models developed using machine learning methods, the majority of studies analysed data on fewer than 1000 patients (Shillan et al., 2019).

It has also been highlighted that machine learning models do not automatically lead to improved performance over traditional statistical methods (Christodoulou et al., 2019), and that many comparisons of machine learning and statistical methods are poorly reported and unfair.

Reporting guidelines are currently under development for diagnostic and prognostic prediction model studies based on artificial intelligence (TRIPOD-AI)(Collins et al., 2021). These guidelines will be an extension to the TRIPOD statement for prediction model studies developed using machine learning and will hopefully empower researchers developing prediction models using machine learning to report key details which will allow readers to evaluate the study quality and interpret the findings, increasing the chances of the model being implemented in practice.

7.6 Conclusions

Prognosis research is a fundamental aspect of medical research, with the ultimate aim of improving future outcomes in people with a given disease or health condition by informing clinical decision making and helping patients understand their risk. The research in this thesis has contributed towards both applied and methodological aspects of prognosis research. Clinically, the thesis has investigated prognostic factors of adverse outcome in MCDA twin pregnancies, and externally validated existing stillbirth prediction models. Methodologically, the work has contributed towards the improvement of reporting and methodology in prognosis research, by providing recommendations for reporting of prediction models in relation to measurement error that may be present in the predictors, and in relation to the timing of predictor measurements and the intended moment of using the prediction model in practice. Finally, the work in this thesis has provided a method for calculating the power of a prognostic factor study when synthesising IPD from multiple studies, which would allow researchers and funders to decide in advance of collecting the IPD whether the project is worth their investment, potentially saving years of wasted time and money. It is hoped this body of work will have a positive impact on improving the quality of prognosis research in the future.

Appendix A

1. Angioli, R., et al., *A Predictive Score for Secondary Cytoreductive Surgery in Recurrent Ovarian Cancer (SeC-Score): A Single-Centre, Controlled Study for Preoperative Patient Selection.* Annals Of Surgical Oncology, 2015. **22**(13): p. 4217-4223.

2. Ankerst, D.P., et al., *Precision Medicine in Active Surveillance for Prostate Cancer: Development of the Canary-Early Detection Research Network Active Surveillance Biopsy Risk Calculator.* European Urology, 2015. **68**(6): p. 1083-1088.

3. Bendifallah, S., et al., *A Predictive Model Using Histopathologic Characteristics of Early-Stage Type 1 Endometrial Cancer to Identify Patients at High Risk for Lymph Node Metastasis.* Annals Of Surgical Oncology, 2015. **22**(13): p. 4224-4232.

4. Black, S.R., et al., *Toward a More Robust Prediction of Pulmonary Embolism in Trauma Patients. A Risk Assessment Model Based Upon 38,000 Patients.* Journal Of Orthopaedic Trauma, 2015.

5. Chang, C., et al., *A joint model based on longitudinal CA125 in ovarian cancer to predict recurrence.* Biomarkers In Medicine, 2015.

6. Chen, J.-Y., et al., *Predicting Non-sentinel Lymph Node Metastasis in a Chinese Breast Cancer Population with 1-2 Positive Sentinel Nodes: Development and Assessment of a New Predictive Nomogram.* World Journal Of Surgery, 2015. **39**(12): p. 2919-2927.

7. Cohen, M.H., et al., *Gender-Related Risk Factors Improve Mortality Predictive Ability of VACS Index Among HIV-Infected Women.* Journal Of Acquired Immune Deficiency Syndromes (1999), 2015. **70**(5): p. 538-544.

8. Corey, K.E., et al., *Development and Validation of an Algorithm to Identify Nonalcoholic Fatty Liver Disease in the Electronic Medical Record.* Digestive Diseases And Sciences, 2015.

9. Coté, G.A., et al., *Development and Validation of a Prediction Model for Admission After Endoscopic Retrograde Cholangiopancreatography.* Clinical Gastroenterology And Hepatology: The Official Clinical Practice Journal Of The American Gastroenterological Association, 2015. **13**(13): p. 2323-2332.e9.

10. Di Filippo, F., et al., *Elaboration of a nomogram to predict non sentinel node status in breast cancer patients with positive sentinel node, intra-operatively assessed with one step nucleic acid amplification method.* Journal Of Experimental & Clinical Cancer Research: CR, 2015. **34**(1): p. 136-136.

11. Du, X.-J., et al., *Neoadjuvant chemotherapy in locally advanced nasopharyngeal carcinoma: Defining high-risk patients who may benefit before concurrent chemotherapy combined with intensity-modulated radiotherapy.* Scientific Reports, 2015. **5**: p. 16664-16664.

12. Dua, A., et al., *Development of a scoring system to estimate mortality in abdominal aortic aneurysms management.* Vascular, 2015. **23**(6): p. 586-591.

13. Englum, B.R., et al., *A Bedside Risk Calculator to Preoperatively Distinguish Follicular Thyroid Carcinoma from Follicular Variant of Papillary Thyroid Carcinoma.* World Journal Of Surgery, 2015. **39**(12): p. 2928-2934.

14. Faget, C., et al., *Value of CT to predict surgically important bowel and/or mesenteric injury in blunt trauma: performance of a preliminary scoring system.* European Radiology, 2015. **25**(12): p. 3620-3628.

15. Horn, S.D., et al., *A Predictive Model for Pressure Ulcer Outcome: The Wound Healing Index.* Advances In Skin & Wound Care, 2015. **28**(12): p. 560-572.

16. Kaymakcalan, M.D., et al., *Risk factors and model for predicting toxicity-related treatment discontinuation in patients with metastatic renal cell carcinoma treated with vascular endothelial growth factor-targeted therapy: Results from the International Metastatic Renal Cell Carcinoma Database Consortium.* Cancer, 2015.

17. Koller, L., et al., *History of previous bleeding and C-reactive protein improve assessment of bleeding risk in elderly patients (≥ 80 years) with myocardial infarction.* Thrombosis And Haemostasis, 2015. **114**(5): p. 1085-1091.

18. Koning, N.R., et al., *Identification of patients at risk for colorectal cancer in primary care: an explorative study with routine healthcare data.* European Journal Of Gastroenterology & Hepatology, 2015. **27**(12): p. 1443-1448.

19. Kusamura, S., et al., *The Role of Ki-67 and Pre-cytoreduction Parameters in Selecting Diffuse Malignant Peritoneal Mesothelioma (DMPM) Patients for Cytoreductive Surgery (CRS) and Hyperthermic Intraperitoneal Chemotherapy (HIPEC).* Annals Of Surgical Oncology, 2015.

20. Lei, M., et al., *Prediction of survival prognosis after surgery in patients with symptomatic metastatic spinal cord compression from non-small cell lung cancer.* BMC Cancer, 2015. **15**(1): p. 853-853.

21. Matsuo, K., et al., *Predictive Factor of Conversion to Laparotomy in Minimally Invasive Surgical Staging for Endometrial Cancer.* International Journal Of Gynecological Cancer: Official Journal Of The International Gynecological Cancer Society, 2015.

22. Nykanen, D.G., et al., *CRISP: Catheterization RISk score for pediatrics: A Report from the Congenital Cardiac Interventional Study Consortium (CCISC).* Catheterization And Cardiovascular Interventions: Official Journal Of The Society For Cardiac Angiography & Interventions, 2015.

23. Olmedilla, L., et al., *Early Measurement of Indocyanine Green Clearance Accurately Predicts Short-Term Outcomes After Liver Transplantation.* Transplantation, 2015.

24. Resch, J.E., et al., *A Preliminary Formula to Predict Timing of Symptom Resolution for Collegiate Athletes Diagnosed With Sport Concussion.* Journal Of Athletic Training, 2015.

25. Rosenkrantz, A.B., et al., *Prostate Cancer: Utility of Whole-Lesion Apparent Diffusion Coefficient Metrics for Prediction of Biochemical Recurrence After Radical Prostatectomy.* AJR. American Journal Of Roentgenology, 2015. **205**(6): p. 1208-1214.

26. Russo, G.I., et al., *Performance of biopsy factors in predicting unfavorable disease in patients eligible for active surveillance according to the PRIAS criteria.* Prostate Cancer And Prostatic Diseases, 2015. **18**(4): p. 338-342.

27. Shaikh, A.Y., et al., *Addition of B-Type Natriuretic Peptide to Existing Clinical Risk Scores Enhances Identification of Patients at Risk for Atrial Fibrillation Recurrence After Pulmonary Vein Isolation.* Critical Pathways In Cardiology, 2015. **14**(4): p. 157-165.

28. Siegel, C.A., et al., *A validated web-based tool to display individualised Crohn's disease predicted outcomes based on clinical, serologic and genetic variables.* Alimentary Pharmacology & Therapeutics, 2015.

29. Spolverato, G., et al., *Can hepatic resection provide a long-term cure for patients with intrahepatic cholangiocarcinoma?* Cancer, 2015. **121**(22): p. 3998-4006.

30. Suh, Y.J., et al., *Prognostic value of SYNTAX score based on coronary computed tomography angiography.* International Journal Of Cardiology, 2015. **199**: p. 460-466.

31. Tada, H., et al., *Predictive score for early diagnosis of acute encephalopathy with biphasic seizures and late reduced diffusion (AESD).* Journal Of The Neurological Sciences, 2015. **358**(1-2): p. 62-65.

32. Takahashi, N., et al., *Small (< 4 cm) Renal Masses: Differentiation of Angiomyolipoma Without Visible Fat From Renal Cell Carcinoma Using Unenhanced and Contrast-Enhanced CT.* AJR. American Journal Of Roentgenology, 2015. **205**(6): p. 1194-1202.

33. Zhou, J., et al., *Multivariate logistic regression analysis of postoperative complications and risk model establishment of gastrectomy for gastric cancer: A single-center cohort report.* Scandinavian Journal Of Gastroenterology, 2016. **51**(1): p. 8-15.

