Keele University

Doctor of Philosophy

Application and development of statistical methods for prediction modelling in healthcare research

- Lucinda Archer -

A thesis submitted to Keele University for the degree of

DOCTOR OF PHILOSOPHY

March, 2025

Abstract

This thesis explores methods and applications for prediction modelling in a healthcare setting, focusing on continuous outcomes, sample size, and external validation of model performance.

It begins by discussing current practices around the dichotomisation of birthweight, along with the issues associated with the dichotomisation of continuous outcomes prior to modelling. Methods are proposed to retain model development on the continuous outcome scale and subsequently generate predicted probabilities for the dichotomised outcome, if needed.

Models for continuous and dichotomised pain score are externally validated, demonstrating large uncertainty in statistical measures of predictive performance due to the small sample size available for validation. This motivates development of new sample size calculations to target precise estimation of performance when externally validating a clinical prediction model with a continuous outcome.

A further external validation shows precision in performance estimates when using individual participant data, combining data from multiple sources to boost the sample size for validating a prediction model for continuous birthweight. Each included cohort surpassed the minimum recommended sample size, based on the newly proposed methods, thus high precision could be expected. However, accounting for heterogeneity in performance across included populations through meta-analysis led to wider confidence intervals for pooled performance statistics than in any individual cohort.

Heterogeneity in model performance is further demonstrated in the external validation of a

prediction model for serious falls. This validation involved utilising electronic health records to assess model performance across a range of general practice populations. Pooled performance estimates, though precise on average, hid the large variation in model performance across practices, giving an unrealistic summary of how the model might perform in practice.

In summary, the thesis demonstrates the importance of methodological rigour within clinical prediction model research, to ensure efficient and rigorous models are produced and evaluated.

Acknowledgements

Firstly, I would like to give thanks to my supervisors, especially for your help with the final push to get this finished. Your involvement and insights have been invaluable. Also to my colleagues and friends at Keele University, whose support has made this time enjoyable and memorable.

To my family: Mum, Dad, Conor, Owen and Suzi, and, of course, the Coven. You rule. Georgie: for being my very best friend, and for making me laugh (and laugh and laugh) when I needed it most. TJ: your beautiful smile has got me through these last two years, never change. Cassie and Louis: thank you for always being there to talk to, and for keeping me sane. Your loving presence has been my biggest source of comfort and motivation. Kelly, for your unwavering support and encouragement. Thank you for believing in me and for putting up with me all these years.

Lastly, to all participants, collaborators, and funders on the projects that have made up this thesis, without whom this research would not have been possible.

Production babies:

Isabelle; Georgie; Rowan; Daisy; Ryan; Evelyn; Jamie; Saul; TJ; Eva; Dylan; Abe; Roisin.

Contents

List of Figures

List of Tables

Abbreviations

AUC	Area Under the (ROC) Curve
BMI	Body Mass Index
CI	Confidence Interval
CIF	Cumulative Incidence Function
CITL	Calibration-in-the-large
DCA	Decision Curve Analysis
DBP	Diastolic Blood Pressure
EFW	Estimated Fetal Weight
EHR	Electronic Health Records
FGR	Fetal Growth Restriction
FP	Fractional Polynomials
GA	Gestational Age
GP	General Practitioner
HR	Hazard Ratio
HTA	Health Technology Assessment
IECV	Internal-External Cross-Validation
IPD	Individual Participant Data
KM	Kaplan-Meier
LP	Linear Predictor
LQ	Lower Quartile
MAR	Missing At Random
MeSH	Medical Subject Headings

MFP	Multivariable Fractional Polynomial
MI	Multiple Imputation
O/E	Observed/Expected ratio
OR	Odds Ratio
PROBAST	Prediction model Risk Of Bias ASsessment Tool
PROGRESS	PROGnosis RESearch Strategy
QUIPS	QUality In Prognosis Studies
RCT	Randomised Controlled Trial
SBP	Systolic Blood Pressure
SD	Standard Deviation
SE	Standard Error
SHR	Sub-distribution Hazard Ratios
SGA	Small for Gestational Age
UQ	Upper Quartile
UK	United Kingdom
WHO	World Health Organisation

CHAPTER 1

Introduction

1 Chapter 1: Introduction

1.1 Key themes of prediction research

Prediction has long been a vital part of research in healthcare and has recently re-emerged as a priority area, with a surge in public and clinician interest in the topic during the COVID-19 pandemic [1]. The increase in popularity has brought many opportunities, along with challenges and research demands.

Prediction research in healthcare refers to the study of clinical outcomes in patients, where the outcome of interest could be the presence of a particular disease or health condition (diagnostic prediction) or the development of some health state in the future (prognostic prediction) [2]. It is a broad research area, covering vital issues such as assessments of baseline risk [3], identification of risk factors [4], development and validation of clinical prediction models [5], and assessment of predictors of treatment effect [6]. The different areas of prediction research are often complementary, for example, with a thorough knowledge of relevant predictors (risk factors) being a vital first-step in developing or improving a clinical prediction model [5].

A major concern in all areas of prediction research is that research quality is often sub-standard; for example, a recent review of COVID-19 prediction models demonstrated that 545 of 606 models were at high risk of bias, with only 5 (0.8%) suitable for potential use in practice [1]. The use of inappropriate or sub-optimal statistical methods is of major concern, leading to a gap between the potential and actual impact of such research on patient outcomes, thus being an important source of research waste. This gap has previously led to the PROGRESS partnership recommending four key themes of prognosis research [3, 4, 5, 6], and subsequent areas for improvement within each.

These concepts extend naturally to the field of diagnostic prediction.

1.1.1 Overall prognosis

Overall, or fundamental, prognosis research covers a variety of important healthcare questions. It involves both describing and explaining clinically important outcomes for people with a particular health condition, in the context of current diagnosis and treatment practices [3]. Such research can supply evidence on how different patterns of diagnosis and treatment can impact future endpoints on a population level, to help inform improvements in the overall quality of healthcare [7]. Assessing overall prognosis in the absence of any clinical care is also important, as it gives vital information towards judging the potential impact of screening programmes for asymptomatic disease, such as breast cancer [8].

An example of an assessment of overall prognosis is seen in a 2020 study summarising changes in years lived with disability for individuals with low back pain (LBP) between 1999 and 2017 [9]. This study concluded that LBP was a leading cause of disability worldwide, and that urgent attention was needed to alleviate the increasing burden and associated impact on health and social care systems, leading to calls for better methods to identify high-risk cases, so that prevention and early intervention could be considered.

1.1.2 Risk factors

A risk factor, or predictor, refers to a characteristic or feature present among a subgroup of the people with a particular health condition, that is predictive of the clinical endpoint of interest [4].

This factor may or may not be causally associated with the outcome, but could separate groups of individuals with different average prognoses to help inform treatment decisions [10]. Such factors may be on the individual level, such as illness perceptions or pain intensity, or at an ecological level (where the exposure of the individual is inferred), including factors such as social deprivation status [11]. Equally, risk factors could be related to treatments received up until the point of prediction, such as previously prescribed medications or surgical features [12]. For example, a 2021 review by Albasri et al. investigated whether antihypertensive treatment was a risk factor for a number of different adverse events, concluding that the current literature showed important associations between antihypertensive medications and the risks of electrolyte abnormalities, acute kidney injury, and syncope [13].

Good quality evidence on the impact of individual risk factors is of great importance, and could be used to refine diagnostic criteria, to inform treatment decisions, or to target preventative interventions [4, 10]. Risk factors also form the building blocks for the development of clinical prediction models, where multiple factors are combined to help predict the value or risk of the clinical outcome for an individual [5].

1.1.3 Prediction models

Clinical prediction models provide individual-level predictions of outcome values or risks to inform patient counselling and facilitate joint clinical decision making [5]. A recent example of a prediction model influencing clinical practice is with the QCOVID model to predict the risk of hospital admission and mortality outcomes from COVID-19 [14]. QCOVID was employed at a national level in the UK, to identify people who may be at high risk of negative outcomes if they were to

catch COVID-19 (assessed using the QCOVID algorithm in their health record data) to prioritise for vaccination and to notify about recommended shielding practices [15].

Prediction model research refers to studies involving the development, validation, and impact assessment of such models [16]. Predicting outcomes on an individual level can allow treatment decisions and monitoring strategies to be tailored to the patient, and their own needs and perceptions, contributing to precision medicine and personalised care [16]. Such outcomes may relate to something current, for example current levels of fat mass in children [17], or in the future, such as the risk of a relapse in depression for those who are currently well [18]. Depending on the context, they may also be referred to as clinical prediction tools, decision tools, diagnostic or prognostic models, risk scores, or prognostic indices, amongst other names, and are in demand in all areas of health, psychology and social care.

Statistical prediction models are typically developed using a multivariable regression framework, which provides an equation to calculate a predicted outcome, conditional on the values of multiple risk factors. Recent years have also seen an increase in the use of approaches attributed to Artificial Intelligence (AI) and Machine Learning (ML), including penalised regression, tree-based methods and deep-learning [19, 20].

1.1.4 Stratified medicine

The final area of prediction research discussed in the PROGRESS framework concerns predictors of treatment effect [6]. Heterogeneity in treatment effects is seen as variation in either the magnitude or direction of the treatment effect across different values for a covariate [21]. Stratified medicine

aims to account for such heterogeneity by tailoring treatment decisions for patients based on their individual-level attributes, to maximise the benefits of treatment while reducing unnecessary costs, both monetary and in terms of harms or potential side-effects from the different treatment paths [22]. Targeting interventions at those most likely to benefit is more important for some treatment-covariate combinations than for others, especially where there is a strong biological rationale for differential treatment effect, and requires good quality evidence of clinically important differences in prognosis across different patient groups [6].

1.2 Need for improved clinical prediction models

A key recommendation of the PROGRESS partnership is for better choice and implementation of statistical methods within prognosis research, especially in regards to both the development and validation of prediction models. Despite this recommendation being published over 10 years ago, recent reviews of prediction modelling studies have concluded that methodological conduct is still poor [20], with a lack of sample size consideration [23] and high risk of bias [1, 24]. Adherence to reporting guidelines has also been shown to be lacking [25], with little improvement after peer review [26], meaning assessment of methodological conduct can be difficult.

With this in mind, this thesis aims to apply and develop high quality methods in the field of prediction modelling. While all areas of prediction research have the potential to offer considerable clinical benefit (or indeed harm, if misused or misunderstood), clinical prediction models in particular offer the opportunity to enhance shared decision making, with increased opportunities for personalised care. Thus, it is vital that high quality methods are used in this area.

This thesis will focus on examples of prediction for clinically important continuous outcomes (such as birthweight, pain intensity, and fat mass), for which the prediction model should ideally give some estimate of the value on its continuous scale. Such models can then be used to predict an individual's expected outcome value (as opposed to the probability of a binary "yes/no" outcome) and provide a basis on which doctors and patients could jointly base their decisions regarding clinical management, if the model was sufficiently accurate. The focus will be how the statistical methods employed while developing and validating clinical prediction models with continuous outcomes could affect the model's performance, and thus it's usefulness or impact as a guide for clinical decision making. Therefore, in the remainder of this Introduction, core methods for prediction modelling with be introduced, and the remainder of the thesis will be signposted.

1.3 Statistical methods for prediction model development

Researchers are faced with many options for modelling approaches when developing a prediction model, many of which are based on a regression framework. Which regression model is most appropriate will generally depend on the format of the outcome to be predicted and the presence of other complexities in data, including incomplete observation of outcomes (censoring) or competing risks. For example, a continuous outcome such as birthweight would be best modelled using a linear regression approach, given all assumptions of the linear regression approach were met. Some of the most commonly used regression models for clinical risk prediction are introduced below.

Non-regression-based methods for prediction modelling are also rising in popularity, including methods attributed to machine learning such as random forests, gradient boosting, and support vector machines (SVM) [19, 20]. Although this thesis is primarily focussed on examples using

regression based methods, findings and recommendations regarding the importance of outcome treatment, sample size requirements, and thorough validation processes extend to all types of prediction models, regardless of underlying modelling approach. The following sections introduce statistical methods for developing prognostic models, in particular for continuous, binary and time-to-event outcomes, all of which will be included in the analyses for the following chapters.

1.3.1 Linear regression for a continuous outcome

When modelling to predict a continuous outcome value, regression analyses are generally based on model equations of the form

$$E(Y_i) = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + ..$$

where α donotes a fixed constant corresponding to the mean outcome value, and the βs are the predictor effects associated with each of the covariates, X_j. The key value of interest for researchers developing a model to predict the continuous outcome, Y_i, is $E(Y_i)$; which gives the expected (or predicted) outcome value for individual i (hereafter referred to as Y_{PREDi}).

While the format of the covariates, X_j, and the methods used to model their predictor effects may vary and thus increase in complexity (as described in the following sections), the standard approach to predicting an outcome on a continuous scale remains unchanged. This is to estimate the parameter values needed to populate a linear regression model in the above form, from data on patients for whom the true (observed) outcome value is already known. Such data is referred to as a model development dataset. Thus the calculated linear regression model would then contain an intercept value (α) and predictor effects (β_1, β_2, β_3, ...), estimated using the observed outcome values, the mean outcome value, and the strengths of the associations between the outcome and

each of the predictor variables (X_{1i}, X_{2i}, X_{3i}, ...) in the model development data.

A simple example of a linear regression model with just one predictor variable would be:

$$Y_{PREDi} = \alpha + \beta_1 X_{1i}$$

Here Y_{PREDi} gives a prediction for the value of the outcome Y_i in individual i. Thus the expected value ($Y_{PRED_{New}}$) of the outcome for a new observation, Y_{New}, can be calculated using the estimated values of the constants α and β_1, and the value of $X_{1_{New}}$ for the new individual.

Given the estimated parameters in the linear regression model are derived from a subset of the population, they are subject to error in their estimation. Equally, not all variation in the outcome will necessarily be explained by the covariates that have been included in the modelling process. Thus, the final prediction for the value of the outcome Y_{PREDi} in an individual i will likely be different from the observed outcome value, even within the development data, thus:

$$Y_i = Y_{PREDi} + e_i$$

where $e_i \sim \mathcal{N}(0, \sigma^2)$.

This normal distribution of the error term is a key assumption of linear regression modelling, being especially important for prediction modelling. Thus, the approach may not be appropriate if the error terms follow a different distribution.

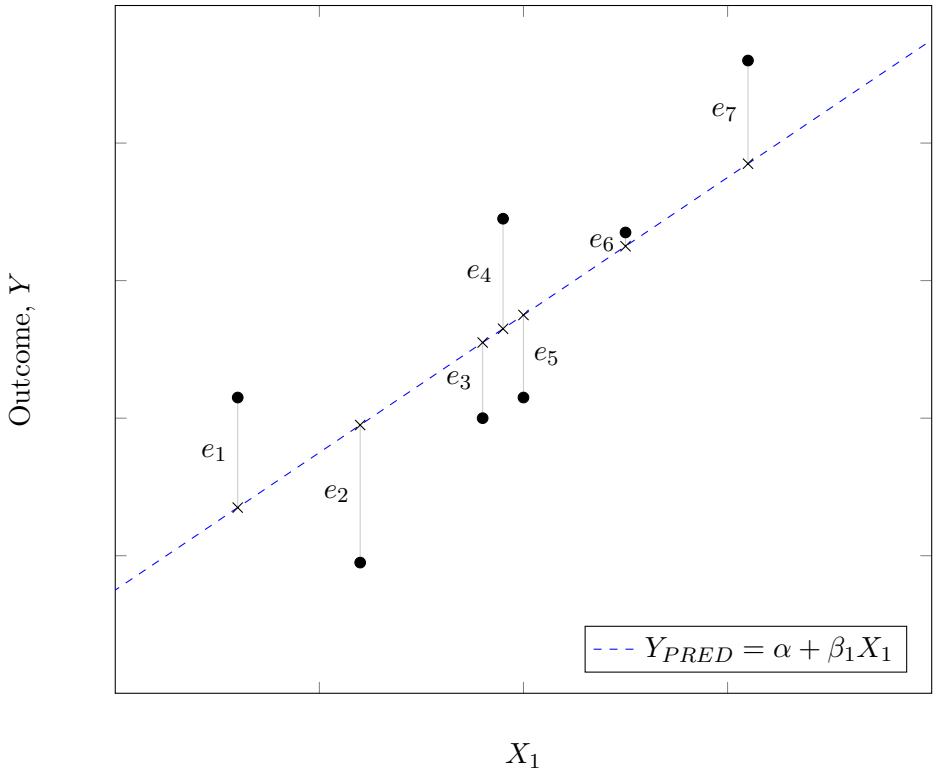


Figure 1.1: Demonstration of the error in estimation from a linear regression model

For a simple linear regression model, it is assumed that the mean value for the outcome, Y_i, is a linear function of the predictor variable X_j, or a combination of multiple X_js. In reality, this linear assumption does not always give a very good approximation of the true relationship, given how complex biological processes and relationships may be. It may be necessary, therefore, to transform the outcome values, the predictor values, or both to a more appropriate scale before this linear assumption is met. Thus the right-hand side of the prediction equation (otherwise known as the linear predictor) can be considerably more complex in reality than is demonstrated above, often with more than just a single predictor (multivariable models), with the potential for non-linear relationships and interactions between predictors once more patient characteristics are included. Linear regression models to predict the continuous outcomes of pain intensity and birthweight are shown in the applied examples of Chapters 3 and 6, respectively.