Appendix B

Appendix Table B.1: Predictors in final models at low risk of important measurement error

Bowel wall discontinuity	% exophytic growth	Extrathyroidal extension	B-type natriuretic peptide (BNP)
Diabetes	Gender	N classification	First wound area
Age	Disease location	SYNTAX groups	Use of anxiolytic
PRD1	Ablation time	Liver cirrhosis	Splenic injury
Inotropic support	Mesenteric stranding	Persistent atrial fibrillation (AF)	Consciousness level after seizures
Uterine size	Entropy	BR2CHADS2	Procedure type
Race	Injury location	Platelet count	Haemoperitoneum
Paralysis	Child-Pugh score	Gleason grade	Endovascular repair
Combined resection	Extrauterine disease	Mechanical ventilation	Number of involved vertebra
Bowel wall thickness	Bile duct obstruction	Blood glucose on admission	Size of SLN metastases
Nodal metastases	Mobility of patients at arrival	Visceral metastases	Invasive neighbouring organs
Left atrial diameter	Stent replacement	No. of major comorbidities	Mesenteric pneumoperitoneum
Single metastatic site	Mean of the bottom 10th percentile ADC	Lesion to kidney CT attenuation difference	Billroth II anastomosis of reconstruction
Peripheral arterial disease	Anterior abdominal wall injury	Injury from motorcycle accident	History of at least one prior negative biopsy
Arrival by helicopter	Renal transplant or dialysis	Resides in nursing home	Patient age at first treatment
Number of positive SLNs	Extraprostatic extension	Surgical volume of surgeons	% Positive core for cancer
Periductal invasion	NOD2 frameshift mutation	Number of CK19 mRNA copies	Cumulative length of positive cores
Aspartate aminotransferase (AST) on admission	Positive lymphovascular invasion (LVSI) status	Number of negative sentinel lymph nodes (SLN)	Multifocal intrahepatic cholangiocellular carcinoma (ICC)
Admission to intensive care unit	Number of months since last biopsy	Number of fatty liver mentions extracted from notes over a lifetime	Coronary computed tomography angiography (CCTA) 1/2/3 vessel disease or left main disease

Short axis diameter	Lesion attenuation on contrast-enhanced Computed Tomography (CT) scan	Triglyceride level within 12 months of NAFLD radiographic	Endoscopic retrograde cholangiopancreatography (ERCP) manometry
Physiologic category	Distant metastases	Hypercoagulability	Sodium level
% involvement of positive cores	Vascular invasion	Renal failure	Histologic subtype
Insulin dependent diabetes	Maximum cancer length	Arterial mesenteric vessel extravasion	Veterans aging cohort study index
Entropy apparent diffusion coefficient (ADC)	Plasma Epstein-Barr virus (EBV) DNA	Intraoperative transfusion	Para-aortic lymphadenectomy
Reduced bowel wall enhancement	Number of previous or concurrent other wounds or ulcers	Admitted for acute hospital stay or emergency department visit report	Lifetime number of non-alcoholic fatty liver disease (NAFLD) icd9 codes

Appendix Table B.2: Predictors in final models at high risk of important measurement error

Key reasons for being at high risk of error	**Predictors included in final models**
Fluctuations in human samples/ biological variability	Serum albumin, Serologic markers, Prostate Specific Antigen (PSA) density, Prostate Specific Antigen (PSA), Ki-67, Human epididymis protein 4 (HE4), Glomerular filtration rate, Emergency room pulse rate, CRUSADE score, C-reactive protein, Creatinine on admission, CA125, Ascites
Inaccuracy of measurement instruments	Body Mass Index (BMI), Myometrial invasion depth, Emergency room pulse rate, Creatinine on admission, Weight, Ascites, International normalised ratio (INR1) , Infection/bioburden
Imperfect recall	Body Mass Index (BMI), Duration of convulsions, Duration of drowsiness, Duration of neck pain, Duration of nervousness, Duration of tingling, History of transactional sex, Area under pain curve, Congestive heart failure, Weight, Previous bleeding, Endoscopic retrograde cholangiopancreatography (ERCP) time, Time developing motor deficits, ImPACT total symptom score, Eastern Cooperative Oncology Group (ECOG) performance status, Depression, Number of non-major comorbidities, Systemic illness/organ failure
Subjective nature of measures	Abdominal pain, Tumour stage, Suboptimal pelvic examination or enlarged uterus during preoperative evaluation, Area under pain curve, hypertension, Clinical stage, Malnutrition, Obesity, Procedure risk category, Pressure ulcer stage, ImPACT total symptom score, Eastern Cooperative Oncology Group (ECOG) performance status, Depression, Pre-catheterisation diagnosis
Laboratory or measurer error	Tumour stage, Suboptimal pelvic examination or enlarged uterus during preoperative evaluation, Myometrial invasion depth, CRUSADE score, CA125, Histologic grade, Primary tumour diameter, Clinical stage, Residual tumour, Endoscopic Retrograde Cholangiopancreatography (ERCP) Time, Tumour size, Pressure ulcer stage, Ascites, International normalised ratio (INR1) , Peritoneal Cancer Index, Infection/bioburden, Operating time and age, Wound (ulcer) age at first encounter

Appendix C

Example code for calculating the power of a planned IPD meta-analysis to evaluate prognostic factor factors prior to collecting the IPD.

Appendix C1: Binary prognostic factor example

```
// Run all the code in one go


// START
     clear all


// define the model as logit-p = alpha + gamma*z with z the
prognostic factor of interest


// Generate matrices M1 and M2 from eq6.9
     mat M1 = (1,0 \ 0,0)
     mat M2 = (1,1 \ 1,1)


// Poynard application - 4 trials, sex covariate
     use "Poynard data.dta", replace


// calculate total sample size of each trial
     gen events = events_C + events_T
     gen total = n_C + n_T
     mkmat total, matrix(n_trans)
     mat n = n_trans'
```

```
// specify gamma = assumed prognostic effect of Z - assumed
common for each trial here
// need to decide what this is going to be based on clinical
evidence or use a range of values
     local exp_gamma = 1.5
     local gamma = ln(exp_gamma)
     mat gamma =  (`gamma', `gamma' , `gamma' , `gamma')


// calculate total percentage male
     gen male_C=round(n_C*percent_male_C/100)
     gen male_T=round(n_T*percent_male_T/100)
     gen male =  male_C + male_T
     gen percent_male = (male / total) * 100


// define number of patients by Z categories (proportion of
z0 AND z1)
     gen n_z0 = total-male
     gen n_z1 = male


     gen prop_z0 = (100-percent_male)/100
     gen prop_z1 = percent_male/100
     mkmat prop_z0, matrix(prob_Z_0_trans)
     mkmat prop_z1, matrix(prob_Z_1_trans)
     mat prob_Z_0 = prob_Z_0_trans'
     mat prob_Z_1 = prob_Z_1_trans'


// create matrices to calculate alpha
     gen logit_risk=logit(events/total)
     mkmat logit_risk, matrix(risk_trans)
     mat l_risk = risk_trans'
     gen n1_n=n_z1/total
     mat n_gamma = (n1_n[1] * `gamma', n1_n[2] * `gamma' ,
     n1_n[3] * `gamma' , n1_n[4] * `gamma')
```

```
// define alpha vector
     mat alpha = l_risk-n_gamma


// count total number of studies
     local studies = _N


// calculate the Variance Inflation Factor (VIF) - to adjust
for other covariates
// set rho=0 if not adjusting for other covariates
     local rho = 0.5
     local VIF = 1/(1-`rho'^2)


// calculate the variance for each study using eq6.9 followed
by eq6.6 (Step 1)
     tempname pow
     tempfile powerdata
     postfile `pow'  study gamma variance se
     using `powerdata', replace


     forvalues i = 1/`studies' {
     mat eqM1_`i' =
     exp(alpha[1,`i'])/((1+exp(alpha[1,`i']))^2)*M1*prob_Z_0[
     1,`i']
     mat eqM2_`i' = exp(alpha[1,`i']+
     gamma[1,`i'])/((1+exp(alpha[1,`i'] +
     gamma[1,`i']))^2)*M2*prob_Z_1[1,`i']
     mat I_`i' = eqM1_`i' + eqM2_`i'
     mat inverseI_`i' = inv(I_`i')
     mat var_`i' = inverseI_`i'/n[1,`i']


     local variance = var_`i'[2,2] * `VIF'


     local se = sqrt(`variance')
     local prog_factor = gamma[1, `i']
```

```
	local study = `i'
	post `pow'  (`study')  (`prog_factor')   (`variance')
	(`se')
	}
	postclose `pow'


// calculate variance of the prognostic factor from the
planned IPD meta-analysis (Step 2) using eq6.15
	use  `powerdata', replace
	gen inv_var = 1/variance
	qui summ inv_var
	local summ_inv_var = r(sum)
	local ma_variance = 1/r(sum)
	local ma_se = sqrt(1/r(sum))


// calculate the power of the planned IPD meta-analysis using
eq6.16 (step 3)
	disp "prognostic factor = " gamma[1,1] _newline
	"ma_variance = " `ma_variance' _newline "ma_se = "
	`ma_se' _newline  "lower = " gamma[1,1] - (1.96*
	sqrt(`ma_variance')) _newline  "upper = " gamma[1,1] +
	(1.96* sqrt(`ma_variance')) _newline  "power = "
	normal(-1.96 + (gamma[1,1] * sqrt(`summ_inv_var')))  +
	normal(-1.96 - (gamma[1,1] * sqrt(`summ_inv_var')))
	_newline "power (%) = " 100 * (normal(-1.96 +
	(gamma[1,1] * sqrt(`summ_inv_var')))  + normal(-1.96 -
	(gamma[1,1] * sqrt(`summ_inv_var'))))



// STOP
```

Appendix C2: Continuous prognostic factor example

```
// Run all the code in one go


// START
     clear all


// Poynard application - 4 trials, age covariate
     use "Poynard data.dta", replace


     tempname pow
     tempfile powerdata
     postfile `pow'  study gamma variance se using
     `powerdata', replace


// first define the logistic equation parameters for each
study


// total sample size of each trial
     gen events = events_C + events_T
     gen total = n_C + n_T
     mkmat total, matrix(n_trans)
     mat n = n_trans'


// define vector alpha = logit(overall risk) - for average Z
     gen logit_prob = logit(events/total)
     mkmat logit_prob, matrix(alpha_trans)
     mat alpha = alpha_trans'
```

```
// specify gamma = assumed prognostic effect of Z - assumed
common for each trial here

     local exp_gamma = 1.04

     local gamma = ln(`exp_gamma')

     mat gamma =  (`gamma', `gamma', `gamma', `gamma')


// calculate the overall mean age

     gen age_mean=((n_C*age_mean_C)+(n_T*age_mean_T)) / total

     gen age_sd = sqrt(((((n_C - 1) * (age_sd_C)^2) + ((n_T -
     1) * (age_sd_T)^2))/(n_C + n_T - 1)) + ((n_C * n_T *
     (age_mean_C - age_mean_T)^2) / ((n_C + n_T) * (n_C + n_T
     - 1))))


     mkmat age_mean , matrix(age_mean_trans)

     mat age_mean = age_mean_trans'

     mkmat age_sd , matrix(age_sd_trans)

     mat age_sd = age_sd_trans'


// count total number of studies

     local studies = _N


// generate large dataset for the simulation

     set obs 1000000

     local obs = 1000000

     set seed 1234

     gen id = _n


// calculate the Variance Inflation Factor (VIF) to adjust
for other covariates

// set rho=0 if not adjusting for other covariates

     local rho = 0.5

     local VIF = 1/(1-`rho'^2)
```

```
// generate the covariate Z of interest; assuming normal
     forvalues i = 1/`studies' {
     gen z`i' = .

     * using the observed mean and sd
     replace z`i' = rnormal(age_mean[1,`i'], age_sd[1, `i'])
     }

// create centered z variable so that intercept is similar to
overall effect
     forvalues i = 1/`studies' {
          qui summ z`i'
     gen z`i'_cent = z`i' - r(mean)
     }

// now generate the 2 by 2 matrix entries corresponding to
logit-p = alpha + gamma*z
// for each study separately
     forvalues i = 1/`studies' {

     gen LP`i' = alpha[1,`i'] + (gamma[1,`i'] * z`i'_cent)

     gen M_11`i' = exp(LP`i')/((1+ exp(LP`i'))^2)
     gen M_12`i' = z`i'_cent*exp(LP`i')/((1+ exp(LP`i'))^2)
     gen M_21`i' = M_12`i'
     gen M_22`i' = z`i'_cent*z`i'_cent*exp(LP`i')/((1+
     exp(LP`i'))^2)
     }

// calculate expected values of the cells and form I for each
study
     forvalues i = 1/`studies' {
     qui summ M_11`i'
     local I_11`i' = r(mean)
```

```
      qui summ M_12`i'
      local I_12`i' = r(mean)
      qui summ M_21`i'
      local I_21`i' = r(mean)
      qui summ M_22`i'
      local I_22`i' = r(mean)
       mat I`i' = (`I_11`i'', `I_12`i'' \ `I_21`i'',
      `I_22`i'')
      mat invI`i' = inv(I`i')

// calculate the variance of each study using eq6.6 (Step 1)
      mat var_`i' = invI`i'/n[1,`i']
      local variance = var_`i'[2,2]*`VIF'


      local se = sqrt(`variance')
      local prog_factor = gamma[1, `i']
      local study = `i'
      post `pow'  (`study')  (`prog_factor')   (`variance')
      (`se')
      }
      postclose `pow'
      use  `powerdata', replace

// calculate the variance of the prognostic factor from the
planned IPD meta-analysis using eq6.15 (Step 2)
      gen inv_var = 1/variance
      qui summ inv_var
      local summ_inv_var = r(sum)
      local ma_variance = 1/r(sum)
      local ma_se = sqrt(1/r(sum))
```

```
// calculate the power of the planned IPD meta-analysis using
eq6.16 (Step 3)

     disp "prognostic factor = " gamma[1,1] _newline
     "ma_variance = " `ma_variance' _newline "ma_se = "
     `ma_se' _newline  "lower = " gamma[1,1] - (1.96*
     sqrt(`ma_variance')) _newline  "upper = " gamma[1,1] +
     (1.96* sqrt(`ma_variance')) _newline  "power = "
     normal(-1.96 + (gamma[1,1] * sqrt(`summ_inv_var')))  +
     normal(-1.96 - (gamma[1,1] * sqrt(`summ_inv_var')))
     _newline "power (%) = " 100 * (normal(-1.96 +
     (gamma[1,1] * sqrt(`summ_inv_var')))  + normal(-1.96 -
     (gamma[1,1] * sqrt(`summ_inv_var'))))



// STOP
```

Appendix C3: Allowing for heterogeneity

```
// run as above until generating inv_var


// specify an assumed value of tau (the between-study
standard deviation of the prognostic factor effect)

     local tau = 0.0075


// calculate the variance of the prognostic factor using
eq6.5

     gen inv_var = 1/(variance+(`tau'^2))

     qui summ inv_var

     local summ_inv_var = r(sum)

     local ma_variance = 1/r(sum)

     local ma_se = sqrt(1/r(sum))


// calculate the power using eq6.17

     disp "prognostic factor = " gamma[1,1] _newline
     "ma_variance = " `ma_variance' _newline "ma_se = "
     `ma_se' _newline  "lower = " gamma[1,1] - (1.96*
     sqrt(`ma_variance')) _newline  "upper = " gamma[1,1] +
     (1.96* sqrt(`ma_variance')) _newline "power = " t(_N-1 ,
     -invt(_N-1, 0.975) + (gamma[1,1] *
     sqrt(`summ_inv_var')))  + t(_N-1 , -invt(_N-1, 0.975) -
     (gamma[1,1] * sqrt(`summ_inv_var'))) _newline "power (%)
     = " 100 * (t(_N-1, -invt(_N-1, 0.975) + (gamma[1,1] *
     sqrt(`summ_inv_var')))  + t(_N-1 , -invt(_N-1, 0.975) -
     (gamma[1,1] * sqrt(`summ_inv_var'))))
```

Reference List

ABDULAZIZ, K. E., PERRY, J. J., YADAV, K., DOWLATSHAHI, D., STIELL, I. G., WELLS, G. A. & TALJAARD, M. 2022. Quality and transparency of reporting derivation and validation prognostic studies of recurrent stroke in patients with TIA and minor stroke: a systematic review. *Diagnostic and Prognostic Research,* 6**,** 9.

ABO-ZAID, G., GUO, B., DEEKS, J. J., DEBRAY, T. P., STEYERBERG, E. W., MOONS, K. G. & RILEY, R. D. 2013. Individual participant data meta-analyses should not ignore clustering. *J Clin Epidemiol,* 66**,** 865-873.e4.

ACOG 2016. Practice Bulletin No. 169: Multifetal Gestations: Twin, Triplet, and Higher-Order Multifetal Pregnancies. *Obstet Gynecol,* 128**,** e131-46.

AHMED, I., SUTTON, A. J. & RILEY, R. D. 2012. Assessment of publication bias, selection bias, and unavailable data in meta-analyses using individual participant data: a database survey. *BMJ,* 344**,** d7762.

AKOLEKAR, R., BOWER, S., FLACK, N., BILARDO, C. M. & NICOLAIDES, K. H. 2011. Prediction of miscarriage and stillbirth at 11-13 weeks and the contribution of chorionic villus sampling. *Prenat Diagn,* 31**,** 38-45.

AKOLEKAR, R., MACHUCA, M., MENDES, M., PASCHOS, V. & NICOLAIDES, K. H. 2016a. Prediction of stillbirth from placental growth factor at 11–13 weeks. *Ultrasound in Obstetrics & Gynecology,* 48**,** 618-623.

AKOLEKAR, R., TOKUNAKA, M., ORTEGA, N., SYNGELAKI, A. & NICOLAIDES, K. H. 2016b. Prediction of stillbirth from maternal factors, fetal biometry and uterine artery Doppler at 19-24 weeks. *Ultrasound Obstet Gynecol,* 48**,** 624-630.

ALI, A., BLACK, D. & SOSLOW, R. A. 2007. Difficulties in assessing the depth of myometrial invasion in endometrial carcinoma. *Int J Gynecol Pathol,* 26**,** 115-23.

ALLOTEY, J., WHITTLE, R., SNELL, K. I. E., SMUK, M., TOWNSEND, R., VON DADELSZEN, P., HEAZELL, A. E. P., MAGEE, L., SMITH, G. C. S., SANDALL, J., THILAGANATHAN, B., ZAMORA, J., RILEY, R. D., KHALIL, A., THANGARATINAM, S. & NETWORK, T. I. C. 2022. External validation of prognostic models to predict stillbirth using International Prediction of Pregnancy Complications (IPPIC) Network database: individual participant data meta-analysis. *Ultrasound in Obstetrics & Gynecology,* 59**,** 209-219.

ALTMAN, D. G., MCSHANE, L. M., SAUERBREI, W. & TAUBE, S. E. 2012. Reporting Recommendations for Tumor Marker Prognostic Studies (REMARK): explanation and elaboration. *PLoS Med,* 9**,** e1001216.

ALTMAN, D. G. & ROYSTON, P. 2000. What do we mean by validating a prognostic model? *Stat Med,* 19**,** 453-73.

ALTMAN, D. G. & ROYSTON, P. 2006. The cost of dichotomising continuous variables. *BMJ,* 332**,** 1080.

ALTMAN, D. G., TRIVELLA, M., PEZZELLA, F., HARRIS, A. L. & PASTORINO, U. 2007. Systematic Review of Multiple Studies of Prognosis: The Feasibility of Obtaining Individual Patient Data. *In:* AUGET, J.-L., BALAKRISHNAN, N., MESBAH, M. & MOLENBERGHS, G. (eds.) *Advances in Statistical Methods for the Health Sciences: Applications to Cancer and AIDS Studies, Genome Sequence Analysis, and Survival Analysis.* Boston, MA: Birkhäuser Boston.

ALTMAN, D. G., VERGOUWE, Y., ROYSTON, P. & MOONS, K. G. 2009. Prognosis and prognostic research: validating a prognostic model. *Bmj,* 338**,** b605.

ÅMARK, H., WESTGREN, M. & PERSSON, M. 2018. Prediction of stillbirth in women with overweight or obesity—A register-based cohort study. *PLOS ONE,* 13**,** e0206940.

ANDAUR NAVARRO, C. L., DAMEN, J. A. A., TAKADA, T., NIJMAN, S. W. J., DHIMAN, P., MA, J., COLLINS, G. S., BAJPAI, R., RILEY, R. D., MOONS, K. G. M. & HOOFT, L. 2022. Completeness of reporting of clinical prediction models developed using supervised machine learning: a systematic review. *BMC Medical Research Methodology,* 22**,** 12.

ARMSTRONG, B. G. 1998. Effect of measurement error on epidemiological studies of environmental and occupational exposures. *Occup Environ Med,* 55**,** 651-6.

ARTUS, M., VAN DER WINDT, D., JORDAN, K. P. & CROFT, P. R. 2014. The clinical course of low back pain: a meta-analysis comparing outcomes in randomised clinical trials (RCTs) and observational studies. *BMC Musculoskelet Disord,* 15**,** 68.

AUPONT, J. E., AKOLEKAR, R., ILLIAN, A., NEONAKIS, S. & NICOLAIDES, K. H. 2016. Prediction of stillbirth from placental growth factor at 19–24 weeks. *Ultrasound in Obstetrics & Gynecology,* 48**,** 631-635.

AUSTIN, P. C. & STEYERBERG, E. W. 2015. The number of subjects per variable required in linear regression analyses. *J Clin Epidemiol,* 68**,** 627-36.

AUSTIN, P. C. & STEYERBERG, E. W. 2017. Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. *Stat Methods Med Res,* 26**,** 796-808.

BEAM, A. L. & KOHANE, I. S. 2018. Big Data and Machine Learning in Health Care. *Jama,* 319**,** 1317-1318.

BHATNAGAR, P., WICKRAMASINGHE, K., WILLIAMS, J., RAYNER, M. & TOWNSEND, N. 2015. The epidemiology of cardiovascular disease in the UK 2014. *Heart*.