Example prediction model developed using linear regression

In 2019, Hudda et al. developed a linear regression model to predict fat-free mass on the continuous kilogram (kg) scale, to indicate a child's adiposity and body composition [17]. This clinical prediction model was to assess fat-free mass in children and adolescents, aged 4 to 15 years, using a combination of five risk factors: the child's height, weight, age, sex and ethnicity. In this example, the continuous outcome was transformed to the natural logarithm scale to better meet the assumptions of the linear regression model. The published model equation is as follows:

$$\begin{aligned}\ln \text{fat-free mass} =& 2.8055 + 0.3073(\text{height}^2) - 10.0155(\text{weight}^{-1}) + 0.004571(\text{weight}) \\ &+ 0.01408(\text{if Black ethnicity}) - 0.06509(\text{if South Asian ethnicity}) \\ &- 0.02624(\text{if other Asian ethnicty}) - 0.01745(\text{if other ethnicty}) \\ &- 0.9180(\ln(\text{age})) + 0.6488(\text{age}^{0.5}) + 0.04723(\text{if male})\end{aligned}$$

where predictor variables of Black, South Asian, other Asian, or other ethnic origins are binary, with value of 1 if individual has the particular origin and 0 otherwise. The child's height, weight and age are all continuous predictors, with height measured in metres, weight in kilograms, and age in years.

1.3.2 Logistic regression for a binary outcome

In practice, clinical prediction models are often used to model a binary outcome: an outcome that can take one of only two possible values. Examples include those that are truly binary, including

mortality or live-birth following In Vitro Fertilisation (IVF) treatment, or those that have been formed by dichotomising outcomes that were originally measured on a continuous scale, such as high pain intensity or low birthweight. Where the follow-up is complete (without censoring, see below) and the length of follow-up is consistent among participants, or where the time until the outcome occurs is not of interest, a common approach for modelling a binary outcome is to use a logistic regression.

While the right hand side of the equation (the linear predictor) is of the same general format as with a linear regression model, the dependant variable (the left hand side) in the equation is instead the logit transformation of the event probability for individual i:

$$logit\left(p_i\right) = \alpha + \boldsymbol{\beta X_i}$$

where $p_i = P\left(Y_i = 1\right)$, the probability that $Y_i = 1$ for the binary outcome Y_i, and Y_i can equal either 0 (to indicate no outcome of interest), or 1 (indicating the outcome of interest ocurred). As previously, the estimated model intercept is denoted by α, $\boldsymbol{X_i} = (X_{1i}, X_{2i}, X_{3i}, ...)^T$ is the vector of values of the predictor variables for individual i, and $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, ...)^T$ is the corresponding vector of coefficients for the model, such that

$$\boldsymbol{\beta X_i} = \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + ...$$

The use of the logit function restricts the values of predicted probabilities to the range of 0 to 1, following a sigmoid shape. To obtain predicted probabilities for individual patients, the inverse-logit transformation is applied to the estimated linear predictor value for that patient, giving:

$$p_i = \frac{exp(LP_i)}{1 + exp(LP_i)}$$

where $LP_i = \alpha + \boldsymbol{\beta X_i}$

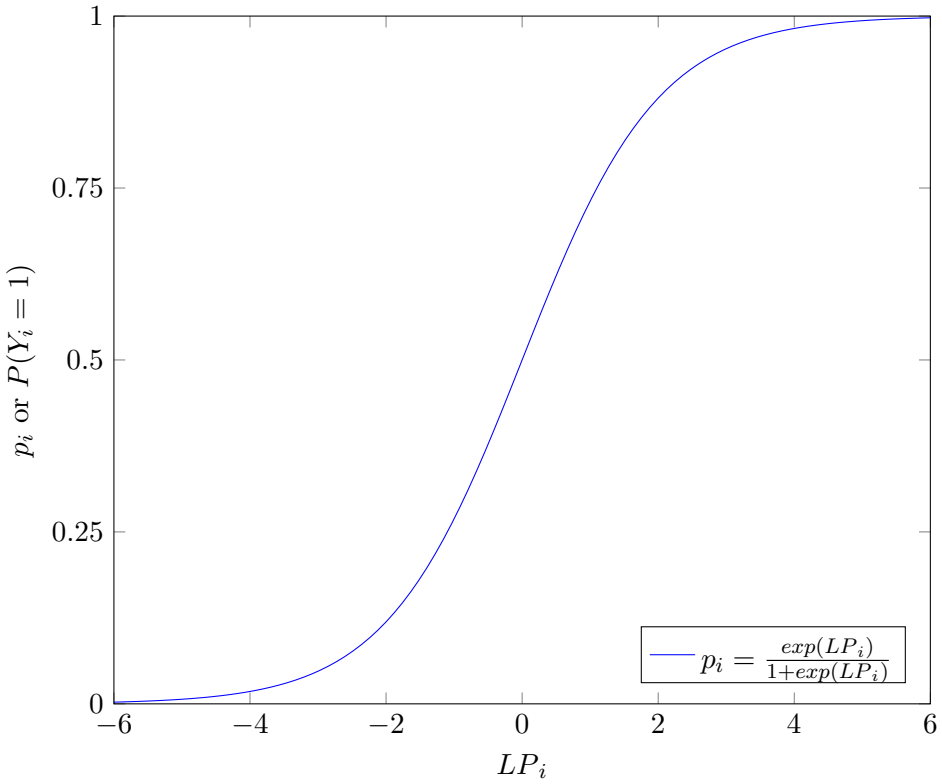


Figure 1.2: The distribution of the outcome probability, p_i, from the range of values of the linear predictor, LP_i.

Example prediction model developed using logistic regression

The MRC CRASH trial used multivariable logistic regression to produce a prediction model to assess the risk of an "unfavourable outcome", defined as death or severe disability, within the six months following traumatic brain injury [27]. Their model was developed on a cohort of 10,008 adults, who were recruited within eight hours of injury and were followed up for six months to establish who experienced the binary outcome of interest. The prediction model incorporated demographic and clinical variables to generate the predicted probability of the "unfavourable outcome", as shown below.

	Regression coefficient	Odds ratio
Age (per 10 year increase)	0.548	1.73
GCS (per one unit decrease)	0.199	1.22
Pupil reactivity		
Both	0	1
One	0.888	2.43
None	1.188	3.28
Major extracranial injury		
No	0	1
Yes	0.482	1.62

Table 1.1: Regression coefficients and odds ratios from an example logistic regression model to predict unfavourable outcomes at 6 months, following traumatic brain injury. GCS = Glasgow coma scale

As with many logistic regression models, the study authors reported Odds Ratios (OR) for the predictor effects instead of the regression coefficients for each variable. The coefficient values can easily be obtained, however, from the reported ORs, as $exp(\beta_j) = OR_j$ for all variables j included in the model (see Table 1.1).

1.3.3 Survival analysis for a time-to-event outcome

Often in prediction modelling, the outcome of interest is not just whether an event occurs or not, as is the case for a logistic regression model, but the time taken until the event takes place. In particular, the time-to-event might be of interest when it is known that an event will not occur for all patients during the study time frame. This may be due to the rarity of an event, or the follow up for a study being too short to observe the event for all participants. For example, when considering

the outcome of live-birth following IVF treatment, some women may never become pregnant. A study would need to follow patients up for many rounds of treatment to get sufficient numbers of live-birth events to reliably build a prognostic model, and many women may not give birth to a live baby within the study time frame.

Censoring

Censoring is a phenomenon often seen in prognostic research, which refers to the situation where the exact time that the event took place is unknown for some participants. This could be due to the study follow up period ending prior to the event taking place for an individual; a separate event occurring prior to the event of interest, preventing it from being able to happen, such as a participant dying in a road traffic accident; or a subject becoming lost to follow up before the research team can observe any event taking place.

There are three main types of censoring (right, left and interval censoring), with right censoring being the most common. Right censoring occurs when the event of interest takes place after the end of follow up (if, indeed, it takes place at all). This means that the true survival time for the participant is larger than what is observed.

$$C_{t_i} < T_i$$

with T_i denoting the true event time for patient i, and C_{t_i} referring to the known censoring time. Prognostic modelling in the presence of left or interval censored data is beyond the scope of this thesis.