BOWLING, A. 2005. Mode of questionnaire administration can have serious effects on data quality. *J Public Health (Oxf),* 27**,** 281-91.

BRAGA, F., FERRARO, S., MOZZI, R. & PANTEGHINI, M. 2014. The importance of individual biology in the clinical use of serum biomarkers for ovarian cancer. *Clinical Chemistry and Laboratory Medicine (CCLM).*

BRAGA, F. & PANTEGHINI, M. 2016. Generation of data on within-subject biological variation in laboratory medicine: An update. *Critical Reviews in Clinical Laboratory Sciences,* 53**,** 313-325.

BURKE, D. L., ENSOR, J. & RILEY, R. D. 2017. Meta-analysis using individual participant data: one-stage and two-stage approaches, and why they may differ. *Stat Med,* 36**,** 855-875.

CANTARUTTI, A., BATEMAN, B. T., HERNANDEZ-DIAZ, S., GRAY, K. J., PATORNO, E., CORRAO, G., DESAI, R. J. & HUYBRECHTS, K. 2018. Algorithms to estimate the timing of pregnancy for stillbirths in pregnancy safety studies. *34th International Conference on Pharmacoepidemiology and Therapeutic Risk Management.* Czech Republic: Pharmacoepidemiology and Drug Safety.

CARROLL, R. J., RUPPERT, D., STEFANSKI, L. A. & CRAINICEANU, C. M. 2006. *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition*, CRC Press.

CHAVANCE, M., DELLATOLAS, G. & LELLOUCH, J. 1992. Correlated nondifferential misclassifications of disease and exposure: application to a cross-sectional study of the relation between handedness and immune disorders. *Int J Epidemiol,* 21**,** 537-46.

CHRISTODOULOU, E., MA, J., COLLINS, G. S., STEYERBERG, E. W., VERBAKEL, J. Y. & VAN CALSTER, B. 2019. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology,* 110**,** 12-22.

CHU, H. & COLE, S. R. 2006. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. *J Clin Epidemiol,* 59**,** 1331-2; author reply 1332-3.

COHEN, J. 1983. The Cost of Dichotomization. *Applied Psychological Measurement,* 7**,** 249-253.

COLLINS, G. S., DHIMAN, P., ANDAUR NAVARRO, C. L., MA, J., HOOFT, L., REITSMA, J. B., LOGULLO, P., BEAM, A. L., PENG, L., VAN CALSTER, B., VAN SMEDEN, M., RILEY, R. D. & MOONS, K. G. 2021. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool

(PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open,* 11**,** e048008.

COLLINS, G. S., REITSMA, J. B., ALTMAN, D. G. & MOONS, K. G. 2015. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Bmj,* 350**,** g7594.

CONDE-AGUDELO, A., BIRD, S., KENNEDY, S., VILLAR, J. & PAPAGEORGHIOU, A. 2015. First- and second-trimester tests to predict stillbirth in unselected pregnant women: a systematic review and meta-analysis. *BJOG: An International Journal of Obstetrics & Gynaecology,* 122**,** 41-55.

COSTA, L. C. M., MAHER, C. G., HANCOCK, M. J., MCAULEY, J. H., HERBERT, R. D. & COSTA, L. O. 2012. The prognosis of acute and persistent low-back pain: a meta-analysis. *Cmaj,* 184**,** E613-24.

COSTE, J., DELECOEUILLERIE, G., COHEN DE LARA, A., LE PARC, J. M. & PAOLAGGI, J. B. 1994. Clinical course and prognostic factors in acute low back pain: an inception cohort study in primary care practice. *Bmj,* 308**,** 577-80.

CROWDER, M., DIXON, M., LEDFORD, A. & ROBINSON, M. 2002. Dynamic Modelling and Prediction of English Football League Matches for Betting. *Journal of the Royal Statistical Society. Series D (The Statistician),* 51**,** 157-168.

CROWSON, C. S., ATKINSON, E. J. & THERNEAU, T. M. 2013. Assessing calibration of prognostic risk scores. *Statistical Methods in Medical Research*.

DAOUST, R., SIROIS, M. J., LEE, J. S., PERRY, J. J., GRIFFITH, L. E., WORSTER, A., LANG, E., PAQUET, J., CHAUNY, J. M. & EMOND, M. 2017. Painful Memories: Reliability of Pain Intensity Recall at 3 Months in Senior Patients. *Pain Res Manag,* 2017**,** 5983721.

DAVIS, C. E. 1976. The effect of regression to the mean in epidemiologic and clinical studies. *Am J Epidemiol,* 104**,** 493-8.

DEBRAY, T. P., VERGOUWE, Y., KOFFIJBERG, H., NIEBOER, D., STEYERBERG, E. W. & MOONS, K. G. 2015. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol,* 68**,** 279-89.

DEBRAY, T. P. A., DAMEN, J. A. A. G., SNELL, K. I. E., ENSOR, J., HOOFT, L., REITSMA, J. B., RILEY, R. D. & MOONS, K. G. M. 2017. A guide to systematic review and meta-analysis of prediction model performance. *BMJ,* 356**,** i6460.

DELANAYE, P., SCHAEFFNER, E., EBERT, N., CAVALIER, E., MARIAT, C., KRZESINSKI, J.-M. & MORANNE, O. 2012. Normal reference values for glomerular filtration rate: what do we really know? *Nephrology Dialysis Transplantation,* 27**,** 2664-2672.

DEMIDENKO, E. 2008. Sample size and optimal design for logistic regression with binary interaction. *Stat Med,* 27**,** 36-46.

DHIMAN, P., MA, J., NAVARRO, C. A., SPEICH, B., BULLOCK, G., DAMEN, J. A., KIRTLEY, S., HOOFT, L., RILEY, R. D., VAN CALSTER, B., MOONS, K. G. M. & COLLINS, G. S. 2021. Reporting of prognostic clinical prediction models based on machine learning methods in oncology needs to be improved. *J Clin Epidemiol,* 138**,** 60-72.

DI, J., LI, X., YANG, J., LI, L. & YU, X. 2022. Bias and Reporting Quality of Clinical Prognostic Models for Idiopathic Pulmonary Fibrosis: A Cross-Sectional Study. *Risk Manag Healthc Policy,* 15**,** 1189-1201.

DIEHM, C., DARIUS, H., PITTROW, D., SCHWERTFEGER, M., TEPOHL, G., HABERL, R. L., ALLENBERG, J. R., BURGHAUS, I. & TRAMPISCH, H. J. 2011. Prognostic value of a low post-exercise ankle brachial index as assessed by primary care physicians. *Atherosclerosis,* 214**,** 364-72.

DONDERS, A. R., VAN DER HEIJDEN, G. J., STIJNEN, T. & MOONS, K. G. 2006. Review: a gentle introduction to imputation of missing values. *J Clin Epidemiol,* 59**,** 1087-91.

DREWS, C. D. & GREELAND, S. 1990. The impact of differential recall on the results of case-control studies. *Int J Epidemiol,* 19**,** 1107-12.

DURRLEMAN, S. & SIMON, R. 1989. Flexible regression models with cubic splines. *Statistics in Medicine,* 8**,** 551-561.

EFRON, B. 1983. Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American statistical association,* 78**,** 316-331.

EFRON, B. & TIBSHIRANI, R. J. 1994. *An introduction to the bootstrap*, CRC press.

ENDERS, C. K. 2008. A Note on the Use of Missing Auxiliary Variables in Full Information Maximum Likelihood-Based Structural Equation Models. *Structural Equation Modeling: A Multidisciplinary Journal,* 15**,** 434-448.

ENSOR, J., BURKE, D. L., SNELL, K. I. E., HEMMING, K. & RILEY, R. D. 2018. Simulation-based power calculations for planning a two-stage individual participant data meta-analysis. *BMC Medical Research Methodology,* 18**,** 41.

ENSOR, J., SNELL, K. I. E. & MARTIN, E. C. 2020. PMCALPLOT: Stata module to produce calibration plot of prediction model performance.

FAMILIARI, A., SCALA, C., MORLANDO, M., BHIDE, A., KHALIL, A. & THILAGANATHAN, B. 2016. Mid-pregnancy fetal growth, uteroplacental Doppler indices and maternal demographic characteristics: role in prediction of stillbirth. *Acta Obstet Gynecol Scand,* 95**,** 1313-1318.

FASP 2015. Fetal anomaly screening programme. *Crown*.

FISHER, D. J. 2015. Two-stage individual participant data meta-analysis and generalized forest plots. *Stata Journal,* 15**,** 369-396.

FISHER, D. J., COPAS, A. J., TIERNEY, J. F. & PARMAR, M. K. 2011. A critical review of methods for the assessment of patient-level interactions in individual participant data meta-analysis of randomized trials, and guidance for practitioners. *J Clin Epidemiol,* 64**,** 949-67.

FLENADY, V., KOOPMANS, L., MIDDLETON, P., FROEN, J. F., SMITH, G. C., GIBBONS, K., COORY, M., GORDON, A., ELLWOOD, D., MCINTYRE, H. D., FRETTS, R. & EZZATI, M. 2011. Major risk factors for stillbirth in high-income countries: a systematic review and meta-analysis. *Lancet,* 377**,** 1331-40.

FMF 2004. The 11-13+6 weeks scan. *Fetal Medicine Foundation*.

FOSGATE, G. T. 2006. Non-differential measurement error does not always bias diagnostic likelihood ratios towards the null. *Emerg Themes Epidemiol,* 3**,** 7.

GARDOSI, J., KADY, S. M., MCGEOWN, P., FRANCIS, A. & TONKS, A. 2005. Classification of stillbirth by relevant condition at death (ReCoDe): population based cohort study. *Bmj,* 331**,** 1113-7.

GEERSING, G. J., BOUWMEESTER, W., ZUITHOFF, P., SPIJKER, R., LEEFLANG, M. & MOONS, K. G. 2012. Search filters for finding prognostic and diagnostic prediction studies in Medline to enhance systematic reviews. *PLoS One,* 7**,** e32844.

GORELICK, M. H. 2006. Bias arising from missing data in predictive models. *J Clin Epidemiol,* 59**,** 1115-23.

GOYAL, N. K., HALL, E. S., GREENBERG, J. M. & KELLY, E. A. 2015. Risk Prediction for Adverse Pregnancy Outcomes in a Medicaid Population. *Journal of women's health (2002),* 24**,** 681-688.

GRASSI, G., BOMBELLI, M., BRAMBILLA, G., TREVANO, F. Q., DELL'ORO, R. & MANCIA, G. 2012. Total Cardiovascular Risk, Blood Pressure Variability and Adrenergic Overdrive in Hypertension: Evidence, Mechanisms and Clinical Implications. *Current Hypertension Reports,* 14**,** 333-338.

GREENLAND, S. & ROBINS, J. M. 1985. Confounding and misclassification. *Am J Epidemiol,* 122**,** 495-506.

GUOLO, A. 2008. Robust techniques for measurement error correction: a review. *Stat Methods Med Res,* 17**,** 555-80.

GUSTAFSON, P. 2003. *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*, CRC Press.

GUYATT, G. H., OXMAN, A. D., KUNZ, R., WOODCOCK, J., BROZEK, J., HELFAND, M., ALONSO-COELLO, P., FALCK-YTTER, Y., JAESCHKE, R., VIST, G., AKL, E. A., POST, P. N., NORRIS, S., MEERPOHL, J., SHUKLA, V. K., NASSER, M. & SCHUNEMANN, H. J. 2011. GRADE guidelines: 8. Rating the quality of evidence--indirectness. *J Clin Epidemiol,* 64**,** 1303-10.

HACK, K. E. A., DERKS, J. B., ELIAS, S. G., FRANX, A., ROOS, E. J., VOERMAN, S. K., BODE, C. L., KOOPMAN-ESSEBOOM, C. & VISSER, G. H. A. 2007. Increased perinatal mortality and morbidity in monochorionic versus dichorionic twin pregnancies: clinical implications of a large Dutch cohort study. *british journal of obstetrics and gynaecology,* 115**,** 58-67.

HANDLER, J. 2009. The Importance of Accurate Blood Pressure Measurement. *The Permanente Journal,* 13**,** 51-54.

HANLEY, J. A. & MCNEIL, B. J. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology,* 143**,** 29-36.

HARBORD, R. M., WHITING, P., STERNE, J. A., EGGER, M., DEEKS, J. J., SHANG, A. & BACHMANN, L. M. 2008. An empirical comparison of methods for meta-analysis of diagnostic accuracy showed hierarchical models are necessary. *J Clin Epidemiol,* 61**,** 1095-103.

HARRELL, F. E. 2001. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*, Springer.

HARRELL, F. E., LEE, K. L. & MARK, D. B. 1996. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine,* 15**,** 361-387.

HARTUNG, J. & KNAPP, G. 2001. A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Stat Med,* 20**,** 3875-89.

HAYBITTLE, J. L., BLAMEY, R. W., ELSTON, C. W., JOHNSON, J., DOYLE, P. J., CAMPBELL, F. C., NICHOLSON, R. I. & GRIFFITHS, K. 1982. A prognostic index in primary breast cancer. *Br J Cancer,* 45**,** 361-6.

HEAZELL, A. E. P., WHITWORTH, M. K., WHITCOMBE, J., GLOVER, S. W., BEVAN, C., BREWIN, J., CALDERWOOD, C., CANTER, A., JESSOP, F., JOHNSON, G., MARTIN, I. & METCALF, L. 2015. Research priorities for stillbirth: process overview and results from UK Stillbirth Priority Setting Partnership. *Ultrasound in Obstetrics & Gynecology,* 46**,** 641-647.

HEMINGWAY, H., CROFT, P., PEREL, P., HAYDEN, J. A., ABRAMS, K., TIMMIS, A., BRIGGS, A., UDUMYAN, R., MOONS, K. G. M., STEYERBERG, E. W., ROBERTS, I., SCHROTER, S., ALTMAN, D. G. & RILEY, R. D. 2013. Prognosis research strategy (PROGRESS) 1: A framework for researching clinical outcomes. *BMJ,* 346.

HERMSEN, L. A., LEONE, S. S., VAN DER WINDT, D. A., SMALBRUGGE, M., DEKKER, J. & VAN DER HORST, H. E. 2011. Functional outcome in older adults with joint pain and comorbidity: design of a prospective cohort study. *BMC Musculoskelet Disord,* 12**,** 241.

HIGGINS, J. P., THOMPSON, S. G. & SPIEGELHALTER, D. J. 2009. A re-evaluation of random-effects meta-analysis. *J R Stat Soc Ser A Stat Soc,* 172**,** 137-159.

HIGGINS, J. P. T., THOMPSON, S. G., DEEKS, J. J. & ALTMAN, D. G. 2003. Measuring inconsistency in meta-analyses. *BMJ,* 327**,** 557-560.

HILBE, J. M. 2011. *Negative Binomial Regression*, Cambridge University Press.

HILL, A. & ROBERTS, J. 1998. Body mass index: a comparison between self-reported and measured height and weight. *J Public Health Med,* 20**,** 206-10.

HILL, J. C., DUNN, K. M., LEWIS, M., MULLIS, R., MAIN, C. J., FOSTER, N. E. & HAY, E. M. 2008. A primary care back pain screening tool: Identifying patient subgroups for initial treatment. *Arthritis Care & Research,* 59**,** 632-641.

HILL, J. C., WHITEHURST, D. G. T., LEWIS, M., BRYAN, S., DUNN, K. M., FOSTER, N. E., KONSTANTINOU, K., MAIN, C. J., MASON, E., SOMERVILLE, S., SOWDEN, G., VOHORA, K. & HAY, E. M. 2011. Comparison of stratified primary care management for low back pain with current best practice (STarT Back): a randomised controlled trial. *The Lancet,* 378**,** 1560-1571.

HINGORANI, A. D., WINDT, D. A. V. D., RILEY, R. D., ABRAMS, K., MOONS, K. G. M., STEYERBERG, E. W., SCHROTER, S., SAUERBREI, W., ALTMAN, D. G. & HEMINGWAY, H. 2013. Prognosis research strategy (PROGRESS) 4: Stratified medicine research. *BMJ : British Medical Journal,* 346.

HOERL, A. E. & KENNARD, R. W. 1970. Ridge Regression: Applications to Nonorthogonal Problems. *Technometrics,* 12**,** 69-82.

HOERL, A. E. & KENNARD, R. W. 2000. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics,* 42**,** 80-86.

HOLDEN, M. A., BURKE, D. L., RUNHAAR, J., VAN DER WINDT, D., RILEY, R. D., DZIEDZIC, K., LEGHA, A., EVANS, A. L., ABBOTT, J. H., BAKER, K., BROWN, J., BENNELL, K. L., BOSSEN, D., BROSSEAU, L., CHAIPINYO, K., CHRISTENSEN, R., COCHRANE, T., DE ROOIJ, M., DOHERTY, M., FRENCH, H. P., HICKSON, S., HINMAN, R. S., HOPMAN-ROCK, M., HURLEY, M. V., INGRAM, C., KNOOP, J., KRAUSS, I., MCCARTHY, C., MESSIER, S. P., PATRICK, D. L., SAHIN, N., TALBOT, L. A., TAYLOR, R., TEIRLINCK, C. H., VAN MIDDELKOOP, M., WALKER, C. & FOSTER, N. E. 2017. Subgrouping and TargetEd Exercise pRogrammes for knee and hip OsteoArthritis (STEER OA): a systematic review update and individual participant data meta-analysis protocol. *BMJ Open,* 7**,** e018971.

HSIEH, F. Y., BLOCH, D. A. & LARSEN, M. D. 1998. A simple method of sample size calculation for linear and logistic regression. *Stat Med,* 17**,** 1623-34.

HUGHES, M. & FRANKS, I. 2015. *Essentials of Performance Analysis in Sport: Second Edition*, Taylor & Francis.

INGUI, B. J. & ROGERS, M. A. M. 2001. Searching for Clinical Prediction Rules in Medline. *Journal of the American Medical Informatics Association,* 8**,** 391-397.

INNESS, P. M. & DORLING, S. 2013. *Operational Weather Forecasting*, Wiley.

JANSSEN, K. J., DONDERS, A. R., HARRELL, F. E., JR., VERGOUWE, Y., CHEN, Q., GROBBEE, D. E. & MOONS, K. G. 2010. Missing covariate data in medical research: to impute is better than to ignore. *J Clin Epidemiol,* 63**,** 721-7.

JOHNSON, D. R. & YOUNG, R. 2011. Toward Best Practices in Analyzing Datasets with Missing Data: Comparisons and Recommendations. *Journal of Marriage and Family,* 73**,** 926-945.

JOLANI, S., DEBRAY, T. P., KOFFIJBERG, H., VAN BUUREN, S. & MOONS, K. G. 2015. Imputation of systematically missing predictors in an individual participant data meta-analysis: a generalized approach using MICE. *Stat Med,* 34**,** 1841-63.

JUREK, A. M., GREENLAND, S., MALDONADO, G. & CHURCH, T. R. 2005. Proper interpretation of non-differential misclassification effects: expectations vs observations. *International Journal of Epidemiology,* 34**,** 680-687.

KAISER, I., DIEHL, K., HEPPT, M. V., MATHES, S., PFAHLBERG, A. B., STEEB, T., UTER, W. & GEFELLER, O. 2022. Reporting Quality of Studies Developing and Validating Melanoma Prediction Models: An Assessment Based on the TRIPOD Statement. *Healthcare (Basel),* 10.

KANG, H. 2013. The prevention and handling of the missing data. *Korean Journal of Anesthesiology,* 64**,** 402-406.

KAYODE, G. A., GROBBEE, D. E., AMOAKOH-COLEMAN, M., ADELEKE, I. T., ANSAH, E., DE GROOT, J. A. H. & KLIPSTEIN-GROBUSCH, K. 2016. Predicting stillbirth in a low resource setting. *BMC pregnancy and childbirth,* 16**,** 274-274.

KHALIL, A., RODGERS, M., BASCHAT, A., BHIDE, A., GRATACOS, E., HECHER, K., KILBY, M. D., LEWI, L., NICOLAIDES, K. H., OEPKES, D., RAINE-FENNING, N., REED, K., SALOMON, L. J., SOTIRIADIS, A., THILAGANATHAN, B. & VILLE, Y. 2016. ISUOG Practice Guidelines: role of ultrasound in twin pregnancy. *Ultrasound Obstet Gynecol,* 47**,** 247-63.

KHUDYAKOV, P., GORFINE, M., ZUCKER, D. & SPIEGELMAN, D. 2015. The impact of covariate measurement error on risk prediction. *Stat Med,* 34**,** 2353-67.

KLEINROUWELER, C. E., CHEONG-SEE, F. M., COLLINS, G. S., KWEE, A., THANGARATINAM, S., KHAN, K. S., MOL, B. W., PAJKRT, E., MOONS, K. G. & SCHUIT, E. 2016. Prognostic models in obstetrics: available, but far from applicable. *Am J Obstet Gynecol,* 214**,** 79-90.e36.

KOBAYASHI, H. 2013. Effect of measurement duration on accuracy of pulse-counting. *Ergonomics,* 56**,** 1940-1944.