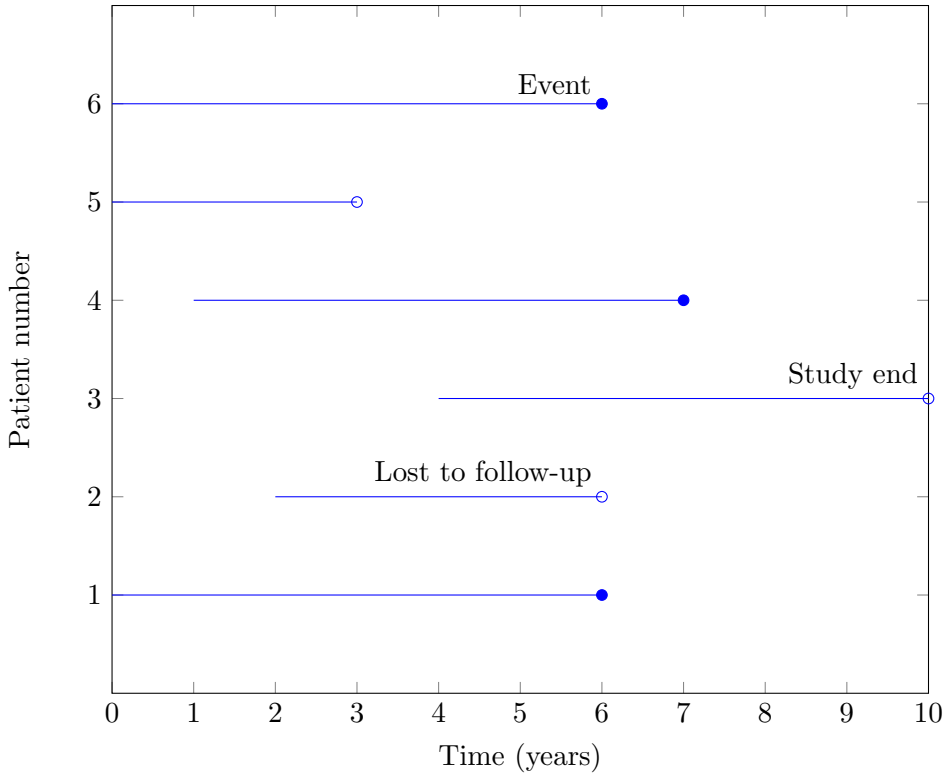


Figure 1.3: Demonstration of right censoring for six patients during study follow up (for a study duration of 10 years), where filled circles indicate the event of interest being observed and hollow circles indicate a right censored observation.

Right censored observations are what make prognostic models for survival data more complex than those with a continuous or binary outcome, as it is highly unlikely that linear or logistic regression models would be suitable. Survival times (the time until the event occurs) are rarely appropriately distributed to meet the requirements of a linear regression model, but even if sufficiently transformed, using a linear regression to predict the continuous time until an event would only be feasible where all participant's event times were known. A linear regression could not model the time to an event that had not yet happened, and thus would be inappropriate for modelling in censored data. Equally, using a logistic regression to model the probability of the event occurring

prior to a particular time point of interest would only be appropriate if all participants who had not experienced the event during follow up were known never to have had the event. Otherwise, those who had not experienced the outcome by the time point of interest would be assumed to be the same as those who would go on never to experience an event: an assumption that is unlikely to hold in practice.

Functions in survival data

Thus, rather than aiming to model the time directly, survival analysis instead focuses on the modelling of important functions of time. The primary function of interest in time-to-event analysis is the survival function, $S(t)$, which defines the probability of an individual surviving up to time t. This can be written as

$$S(t) = P(T > t)$$

where T is the survival time and $0 < t < \infty$. Parametric distributions used to model the underlying shape of the survival function are most commonly of exponential ($\exp(-\lambda t)$) or Weibull ($\exp(-(\lambda t)^k)$) form, and are difficult to estimate in real-life data. In practice, survival functions are often estimated using non-parametric methods, such as using Kaplan-Meier (KM) estimates. These estimates are calculated based on the survival probabilities at each distinct event time: points at which the function drops, creating the step function characteristic of non-parametric survival functions. This is in contrast to the smooth functions achieved when using parametric estimates.

The hazard function, $h(t)$, is also frequently used in survival analysis and refers to the instantaneous failure rate at time t, or the probability that the event occurs in the time interval immediately

following time t given the subject has survived up until that time. It is defined by

$$h(t) = \lim_{\delta \to 0} \frac{P(t \leq T < t + \delta | t \leq T)}{\delta}$$

where δ is some small interval of time, with length tending to zero.

Different shapes of the underlying hazard function (as demonstrated in Figure 1.4) are indicative of different survival distributions over time, and are associated with different types of clinical event or time frame. For example, a hazard function that increases over time implies an increased risk of the event as time passes, such as an increased risk of death in an ageing population. The survival function associated with such a hazard would drop more steeply as time progressed, as a higher proportion of the surviving population died at each time point.

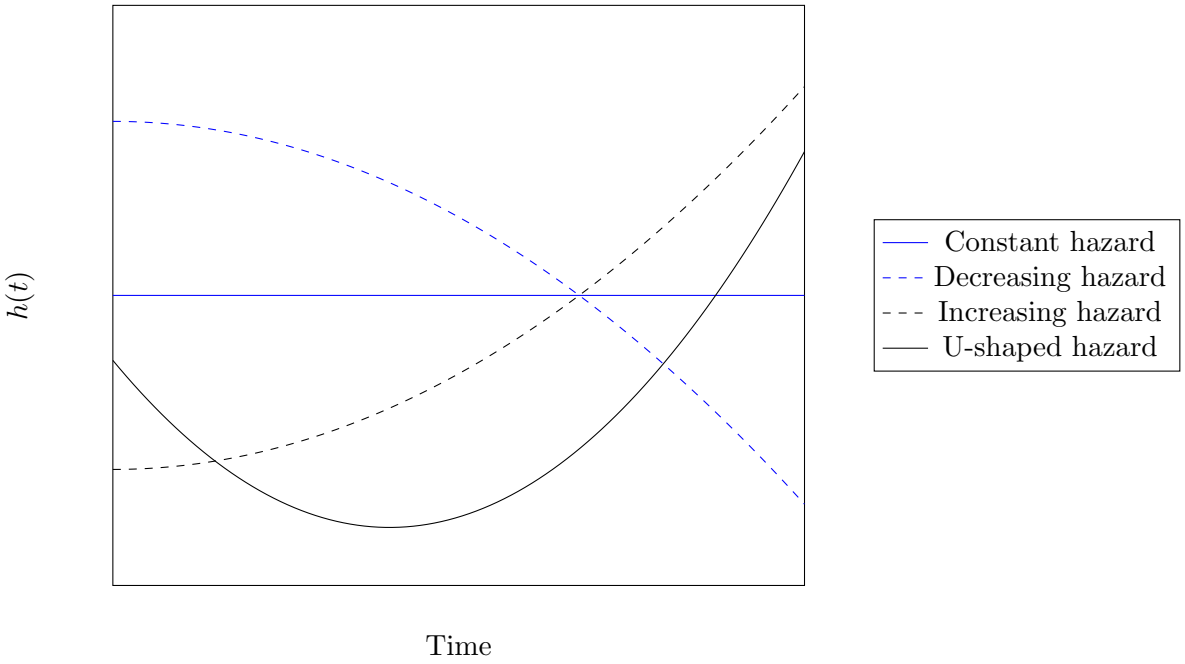


Figure 1.4: Theoretical hazard functions over time, showing how the risk of an event occurring may differ at different times over the course of follow up

Indeed, the hazard and survival functions are closely related and can be obtained from one another mathematically, as

$$S(t) = e^{-\int_0^t h(u)\,\mathrm{d}u}$$

Figure 1.5 shows a simple theoretical survival curve associated with a constant hazard function (the risk of an event does not change over time), with a $h(t) = \frac{1}{6}$ chance of the event occurring at any given time point. The corresponding smooth survival function is $S(t) = e^{-\frac{1}{6}t}$, which is shown alongside non-parametric Kaplan-Meier [28] estimates from simulated data following this hazard distribution, for comparison.

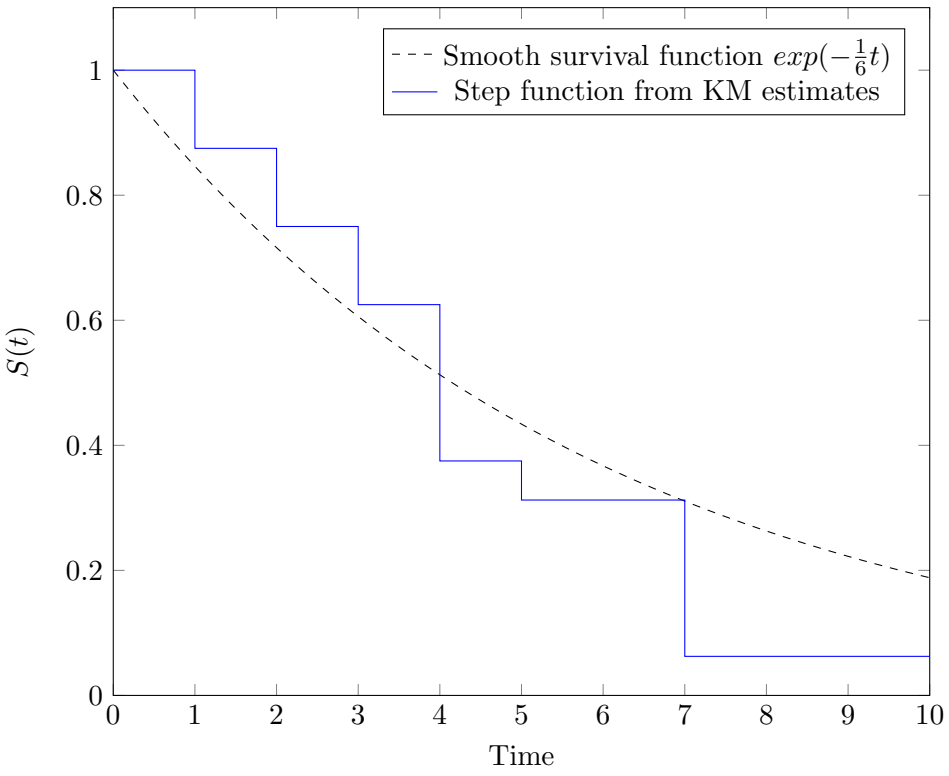


Figure 1.5: Theoretical survival curve compared to Kaplan-Meier estimates for a simple simulated example