KOUROU, K., EXARCHOS, T. P., EXARCHOS, K. P., KARAMOUZIS, M. V. & FOTIADIS, D. I. 2015. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal,* 13**,** 8-17.

KOVALCHIK, S. A. & CUMBERLAND, W. G. 2012. Using aggregate data to estimate the standard error of a treatment-covariate interaction in an individual patient data meta-analysis. *Biom J,* 54**,** 370-84.

KRISTENSEN, P. 1992. Bias from nondifferential but dependent misclassification of exposure and outcome. *Epidemiology,* 3**,** 210-5.

KUSS, O., HOYER, A. & SOLMS, A. 2014. Meta-analysis for diagnostic accuracy studies: a new statistical model using beta-binomial distributions and bivariate copulas. *Stat Med,* 33**,** 17-30.

LANGAN, D., HIGGINS, J. P. T., JACKSON, D., BOWDEN, J., VERONIKI, A. A., KONTOPANTELIS, E., VIECHTBAUER, W. & SIMMONDS, M. 2019. A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Res Synth Methods,* 10**,** 83-98.

LASH, T. L. & FINK, A. K. 2003. Re: "Neighborhood environment and loss of physical function in older adults: evidence from the Alameda County Study". *Am J Epidemiol,* 157**,** 472-3.

LECKIE, G. & CHARLTON, C. 2013. runmlwin: A Program to Run the MLwiN Multilevel Modeling Software from within Stata. *2013,* 52**,** 40.

LEE, H. S. & KRISCHER, J. P. 2017. A new framework for prediction and variable selection for uncommon events in a large prospective cohort study. *Model Assist Stat Appl,* 12**,** 227-237.

LEEFLANG, M. M., DEEKS, J. J., RUTJES, A. W., REITSMA, J. B. & BOSSUYT, P. M. 2012. Bivariate meta-analysis of predictive values of diagnostic tests can be an alternative to bivariate meta-analysis of sensitivity and specificity. *J Clin Epidemiol,* 65**,** 1088-97.

LEVIS, B., BENEDETTI, A., LEVIS, A. W., IOANNIDIS, J. P. A., SHRIER, I., CUIJPERS, P., GILBODY, S., KLODA, L. A., MCMILLAN, D., PATTEN, S. B., STEELE, R. J., ZIEGELSTEIN, R. C., BOMBARDIER, C. H., DE LIMA OSÓRIO, F., FANN, J. R., GJERDINGEN, D., LAMERS, F., LOTRAKUL, M., LOUREIRO, S. R., LÖWE, B., SHAABAN, J., STAFFORD, L., VAN WEERT, H., WHOOLEY, M. A., WILLIAMS, L. S., WITTKAMPF, K. A., YEUNG, A. S. & THOMBS, B. D. 2017. Selective Cutoff Reporting in Studies of Diagnostic Test Accuracy: A Comparison of Conventional and Individual-Patient-Data Meta-Analyses of the Patient Health Questionnaire-9 Depression Screening Tool. *Am J Epidemiol,* 185**,** 954-964.

LICHT-STRUNK, E., BEEKMAN, A. T., DE HAAN, M. & VAN MARWIJK, H. W. 2009. The prognosis of undetected depression in older general practice patients. A one year follow-up study. *J Affect Disord,* 114**,** 310-5.

LITTLE, R. J. A. 1992. Regression With Missing X's: A Review. *Journal of the American Statistical Association,* 87**,** 1227-1237.

MACASKILL, P. 2004. Empirical Bayes estimates generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis. *J Clin Epidemiol,* 57**,** 925-32.

MACASKILL, P., GATSONIS, C. A., DEEKS, J. J., HARBORD, R. M. & TAKWOINGI, Y. 2010. Chapter 10: Analysing and Presenting Results. *In:* DEEKS, J. J., BOSSUYT, P. M. & GATSONIS, C. A. (eds.) *Cochrane Handbook for Systematic Reviews of Diagostic Test Accuracy.* Version 1.0 ed.

MACKIE, F. L., MORRIS, R. K. & KILBY, M. D. 2017. The prediction, diagnosis and management of complications in monochorionic twin pregnancies: the OMMIT (Optimal Management of Monochorionic Twins) study. *BMC Pregnancy Childbirth,* 17**,** 153.

MACKIE, F. L., WHITTLE, R., MORRIS, R. K., HYETT, J., RILEY, R. D. & KILBY, M. D. 2019. First-trimester ultrasound measurements and maternal serum biomarkers as prognostic factors in monochorionic twins: a cohort study. *Diagn Progn Res,* 3**,** 9.

MACY, E. M., HAYES, T. E. & TRACY, R. P. 1997. Variability in the measurement of C-reactive protein in healthy subjects: implications for reference intervals and epidemiological applications. *Clinical Chemistry,* 43**,** 52-58.

MALACOVA, E., TIPPAYA, S., BAILEY, H. D., CHAI, K., FARRANT, B. M., GEBREMEDHIN, A. T., LEONARD, H., MARINOVICH, M. L., NASSAR, N., PHATAK, A., RAYNES-GREENOW, C., REGAN, A. K., SHAND, A. W., SHEPHERD, C. C. J., SRINIVASJOIS, R., TESSEMA, G. A. & PEREIRA, G. 2020. Stillbirth risk prediction using machine learning for a large cohort of births from Western Australia, 1980–2015. *Scientific Reports,* 10**,** 5354.

MALLEN, C. D., PEAT, G., THOMAS, E., DUNN, K. M. & CROFT, P. R. 2007. Prognostic factors for musculoskeletal pain in primary care: a systematic review. *Br J Gen Pract,* 57**,** 655-61.

MALLEN, C. D., PEAT, G., THOMAS, E., WATHALL, S., WHITEHURST, T., CLEMENTS, C., BAILEY, J., GRAY, J. & CROFT, P. R. 2006. The assessment of the prognosis of musculoskeletal conditions in older adults presenting to general practice: a research protocol. *BMC Musculoskelet Disord,* 7**,** 84.

MANKTELOW, B. M., SMITH, L. K., EVANS, T. A., HYMAN-TAYLOR, P., KURINCZUK, J. J., FIELD, D. J., SMITH, P. W. & DRAPER, E. S. 2015. Perinatal Mortality Surveillance Report UK Perinatal Deaths for births from January to December 2013. *In:* COLLABORATION, M.-U. (ed.). Leicester: The Infant Mortality and Morbidity Group, Department of Health Sciences, University of Leicester.

MANTYSELKA, P., KUMPUSALO, E., AHONEN, R. & TAKALA, J. 2001. Patients' versus general practitioners' assessments of pain intensity in primary care patients with non-cancer pain. *Br J Gen Pract,* 51**,** 995-7.

MASTRODIMA, S., AKOLEKAR, R., YERLIKAYA, G., TZELEPIS, T. & NICOLAIDES, K. H. 2016. Prediction of stillbirth from biochemical and biophysical markers at 11-13 weeks. *Ultrasound Obstet Gynecol,* 48**,** 613-617.

MATHEMATICS STACK EXCHANGE. 2018. *How do I combine standard deviations of two groups?* [Online]. Available: https://math.stackexchange.com/questions/2971315/how-do-i-combine-standard-deviations-of-two-groups [Accessed 30/03/2023 2023].

MITCHELL, T. M. 1997. *Machine learning*, McGraw-hill New York.

MOLDENHAUER, J. S. & JOHNSON, M. P. 2015. Diagnosis and Management of Complicated Monochorionic Twins. *Clin Obstet Gynecol,* 58**,** 632-42.

MOONS, K. G., ALTMAN, D. G., VERGOUWE, Y. & ROYSTON, P. 2009a. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *Bmj,* 338**,** b606.

MOONS, K. G., DE GROOT, J. A., BOUWMEESTER, W., VERGOUWE, Y., MALLETT, S., ALTMAN, D. G., REITSMA, J. B. & COLLINS, G. S. 2014. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med,* 11**,** e1001744.

MOONS, K. G., KENGNE, A. P., GROBBEE, D. E., ROYSTON, P., VERGOUWE, Y., ALTMAN, D. G. & WOODWARD, M. 2012a. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart,* 98**,** 691-8.

MOONS, K. G., KENGNE, A. P., WOODWARD, M., ROYSTON, P., VERGOUWE, Y., ALTMAN, D. G. & GROBBEE, D. E. 2012b. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart,* 98**,** 683-90.

MOONS, K. G., ROYSTON, P., VERGOUWE, Y., GROBBEE, D. E. & ALTMAN, D. G. 2009b. Prognosis and prognostic research: what, why, and how? *Bmj,* 338**,** b375.

MOONS, K. G. M., ALTMAN, D. G., REITSMA, J. B., IOANNIDIS, J. P. A., MACASKILL, P., STEYERBERG, E. W., VICKERS, A. J., RANSOHOFF, D. F. & COLLINS, G. S. 2015. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and ElaborationThe TRIPOD Statement: Explanation and Elaboration. *Annals of Internal Medicine,* 162**,** W1-W73.

MORI, H., KOBARA, H., TSUSHIMI, T., NISHIYAMA, N., FUJIHARA, S. & MASAKI, T. 2015. Unavoidable Human Errors of Tumor Size Measurement during Specimen Attachment after Endoscopic Resection: A Clinical Prospective Study. *PLoS ONE,* 10**,** e0121798.

NEILSON, J. & KILBY, M. 2008. Management of monochorionic twin pregnancy: Green-top guideline no. 51. *Royal College of Obstetricians & Gynaecologists*.

NICE 2011. Multiple Pregnancy: The Management of Twin and Triplet Pregnancies in the Antenatal Period. *NICE clinical guideline.* Manchester: National Institute for Health and Clinical Excellence.

NIJMAN, S. W. J., GROENHOF, T. K. J., HOOGLAND, J., BOTS, M. L., BRANDJES, M., JACOBS, J. J. L., ASSELBERGS, F. W., MOONS, K. G. M. & DEBRAY, T. P. A. 2021. Real-time imputation of missing predictor values improved the application of prediction models in daily practice. *J Clin Epidemiol,* 134**,** 22-34.

NIKOLOULOPOULOS, A. K. 2017. A vine copula mixed effect model for trivariate meta-analysis of diagnostic test accuracy studies accounting for disease prevalence. *Stat Methods Med Res,* 26**,** 2270-2286.

OEPKES, D. & SUETERS, M. 2017. Antenatal fetal surveillance in multiple pregnancies. *Best Pract Res Clin Obstet Gynaecol,* 38**,** 59-70.

OGUNDIMU, E. O., ALTMAN, D. G. & COLLINS, G. S. 2016. Adequate sample size for developing prediction models is not simply related to events per variable. *J Clin Epidemiol,* 76**,** 175-82.