Cox proportional hazard model

When comparing survival in two groups of individuals, where the hazards in the groups are proportional (i.e., the difference in hazards between the two groups is constant over time), Cox proposed a regression model can be formulated such that

$$h_1(t) = h_0(t)exp(\beta \mathbf{X})$$

where, $h_1(t)$ and $h_0(t)$ are the hazard functions for the groups, X is a binary indicator for group membership, and $exp(\beta)$ gives an estimate of the relative difference in hazard between the two groups [29].

A fully parametric approach to modelling this relationship would need to specify the distributional form of $h_0(t)$, which can be difficult in practice, just as specifying the form of $S(t)$ is complex. Cox proportional hazards regression offers a semi-parametric approach to modelling survival data, where the baseline hazard ($h_0(t)$) is left unspecified [29, 30]. Alternatively, parametric or flexible parametric approaches might be adopted where some distribution of the baseline hazard of the event over time can be assumed [31, 32]. Flexible parametric approaches to modelling survival are beyond the scope of this thesis.

Competing risks

In standard survival analysis, for example when using the Cox proportional hazard model (described above), patients are assumed to experience only one type of event. In practice, follow up may end with the occurrence of one of many possible events, some of which preclude the event of interest. When some alternative event could occur prior to the event of interest and, in doing so, prevent the event of interest from happening, it is known as a competing risk.

For example, when a patient dies during follow up from a cause unrelated to the event of interest, the event of interest has not occurred, nor can it ever take place. This is different to ordinary right censoring, where it is still possible that an event could occur (unobserved) after follow up has ended. The concept of competing risks is highly relevant in prognostic modelling with a long prediction-horizon, for example in older populations where death from unrelated causes is more likely, as in the applied example in Chapter 7. Further discussion of methods to account for competing risks in survival analysis is included in Chapter 7.

1.4 Modelling complex predictors

1.4.1 Non-linear trends in covariates

Continuous variables such as age, weight or blood pressure, are regularly included as covariates when developing a prognostic model, regardless of outcome type. While standard regression approaches would assume a linear relationship (discussed above, in Section 1.3.1), the observed relationship between these continuous predictors and the outcome of interest is often in fact non-linear [33]. There are many options for addressing the non-linearity of predictor-outcome associations, for example a simple transformation of the predictor may be sufficient to achieve linearity. A common approach to combat non-linearity in practice is to categorise continuous measurements into two or more categories and to model the predictor as a categorical measurement. This approach is widely discouraged as it is inefficient and often biologically implausible [34], thus alternative methods for modelling non-linear trends using transformations, fractional polynomials or restricted cubic splines are preferred [33, 35, 36].

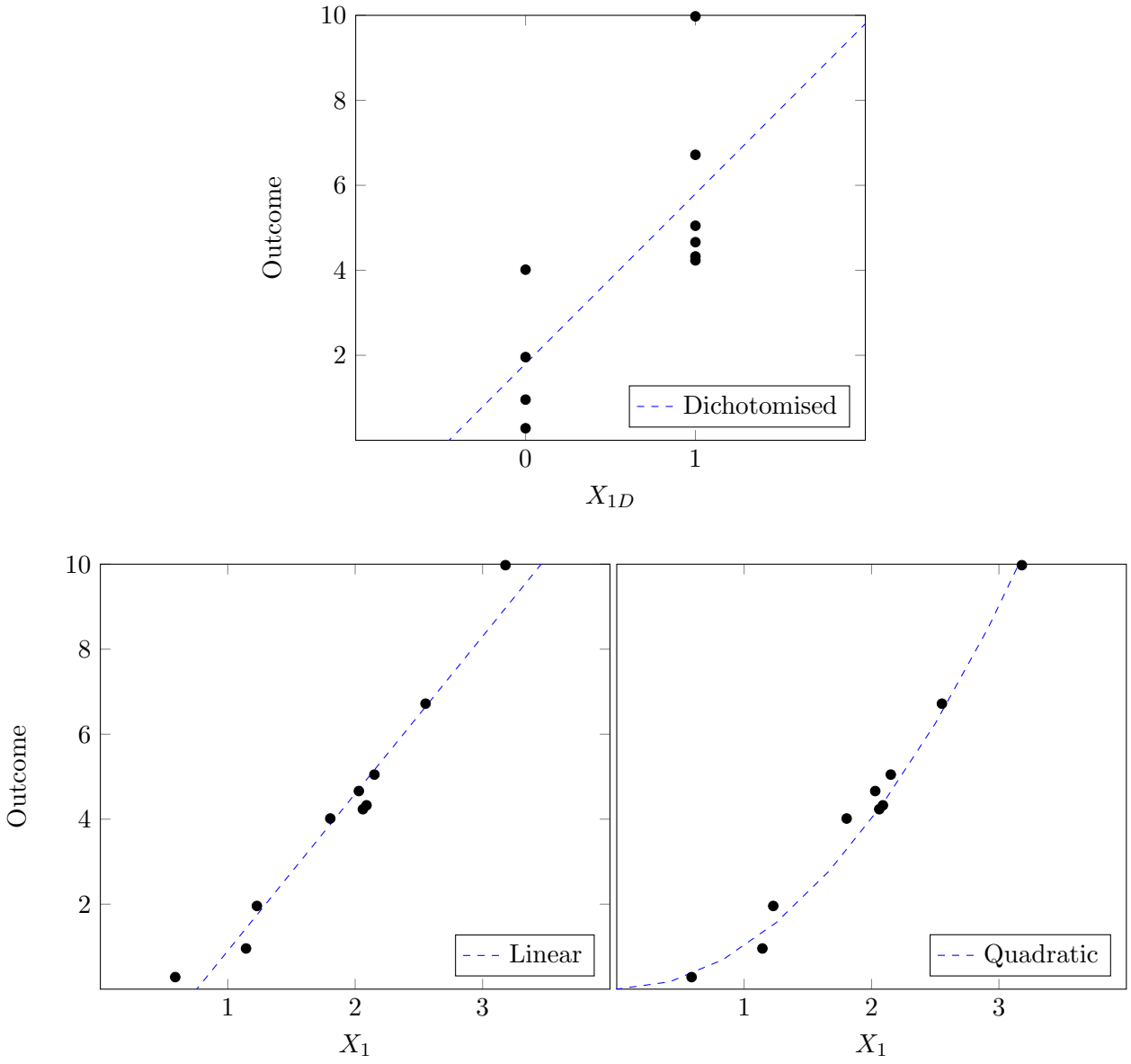


Figure 1.6: A visual comparison of modelling a continuous covariate in its dichotomised, linear and quadratic forms

Fractional polynomials

Relationships between continuous covariates and the outcome may be complicated, and are often of an unknown form in experimental data. A key feature of these underlying relationships is that they are smooth, or are subject to very little noise, meaning that they can be well represented by a

functional approximation. Fractional polynomials in regression modelling allow for far more flexible curve shapes in the modelling of non-linear relationships, over simple non-linear transformations of the covariate [35, 37, 38, 39]. This is achieved by using a combination of positive, negative and non-integer powers, with continuous predictors transformed using up to m different powers from a predefined set:

$$\{-2, -1, -0.5, 0, 0.5, 1, 2, ..., max(3, m)\}$$

where the power 0 refers to the natural logarithm, such that ${X_1}^{(0)} = \ln X_1$.

In the field of medical statistics, it is rare that the true underlying relationship between a continuous predictor and the outcome will require more than three fractional polynomial terms, thus the maximum power included in the predefined set is three [35]. As such, possible fractional polynomial transformations for a continuous predictor X_1 include combinations of the following:

$$\{\frac{1}{{X_1}^2}, \frac{1}{X_1}, \frac{1}{\sqrt{X_1}}, \ln X_1, \sqrt{X_1}, X_1, {X_1}^2, {X_1}^3\}$$

Restricted cubic splines

Cubic splines can be used to generate flexible functions that are able to better fit more complicated curved relationships between a predictor and the outcome [30, 36]. To achieve this, a series of cubic functions are fit along the range of the predictor, joined at points referred to as "knots". To ensure that the overall function is smooth over the full range of predictor values, cubic functions are required to join smoothly at each of the knot positions. This is achieved by specifying that the first and second derivatives of the functions are equal at the knot points. To ensure the cubic splines remain stable at the extremes, where there is often limited data with

which to derive estimates, the functions are also constrained to be linear in the tails, hence the term "restricted".

While increased complexity in the modelling of continuous predictors can vastly improve model fit in the development data, these modelling approaches can also increase the chance of overfitting to the development data, reducing the generalisability of the model to new data (a vital trait for a prediction model to be used in new patients). This overfitting occurs when the prediction model includes some modelling of the noise within the model development data. For example, a curved predictor-outcome relationship may be included in the model that is, in fact, more complex than the true underlying relationship, as the model also includes the shape of chance variations in the predictor (or outcome) measurement. Thus choosing the level of complexity required when modelling non-linear trends involves a fine balance between allowing just enough complexity to match the underlying relationship, without also inadvertently modelling noise.

1.4.2 Interactions

Where there are multiple covariates in a model, the effect of a combination of two or more of these covariates, above and beyond what would be expected for each of the factors alone, is known as an interaction effect [30]. A two-way interaction is said to be present where the change in the mean outcome value for two levels of a covariate, X_1, is different for different levels or values of a second covariate, X_2. This notion extends to three-way interactions and beyond, although higher order interactions may require very large amounts of data to assess. Adding interaction terms to a multivariable regression model can help to account for the complex, real-world relationships that exist between predictor variables and allows for more accurate modelling of the outcome.

Of particular interest in prediction research is the way that treatments may interact with patient demographics, such that a patient's traits might be expected to impact the effectiveness of a treatment or intervention (also known as "treatment-covariate interactions") [40, 41, 42]. Research involving such interactions may inform the targeted use of certain treatments in particular patient groups, allowing tailored treatment for maximum patient benefit. Accounting for such interactions in a prognostic model allows for much more accurate risk prediction than when only including the treatment and demographic predictors independently, without also allowing for the additional prognostic effect of their interaction.

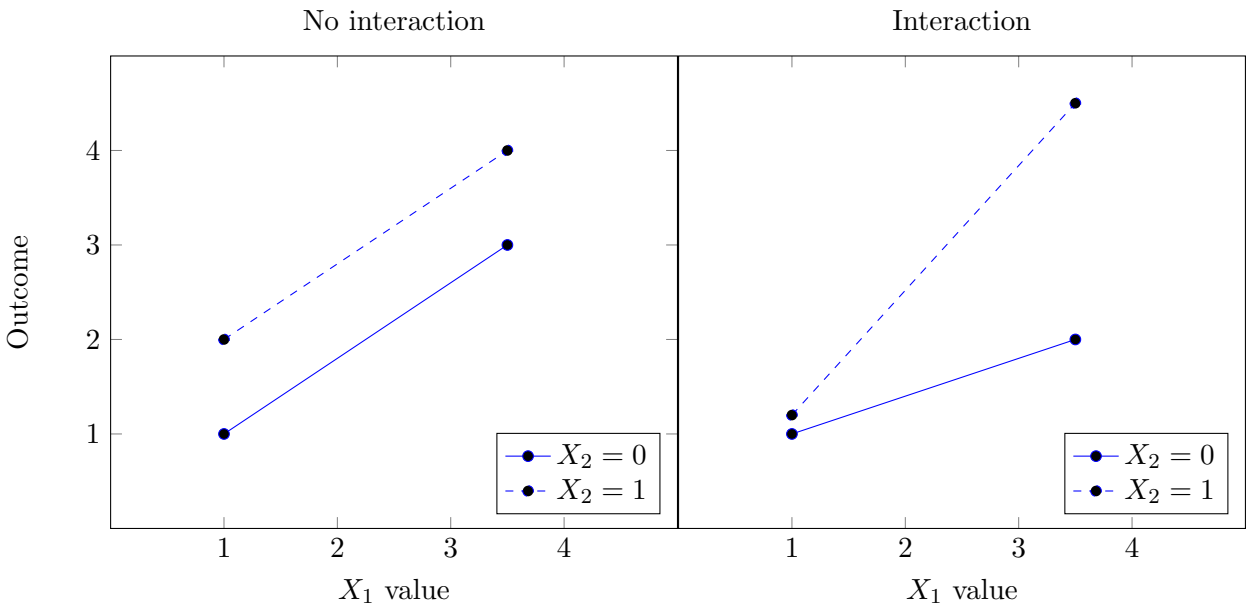


Figure 1.7: Demonstration of an interaction between a continuous (X_1) and a binary (X_2) covariate, in comparison to the case where no interaction is present

1.4.3 Time-varying predictor variables

While some predictor values, and their association with the outcome, are assumed to remain constant over time, for others this assumption is unlikely to hold in practice. Often prediction is

based on patient characteristics that are known at a single baseline time point, whereas patients will often have repeated contact with healthcare services. For example, they may then have multiple measurements of features such as biomarker levels in the blood, differing treatments received over time, or clinical histories of relapse/recurrence events. Predictions including time-varying predictors should include the measurement recorded at the time of intended model application [43].

Where changes in a predictor's value over time are expected to be informative, these changes can be incorporated into the prediction model in a number of different ways [44]. Calculating predictions to account for new or changing predictor information over time include more complicated analysis techniques, and require rich sources of data with detailed histories of changes in predictor values [45]. Further details on the consideration of time-varying predictors is included in the discussion of Chapter 6.

1.5 Assessing predictive performance

1.5.1 Internal validation

Once a model has been developed, its predictive performance must be thoroughly evaluated to ensure that it is fit for purpose, prior to recommending it for use in clinical practice. Internal validation processes aim to quantify a model's predictive performance in data from the same underlying population as was used to develop it [46, 47]. A model which performs poorly in its own development data is unlikely to perform well externally (in new patients), and thus internal validation is an important step in a model's validation process, to avoid research waste at the external validation or impact assessment stages.

The "apparent" performance of a prediction model refers to how well the model performs in the same data as was used in its development. The performance of a model it its own development data will usually be better than its performance in any other dataset, even where the new data contains patients from an identical population. As such, any performance measure calculated for a model in the development data is likely to be optimistic, giving an unrealistically high expectation of the model's accuracy [48]. The difference between apparent performance in the model development data and performance in new individuals is expected to be smaller where extremely large sample sizes are used in model development, or when adjustments for overfitting were made during the modelling process, as with penalised regression approaches [49, 50].

It is common to see researchers opting to randomly split their data into a model development (or "training") dataset, and a hold-out, unseen validation (or "test") dataset, where the latter is used to estimate the performance of the model developed in the former [51, 47]. In general, this approach is known to be inefficient and sub-optimal use of data, resulting in a smaller sample size and thus less information for model development [52, 53]. Given the random split is expected to result in two samples from identical populations, the model performance in this hold-out sample is also expected to be, on average, the same as in the development sample, and so does not give a good representation of how generalisable the model is to a new setting [51]. A better approach is to use all data for model development, with some form of internal validation to assess the level of optimism in predictive performance present in the apparent validation statistics [51, 52].

Such internal validation methods can give an indication of the extent to which a model performs better in its own development data than it would be expected to perform in new patients. This

knowledge allows for calculation of optimism-adjusted performance estimates to give a more realistic indication of how well the model might perform in new data [48]. Internal validation also allows for assessment and correction for overfitting of a model to the development data, which would likely lead to poor external performance of a model if uncorrected by appropriate shrinkage [54, 55, 50]. Recommended methods of internal validation include the following [30]:

Cross-validation

During the internal validation stage, data are split into two sub-samples, similar to the split-sample method mentioned above. A small portion of the data is reserved for the validation of an example model developed on the remaining data, using the same modelling procedures as were used in the full model development. Following development of this example model in the "training" data, the example model's performance can be estimated in this remaining validation sample (the "test" data), to give an indication of how the model might perform externally, in new data from the same underlying population. This splitting process should be repeated numerous times (for example, with 10-fold cross validation, splitting the data 10 times), with a different model being developed and validated for each data split. Performance across all data splits is then summarised, providing an estimate of how well the final model, developed on the full development data (without splitting) would perform externally [30, 56].

Bootstrapping

A new sample of participants is obtained by sampling, with replacement, from the original dataset, giving a sample of the same size as the original development data [30]. This new sample may contain some of the original participants multiple times, while other participants are omitted altogether, thus this new sample is fundamentally comprised of individuals representing

the underlying development population. The full model development process is repeated within each bootstrap sample, including all modelling procedures and decision making processes, such as multiple imputation, variable selection, and assessment of non-linear trends [48]. Predictive performance for this new model is then tested in both the bootstrap sample and in the original dataset. Performance in the two are compared to give an estimate of the optimism for each performance statistic for that bootstrap model. This process is repeated multiple times, with the average optimism for each performance statistic being taken across all bootstrap samples to give an estimate of the optimism expected to have been present in the assessment of the original model. Performance statistic values for the original model can then be adjusted for this optimism, to better reflect the expected performance in external data.