ORGANIZATION, W. H. 2006. Neonatal and perinatal mortality : country, regional and global estimates. Geneva: World Health Organization.

PAYNE, B. A., GROEN, H., UKAH, U. V., ANSERMINO, J. M., BHUTTA, Z., GROBMAN, W., HALL, D. R., HUTCHEON, J. A., MAGEE, L. A. & VON DADELSZEN, P. 2015. Development and internal validation of a multivariable model to predict perinatal death in pregnancy hypertension. *Pregnancy Hypertension: An International Journal of Women's Cardiovascular Health,* 5**,** 315-321.

PEAKE, M. & WHITING, M. 2006. Measurement of Serum Creatinine – Current Status and Future Goals. *Clinical Biochemist Reviews,* 27**,** 173-184.

PEDUZZI, P., CONCATO, J., KEMPER, E., HOLFORD, T. R. & FEINSTEIN, A. R. 1996. A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology,* 49**,** 1373-1379.

PICARD, R. R. & BERK, K. N. 1990. Data splitting. *The American Statistician,* 44**,** 140-147.

PODSAKOFF, P. M., MACKENZIE, S. B., LEE, J. Y. & PODSAKOFF, N. P. 2003. Common method biases in behavioral research: a critical review of the literature and recommended remedies. *J Appl Psychol,* 88**,** 879-903.

POLLEY, M.-Y. C., LEUNG, S. C. Y., MCSHANE, L. M., GAO, D., HUGH, J. C., MASTROPASQUA, M. G., VIALE, G., ZABAGLO, L. A., PENAULT-LLORCA, F., BARTLETT, J. M. S., GOWN, A. M., SYMMANS, W. F., PIPER, T., MEHL, E., ENOS, R. A., HAYES, D. F., DOWSETT, M. & NIELSEN, T. O. 2013. An International Ki67 Reproducibility Study. *JNCI Journal of the National Cancer Institute,* 105**,** 1897-1906.

POYNARD, T., CALÈS, P., PASTA, L., IDEO, G., PASCAL, J. P., PAGLIARO, L. & LEBREC, D. 1991. Beta-adrenergic-antagonist drugs in the prevention of gastrointestinal bleeding in patients with cirrhosis and esophageal varices. An analysis of data and prognostic factors in 589 patients from four randomized clinical trials. Franco-Italian Multicenter Study Group. *N Engl J Med,* 324**,** 1532-8.

PRENTICE, R. L. 1982. Covariate Measurement Errors and Parameter Estimation in a Failure Time Regression Model. *Biometrika,* 69**,** 331-342.

PUTTER, H., FIOCCO, M. & STIJNEN, T. 2010. Meta-analysis of diagnostic test accuracy studies with multiple thresholds using survival methods. *Biom J,* 52**,** 95-110.

QUINTERO, R. A., MORALES, W. J., ALLEN, M. H., BORNICK, P. W., JOHNSON, P. K. & KRUGER, M. 1999. Staging of twin-twin transfusion syndrome. *J Perinatol,* 19**,** 550-5.

RADANOV, B. P., DI STEFANO, G., SCHNIDRIG, A. & BALLINARI, P. 1991. Role of psychosocial stress in recovery from common whiplash [see comment]. *Lancet,* 338**,** 712-5.

RCPCH 2016. Early years - UK-WHO growth charts and resources.

REDDY, U. M., LAUGHON, S. K., SUN, L., TROENDLE, J., WILLINGER, M. & ZHANG, J. 2010. Prepregnancy Risk Factors for Antepartum Stillbirth in the United States. *Obstetrics & Gynecology,* 116**,** 1119-1126.

REINHARD, M., ERLANDSEN, E. J. & RANDERS, E. 2009. Biological variation of cystatin C and creatinine. *Scand J Clin Lab Invest,* 69**,** 831-6.

RESCHE-RIGON, M. & WHITE, I. R. 2016. Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *Stat Methods Med Res*.

RILEY, R. D., ENSOR, J., SNELL, K. I. E., DEBRAY, T. P. A., ALTMAN, D. G., MOONS, K. G. M. & COLLINS, G. S. 2016. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *BMJ,* 353**,** i3140.

RILEY, R. D., ENSOR, J., SNELL, K. I. E., HARRELL, F. E., MARTIN, G. P., REITSMA, J. B., MOONS, K. G. M., COLLINS, G. & VAN SMEDEN, M. 2020. Calculating the sample size required for developing a clinical prediction model. *BMJ,* 368**,** m441.

RILEY, R. D., HATTLE, M., COLLINS, G. S., WHITTLE, R. & ENSOR, J. 2022. Calculating the power to examine treatment-covariate interactions when planning an individual participant data meta-analysis of randomized trials with a binary outcome. *Stat Med*.

RILEY, R. D., HAYDEN, J. A., STEYERBERG, E. W., MOONS, K. G. M., ABRAMS, K., KYZAS, P. A., MALATS, N., BRIGGS, A., SCHROTER, S., ALTMAN, D. G., HEMINGWAY, H. & FOR THE, P. G. 2013. Prognosis Research Strategy (PROGRESS) 2: Prognostic Factor Research. *PLoS Med,* 10**,** e1001380.

RILEY, R. D., HIGGINS, J. P. T. & DEEKS, J. J. 2011. Interpretation of random effects meta-analyses. *BMJ,* 342**,** d549.

RILEY, R. D., SAUERBREI, W. & ALTMAN, D. G. 2009. Prognostic markers in cancer: the evolution of evidence from single studies to meta-analysis, and beyond. *Br J Cancer,* 100**,** 1219-29.

RILEY, R. D., SNELL, K. I., ENSOR, J., BURKE, D. L., HARRELL JR, F. E., MOONS, K. G. & COLLINS, G. S. 2019a. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Statistics in Medicine,* 38**,** 1276-1296.

RILEY, R. D., SNELL, K. I. E., MARTIN, G. P., WHITTLE, R., ARCHER, L., SPERRIN, M. & COLLINS, G. S. 2021a. Penalization and shrinkage methods produced unreliable clinical prediction models especially when sample size was small. *J Clin Epidemiol,* 132**,** 88-96.

RILEY, R. D., TIERNEY, J. & STEWART, L. A. 2021b. *Individual Participant Data Meta-Analysis: A Handbook for Healthcare Research,* Chichester, Wiley.

RILEY, R. D., VAN DER WINDT, D., CROFT, P. & MOONS, K. G. M. 2019b. *Prognosis Research in Health Care: Concepts, Methods, and Impact*, Oxford University Press.

ROEHRBORN, C. G., PICKENS, G. J. & CARMODY, T., 3RD 1996. Variability of repeated serum prostate-specific antigen (PSA) measurements within less than 90 days in a well-defined patient population. *Urology,* 47**,** 59-66.

ROLAND, M. & MORRIS, R. 1983. A study of the natural history of back pain. Part I: development of a reliable and sensitive measure of disability in low-back pain. *Spine (Phila Pa 1976),* 8**,** 141-4.

ROSELLA, L. C., COREY, P., STUKEL, T. A., MUSTARD, C., HUX, J. & MANUEL, D. G. 2012. The influence of measurement error on calibration, discrimination, and overall estimation of a risk prediction model. *Population Health Metrics,* 10**,** 20.

ROTHMAN, K. J., GREENLAND, S. & LASH, T. L. 2008. *Modern Epidemiology*, Wolters Kluwer Health/Lippincott Williams & Wilkins.

ROYSTON, P., ALTMAN, D. G. & SAUERBREI, W. 2006. Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in Medicine,* 25**,** 127-141.

ROYSTON, P., AMBLER, G. & SAUERBREI, W. 1999. The use of fractional polynomials to model continuous risk variables in epidemiology. *Int J Epidemiol,* 28**,** 964-74.

ROYSTON, P., MOONS, K. G. M., ALTMAN, D. G. & VERGOUWE, Y. 2009. Prognosis and prognostic research: Developing a prognostic model. *BMJ,* 338.

ROYSTON, P. & SAUERBREI, W. 2008. *Multivariable Model - Building: A Pragmatic Approach to Regression Anaylsis based on Fractional Polynomials for Modelling Continuous Variables*, Wiley.

RUBIN, D. B. 1987. *Multiple Imputation for Nonresponse in Surveys.,* New York, Wiley.

RUTTER, C. M. & GATSONIS, C. A. 2001. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Stat Med,* 20**,** 2865-84.

SAWERS, L. 2013. Measuring and modelling concurrency. *Journal of the International AIDS Society,* 16**,** 17431.

SCHEELE, J., LUIJSTERBURG, P. A., FERREIRA, M. L., MAHER, C. G., PEREIRA, L., PEUL, W. C., VAN TULDER, M. W., BOHNEN, A. M., BERGER, M. Y., BIERMA-ZEINSTRA, S. M. & KOES, B. W. 2011. Back complaints in the elders (BACE); design of cohort studies in primary care: an international consortium. *BMC Musculoskelet Disord,* 12**,** 193.

SCHMOOR, C., SAUERBREI, W. & SCHUMACHER, M. 2000. Sample size considerations for the evaluation of prognostic factors in survival analysis. *Statistics in Medicine,* 19**,** 441-452.

SCOTT, J. & HUSKISSON, E. C. 1979. Accuracy of subjective measurements made with or without previous scores: an important source of error in serial measurement of subjective states. *Annals of the Rheumatic Diseases,* 38**,** 558-559.

SEPULVEDA, W., SEBIRE, N. J., HUGHES, K., ODIBO, A. & NICOLAIDES, K. H. 1996. The lambda sign at 10-14 weeks of gestation as a predictor of chorionicity in twin pregnancies. *Ultrasound Obstet Gynecol,* 7**,** 421-3.

SHILLAN, D., STERNE, J. A. C., CHAMPNEYS, A. & GIBBISON, B. 2019. Use of machine learning to analyse routinely collected intensive care unit data: a systematic review. *Critical Care,* 23**,** 284.

SIDIK, K. & JONKMAN, J. N. 2002. A simple confidence interval for meta-analysis. *Stat Med,* 21**,** 3153-9.

SIMMONDS, M. C. & HIGGINS, J. P. 2007. Covariate heterogeneity in meta-analysis: criteria for deciding between meta-regression and individual patient data. *Stat Med,* 26**,** 2982-99.

SIMMONDS, M. C., HIGGINS, J. P. T., STEWART, L. A., TIERNEY, J. F., CLARKE, M. J. & THOMPSON, S. G. 2005. Meta-analysis of individual patient data from randomized trials: a review of methods used in practice. *Clinical Trials,* 2**,** 209-217.