1.5.2 External validation

Prior to implementation, it is important to evaluate a model's predictive performance in new data, independent to that used to develop the model [57, 58, 59]. This process is known as external validation and is essential to the uptake of a model in practice, especially where the sample size for model development was small or perhaps unrepresentative of the general population of interest [58, 60]. Specifically, external validation indicates how the model performs in new data representative of the population to which the model is intended to be applied in practice. This new data may be from a population that is very similar to the model development, or equally could stem from an entirely new population where the model's expected performance is unknown [59]. External validation can be used to assess two different attributes of the model [61, 46], depending on which of these types of population is used for the validation:

Reproducibility refers to whether a model performs sufficiently well in new individuals from the same target population. If the observed relationships between the model predictors and the outcome of interest were true (rather than being observed by chance in the model development data), the same relationships would be present in other individuals from that population, and thus the model would be expected to perform well across new samples. Reproducibility can also be assessed through resampling techniques in the development data, such as internal validation via bootstrapping [61].

Transportability refers to whether the performance of the model is consistent in new samples from a different but related population [62]. For example, a model may have been developed for a particular outcome in an adult only population, while researchers are interested in whether the model can be used to identify that same outcome in children. Model transportability can only be assessed with external validation [61].

External validation involves applying the developed model to every individual in the new dataset, to generate a predicted outcome value or risk for everyone [59]. For this, the external data must include all predictor variables needed to calculate predictions as well as a measurement of the outcome of interest. The equation of the prediction model is also required in full, including all coefficients (or predictor effects) as well as an estimate of baseline risk in the development population (the model intercept or baseline hazard, depending on model type).

Preferably a model would be assessed in multiple datasets, to give some idea of how well it performs in different settings or patient groups. Such new settings could include different geographical areas or different medical settings. Often electronic health records (EHR) will include patient

information from multiple centres (for example different General Practitioner (GP) practices), giving the opportunity to assess how well a model performs in multiple locations, with different case-mixes, from just one data source [63, 56].

Where multiple external validation assessments have been conducted, meta-analysis methods can be used to estimate average model performance [64, 65, 66]. Assessing model performance across different datasets or populations in this way gives an indication of the spread of model performance over different patient groups, which is of particular importance when individual external validation populations are small [67]. What constitutes an adequate sample size for external validation of a prediction model is discussed further in Chapter 4.

1.5.3 Calibration

The calibration of a prediction model refers to the agreement between the predicted value or risk of the outcome (Y_{PREDi}, p_i) and the observed outcome value (Y_i) across individuals. The model's calibration performance should be estimated during both internal and external validation of a prediction model and is most clearly demonstrated graphically through a calibration plot, where the predicted outcomes are plotted on the horizontal axis, against observed outcomes on the vertical axis, as shown in Figure 1.8. A smoothed calibration curve can be fitted through all of the data-points and presented on the plot to give a view of calibration performance across the full range of the data.

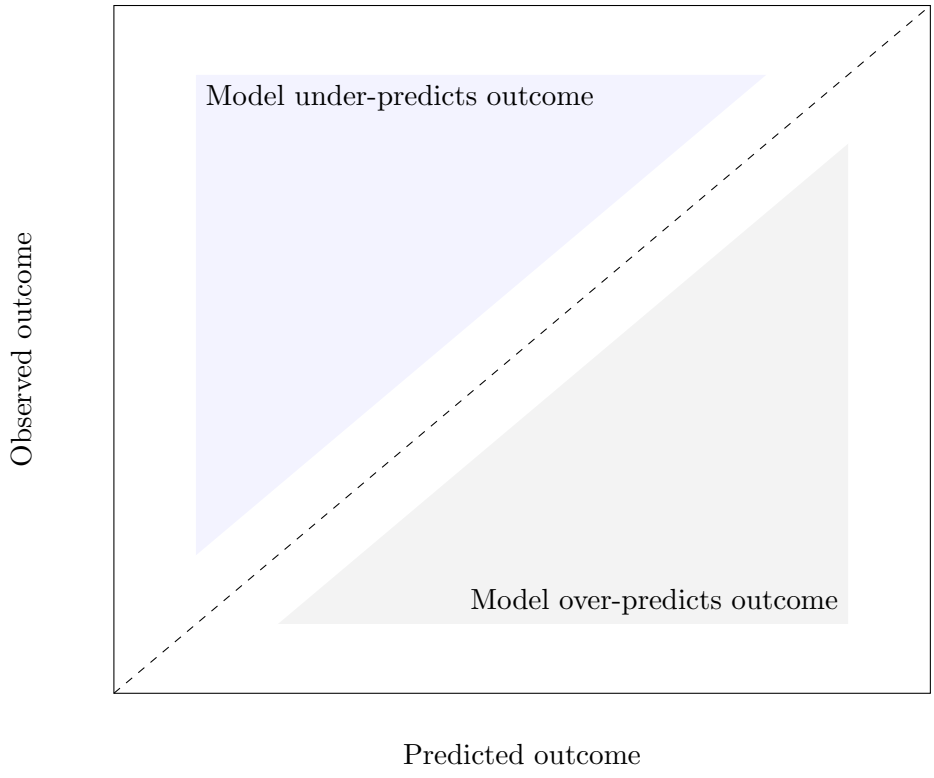


Figure 1.8: Generic calibration plot format, showing predicted against observed outcome (values or probabilities). Ideal calibration is indicated by the 45° reference line where predicted outcomes exactly equal observed outcomes.

Ideally, predicted outcomes should show close agreement with observed outcomes for all individuals, across the full range of predicted values, and should not be systematically over- or under-estimated. Where a specific range of outcome values are of particular clinical importance, good calibration in that range alone may be sufficient for the model to supply clinical benefit, even if some miscalibration is evident elsewhere. For example, a model to predict birthweight might be intended to identify any abnormal fetal growth, in which case the model calibration in the extremes (for those with particularly high or particularly low predicted birthweight) would be of most clinical importance, and miscalibration in these regions of the plot would be of most concern.

Calibration slope

To quantify the calibration performance of a prediction model, a calibration model should be fitted to the validation data (whether this be a sample of the development data during internal validation, or an independent dataset during external validation) using standard estimation methods for the appropriate regression model, for example using restricted maximum likelihood estimation to perform a linear regression, where the outcome is continuous. This calibration model should be of the form:

$$Y_i = \alpha_{cal} + \lambda_{cal}(Y_{PREDi}) + e_{cali}$$

$$e_{cali} \sim \mathcal{N}(0, {\sigma^2}_{cal})$$

The parameter λ_{cal} represents the calibration slope, which measures the agreement between the predicted and observed outcome values across the whole range of predicted values. This value can be interpreted in the same way as any other regression coefficient, namely it is the increase in the observed outcome value, for each one unit increase in the predicted outcome. Thus, the ideal value for λ_{cal} is one. A λ_{cal} below one indicates that predictions are too extreme (predictions above the mean are too high, while predictions below the mean are too low) and a λ_{cal} greater than one suggests that the range of predictions is too narrow (predictions above the mean are too low, while predictions below the mean are too high). The term ${\sigma^2}_{cal}$ gives the residual variance of the calibration model.

Calibration-in-the-large

Systematic over- or under-prediction of a model is possible even when the calibration slope is perfect (equal to one). Therefore, the calibration slope should always be considered alongside calibration plots and calibration-in-the-large (CITL). The CITL measures the agreement between the predicted and observed outcomes on average. For example, with a continuous outcome, this is

a comparison of the mean predicted outcome value ($\bar{Y}_{PRED}$) and the mean observed outcome value ($\bar{Y}$), and can be measured in the validation data simply by calculating

$$CITL_{val} = \bar{Y} - \bar{Y}_{PRED}$$

More generally, CITL is equivalent to estimating the intercept term, α_{cal}, in a calibration model where the slope, λ_{cal}, is forced to equal one. The ideal value for $CITL_{val}$ is therefore zero, where the mean predicted and observed outcomes are exactly equal.

Ratio of Observed to Expected values

An alternative measure of calibration-in-the-large is the Observed to Expected ratio, O/E, which is the ratio of the total observed outcomes (O) in the data, against the total predicted outcomes when using the prediction model (E). For a binary outcome, this is equivalent to the observed risk of having the outcome in the validation data (O/N) divided by the average predicted risk from the model ($\frac{\sum_{i=1}^{N} p_i}{N}$). The ideal value of O/E is therefore also one, with values greater than one implying that observed outcomes are bigger than expected, so the model is under-predicting outcome probabilities, while values less than one indicate that observed outcomes are smaller than expected, so the model is over-predicting on average.