SMITH, G. C. S., YU, C. K. H., PAPAGEORGHIOU, A. T., CACHO, A. M., NICOLAIDES, K. H. & FETAL MEDICINE FOUNDATION SECOND TRIMESTER SCREENING, G. 2007. Maternal uterine artery Doppler flow velocimetry and the risk of stillbirth. *Obstetrics and gynecology,* 109**,** 144-151.

SNELL, K. I., ENSOR, J., DEBRAY, T. P., MOONS, K. G. & RILEY, R. D. 2018. Meta-analysis of prediction model performance across multiple studies: Which scale helps ensure between-study normality for the C-statistic and calibration measures? *Stat Methods Med Res,* 27**,** 3505-3522.

SNELL, K. I. E., ARCHER, L., ENSOR, J., BONNETT, L. J., DEBRAY, T. P. A., PHILLIPS, B., COLLINS, G. S. & RILEY, R. D. 2021. External validation of clinical prediction models: simulation-based sample size calculations were more reliable than rules-of-thumb. *J Clin Epidemiol,* 135**,** 79-89.

SORAHAN, T. & GILTHORPE, M. S. 1994. Non-differential misclassification of exposure always leads to an underestimate of risk: an incorrect conclusion. *Occup Environ Med,* 51**,** 839-40.

STATON, L. J., PANDA, M., CHEN, I., GENAO, I., KURZ, J., PASANEN, M., MECHABER, A. J., MENON, M., O'RORKE, J., WOOD, J., ROSENBERG, E., FAESLIS, C., CAREY, T., CALLESON, D. & CYKERT, S. 2007. When race matters: disagreement in pain perception between patients and their physicians in primary care. *J Natl Med Assoc,* 99**,** 532-8.

STERNE, J. A. C., WHITE, I. R., CARLIN, J. B., SPRATT, M., ROYSTON, P., KENWARD, M. G., WOOD, A. M. & CARPENTER, J. R. 2009. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ,* 338.

STEYERBERG, E. 2010. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*, Springer New York.

STEYERBERG, E. W., MOONS, K. G., VAN DER WINDT, D. A., HAYDEN, J. A., PEREL, P., SCHROTER, S., RILEY, R. D., HEMINGWAY, H. & ALTMAN, D. G. 2013. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med,* 10**,** e1001381.

STREINER, D. L., NORMAN, G. R. & CAIRNEY, J. 2014. *Health Measurement Scales: A practical guide to their development and use: A practical guide to their development and use*, OUP Oxford.

SUN, G. W., SHOOK, T. L. & KAY, G. L. 1996. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. *J Clin Epidemiol,* 49**,** 907-16.

THOMPSON, S., KAPTOGE, S., WHITE, I., WOOD, A., PERRY, P. & DANESH, J. 2010. Statistical methods for the time-to-event analysis of individual participant data from multiple epidemiological studies. *Int J Epidemiol,* 39**,** 1345-59.

TIBSHIRANI, R. 1996. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological),* 58**,** 267-288.

TOWNSEND, R., MANJI, A., ALLOTEY, J., HEAZELL, A., JORGENSEN, L., MAGEE, L., MOL, B., SNELL, K., RILEY, R., SANDALL, J., SMITH, G., PATEL, M., THILAGANATHAN, B., VON DADELSZEN, P., THANGARATINAM, S. & KHALIL, A. 2021a. Can risk prediction models help us individualise stillbirth prevention? A systematic review and critical appraisal of published risk models. *BJOG: An International Journal of Obstetrics & Gynaecology,* 128**,** 214-224.

TOWNSEND, R., SILEO, F., ALLOTEY, J., DODDS, J., HEAZELL, A., JORGENSEN, L., KIM, V., MAGEE, L., MOL, B., SANDALL, J., SMITH, G., THILAGANATHAN, B., VON DADELSZEN, P., THANGARATINAM, S. & KHALIL, A. 2021b. Prediction of stillbirth: an umbrella review of evaluation of prognostic variables. *BJOG: An International Journal of Obstetrics & Gynaecology,* 128**,** 238-250.

TRUDELL, A. S., TUULI, M. G., COLDITZ, G. A., MACONES, G. A. & ODIBO, A. O. 2017. A stillbirth calculator: Development and internal validation of a clinical prediction model to quantify stillbirth risk. *PLoS One,* 12**,** e0173461.

TSO, E., ELSON, P., VANLENTE, F. & MARKMAN, M. 2006. The "real-life" variability of CA-125 in ovarian cancer patients. *Gynecol Oncol,* 103**,** 141-4.

TUXEN, M. K., SOLETORMOS, G., PETERSEN, P. H., SCHIOLER, V. & DOMBERNOWSKY, P. 1999. Assessment of biological variation and analytical imprecision of CA 125, CEA, and TPA in relation to monitoring of ovarian cancer. *Gynecol Oncol,* 74**,** 12-22.

UNAL, B., CRITCHLEY, J. A. & CAPEWELL, S. 2004. Explaining the decline in coronary heart disease mortality in England and Wales between 1981 and 2000. *Circulation,* 109**,** 1101-7.

VAN CALSTER, B., MCLERNON, D. J., VAN SMEDEN, M., WYNANTS, L., STEYERBERG, E. W., BOSSUYT, P., COLLINS, G. S., MACASKILL, P., MCLERNON, D. J., MOONS, K. G. M., STEYERBERG, E. W., VAN CALSTER, B., VAN SMEDEN, M., VICKERS, ANDREW J., ON BEHALF OF TOPIC GROUP 'EVALUATING DIAGNOSTIC, T. & PREDICTION MODELS' OF THE, S. I. 2019. Calibration: the Achilles heel of predictive analytics. *BMC Medicine,* 17**,** 230.

VAN DER PLOEG, T., AUSTIN, P. C. & STEYERBERG, E. W. 2014. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Medical Research Methodology,* 14**,** 137.

VAN DER WINDT, D. A., KOES, B. W., DEVILLE, W., BOEKE, A. J., DE JONG, B. A. & BOUTER, L. M. 1998. Effectiveness of corticosteroid injections versus physiotherapy for treatment of painful stiff shoulder in primary care: randomised trial. *Bmj,* 317**,** 1292-6.

VAN WALRAVEN, C., DAVIS, D., FORSTER, A. J. & WELLS, G. A. 2004. Time-dependent bias was common in survival analyses published in leading clinical journals. *J Clin Epidemiol,* 57**,** 672-82.

VELLAMKONDU, A., VASUDEVA, A., BHAT, R. G., KAMATH, A., AMIN, S. V., RAI, L. & KUMAR, P. 2017. Risk Assessment at 11-14-Week Antenatal Visit: A Tertiary Referral Center Experience from South India. *J Obstet Gynaecol India,* 67**,** 421-427.

VERGOUWE, Y., STEYERBERG, E. W., EIJKEMANS, M. J. & HABBEMA, J. D. 2005. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *J Clin Epidemiol,* 58**,** 475-83.

VICKERS, A. J., CRONIN, A. M., ELKIN, E. B. & GONEN, M. 2008. Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. *BMC Med Inform Decis Mak,* 8**,** 53.

VICKERS, A. J. & ELKIN, E. B. 2006. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making,* 26**,** 565-74.

VICKERS, A. J., VAN CALSTER, B. & STEYERBERG, E. W. 2016. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ,* 352**,** i6.

VON KORFF, M. 2013. Tailoring chronic pain care by brief assessment of impact and prognosis: comment on "Point-of-care prognosis for common musculoskeletal pain in older adults". *JAMA Intern Med,* 173**,** 1126-7.

VON KORFF, M., DEYO, R. A., CHERKIN, D. & BARLOW, W. 1993. Back pain in primary care. Outcomes at 1 year. *Spine (Phila Pa 1976),* 18**,** 855-62.

VON KORFF, M., ORMEL, J., KEEFE, F. J. & DWORKIN, S. F. 1992. Grading the severity of chronic pain. *Pain,* 50**,** 133-49.

WALKER, A. M. & BLETTNER, M. 1985. Comparing imperfect measures of exposure. *Am J Epidemiol,* 121**,** 783-90.

WARDENAAR, K. J., CONRADI, H. J. & DE JONGE, P. 2014. Data-driven course trajectories in primary care patients with major depressive disorder. *Depress Anxiety,* 31**,** 778-86.

WESSLER, B. S., LANA LAI, Y. H., KRAMER, W., CANGELOSI, M., RAMAN, G., LUTZ, J. S. & KENT, D. M. 2015. Clinical Prediction Models for Cardiovascular Disease: The Tufts PACE CPM Database. *Circulation. Cardiovascular quality and outcomes,* 8**,** 368-375.

WHITE, I., FROST, C. & TOKUNAGA, S. 2001. Correcting for measurement error in binary and continuous variables using replicates. *Stat Med,* 20**,** 3441-57.

WHITE, I. R., ROYSTON, P. & WOOD, A. M. 2011. Multiple imputation using chained equations: Issues and guidance for practice. *Stat Med,* 30**,** 377-99.

WHITEHEAD, A. & WHITEHEAD, J. 1991. A general parametric approach to the meta-analysis of randomized clinical trials. *Stat Med,* 10**,** 1665-77.

WHITNEY, C. W. & VON KORFF, M. 1992. Regression to the mean in treated versus untreated chronic pain. *Pain,* 50**,** 281-5.

WHITTEMORE, A. S. 1981. Sample Size for Logistic Regression with Small Response Probability. *Journal of the American Statistical Association,* 76**,** 27-32.

WHITTLE, R., PEAT, G., BELCHER, J., COLLINS, G. S. & RILEY, R. D. 2018. Measurement error and timing of predictor values for multivariable risk prediction models are poorly reported. *Journal of Clinical Epidemiology,* 102**,** 38-49.

WHITTLE, R., ROYLE, K.-L., JORDAN, K. P., RILEY, R. D., MALLEN, C. D. & PEAT, G. 2017. Prognosis research ideally should measure time-varying predictors at their intended moment of use. *Diagnostic and Prognostic Research,* 1**,** 1.

WINKEL, P., STATLAND, B. E. & BOKELUND, H. 1974. Factors contributing to intra-individual variation of serum constituents: 5. Short-term day-to-day and within-hour variation of serum constituents in healthy subjects. *Clin Chem,* 20**,** 1520-7.

WOLFF, R. F., MOONS, K. G. M., RILEY, R. D., WHITING, P. F., WESTWOOD, M., COLLINS, G. S., REITSMA, J. B., KLEIJNEN, J. & MALLETT, S. 2019. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Ann Intern Med,* 170**,** 51-58.

YERLIKAYA, G., AKOLEKAR, R., MCPHERSON, K., SYNGELAKI, A. & NICOLAIDES, K. H. 2016. Prediction of stillbirth from maternal demographic and pregnancy characteristics. *Ultrasound Obstet Gynecol,* 48**,** 607-612.

ZOU, H. & HASTIE, T. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 67**,** 301-320.