1.5.4 Discrimination

A prediction model for some event of interest is said to discriminate well if it separates well between those who go on to have the event and those who do not. To achieve such separation, a model should assign higher predicted risks to individuals who go on to experience the event. A popular measure of discrimination for prediction models is the Concordance (C)-statistic.

C-statistic

The C-statistic (or C-index) can be interpreted as the proportion of concordant pairs out of the total number of possible patient pairings where one went on to experience the event and the other did not [30]. A patient pair is described as concordant if the patient experiencing the event has the higher predicted probability of the two. For a binary outcome, the C-statistic is the equivalent to the area under the Receiver Operating Characteristic (ROC) curve [68]. The maximum value of one indicates perfect discriminative ability (for all relevant pairings, the patient with the outcome had the higher predicted probability from the model), while a value of 0.5 indicates the model works no better than chance.

1.5.5 Clinical utility

While measures of calibration and discrimination are vital in understanding how well the model predicts, they give no indication of the likely impact of using the model in clinical practice, in terms of the proportions of patients expected to directly benefit from the model's use. It is possible that a model could show excellent calibration and discrimination performance, but still offer no clinical benefit above the current treatment strategy. Equally, a model that shows relatively poor calibration or discrimination could still be clinically useful. When the intended use of a prediction model is to complement clinical decision making, the model should also be assessed for the overall clinical consequences of its use, for patients and for healthcare services [59]. This is known as the model's clinical utility, and can be quantified using net benefit, among other measures.

Net benefit

The Net Benefit of a prediction model gives a measure of the benefits arising from the model's

use (such as improved patient outcomes), weighed against potential harms (for example, a risk of adverse reactions) [69, 70, 71]. This comes from the difference between the number of true-positive (TP) and false-positive (FP) results that arise from using the model to make treatment allocations, where the latter is weighted by a factor representing the "cost" of a false-positive result relative to a true-positive. This weighting (known as the "exchange rate") is based on the threshold probability used to assign treatment, which reflects a clinician's willingness to accept a certain number of false-positives for each true-positive the model identifies.

$$NB_{p_t} = \frac{TP}{n} - \frac{FP}{n}\left(\frac{p_t}{1-p_t}\right)$$

where p_t is the threshold probability from the model, used to indicate a change in treatment pathway, NB_{p_t} is the net benefit at a given p_t, and n is the total sample size.

The net benefit across a range of different threshold probabilities is often visualised through a decision curve analysis (DCA), where the net benefit associated with using the model is plotted across a range of clinically relevant threshold probabilities and compared to the net benefit that would arise from other treatment strategies. These alternative strategies may include other prediction models, or approaches at the extremes, where either everyone or no one in the population is treated as though they were at high risk of the outcome (known as "treat all" and "treat none" scenarios respectively, see Figure 1.9).

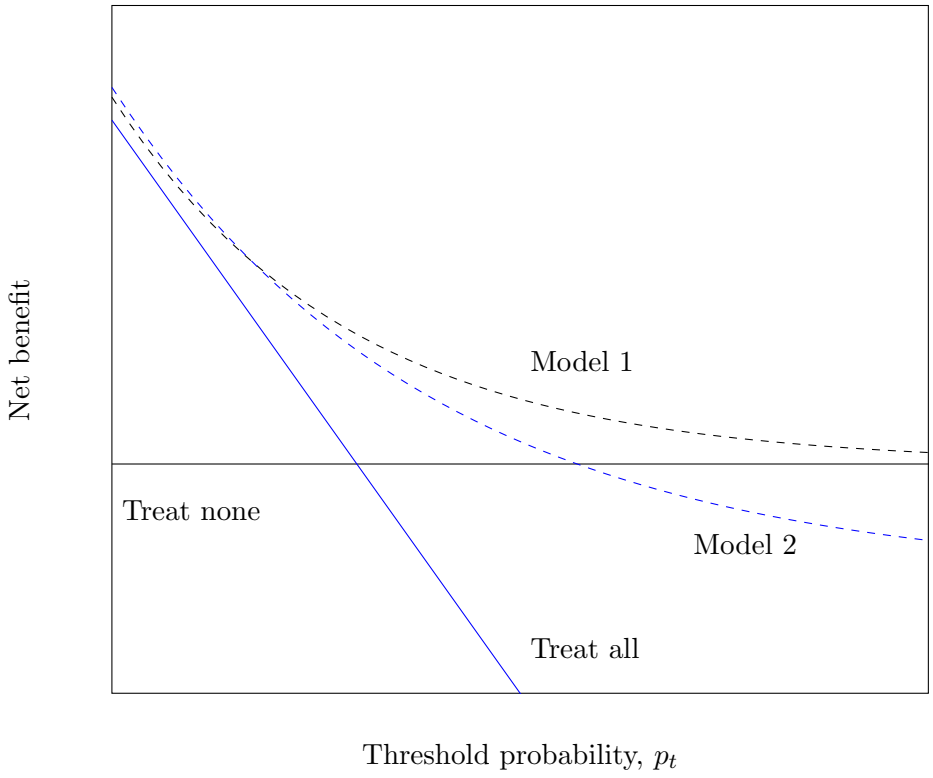


Figure 1.9: Decision curve plotting net benefit against threshold probability for two competing prediction models. Both models show net benefit over-and-above the "treat all" strategy over the full range of threshold probabilities, while Model 2 becomes less favourable than the "treat none" strategy for higher values of p_t

Regarding the "treat none" approach, all patients are treated as though they were at low risk of the outcome, so none are given the treatment reserved for high risk cases. All patients are assumed to be "negative", thus the numbers of true-positives and false-positives are both equal to zero. Thus the net benefit of the "treat none" approach is equal to zero across the full range of probability thresholds, regardless of the weighting attributed to a false-positive (the exchange rate).

When "treating all", every patient is treated as though they were at high risk, thus all truely "positive" patients are correctly treated per this strategy, and so are a true-positive in terms of net

benefit calculation. Every other patient (those who are truly "negative") is also treated as though they are at high risk and so are then included as false-positives, thus $FP = n - TP$. For fair comparison to the model, these false-positives are weighted by the exchange rate, as above, thus the "treat all" strategy only supplies a positive net benefit where the threshold probability (p_t) is lower than the true outcome prevalence in the sample ($\frac{TP}{n}$):

$$\begin{aligned}
NB_{p_t} &> 0 \\
\frac{TP}{n} - \frac{FP}{n}\left(\frac{p_t}{1-p_t}\right) &> 0 \\
\frac{TP}{n} &> \frac{FP}{n}\left(\frac{p_t}{1-p_t}\right) \\
TP &> (n - TP)\left(\frac{p_t}{1-p_t}\right) \\
TP\left(1 + \frac{p_t}{1-p_t}\right) &> n\left(\frac{p_t}{1-p_t}\right) \\
TP\left(\frac{1}{1-p_t}\right) &> n\left(\frac{p_t}{1-p_t}\right) \\
\frac{TP}{n} &> p_t
\end{aligned}$$

A higher net benefit indicates a higher level of benefit from the treatment decision approach, net of harm from false-positives. Therefore, a model should ideally show net benefit over and above all other strategies in the range of threshold probabilities that are most clinically relevant, with these clinically relevant thresholds chosen *a priori* through consultation with patients, clinicians, and other stakeholders.

1.6 Prediction modelling research involving meta-analysis

The availability of multiple datasets from different studies, brings many opportunities in prediction modelling research. Combining patient-level data across multiple studies can allow for model development and assessment of model performance in a wide variety of locations and settings, and

allow for larger sample sizes for analyses [64, 67].

As mentioned previously, it is widely recommended that a prediction model should be externally validated to assess its predictive performance. While one external validation gives useful information on out-of-sample performance, multiple external validations across many studies can give an estimate of the range of likely model performance across different populations [63, 56]. Model calibration and discrimination, in particular, are highly dependant on the patient spectrum, and as such their estimates are likely to vary across different validation populations. Where Individual Participant Data (IPD) is available from multiple studies for model validations, meta-analysis of calibration and discrimination [65, 72], or clinical utility measures [73] can be used to summarise overall performance, giving estimates of heterogeneity in model performance across multiple settings and possibly identifying populations or sub-populations where the model performs inadequately [63].

All applied chapters in this thesis incorporate IPD meta-analysis techniques in their methods, as all examples include clustered data of some form. These include examples where data from multiple studies was combined for model development (Chapter 3), or external validation (Chapter 5), and use of Electronic Health Records (EHR), where data are naturally clustered by GP practice (Chapter 6).

1.7 Current challenges and limitations in prediction modelling research

A recent flurry of systematic reviews into the methodological practices and reporting of prediction modelling studies has shown little improvement in quality since the publication of the PROGRESS

recommendations in 2013. Many prediction modelling studies still suffer from shortcomings in their statistical analyses, resulting in models that are at a high risk of bias [1, 74, 20]. In particular, factors contributing to risk of bias include small study size, poor handling of missing data, and failure to deal with overfitting [74, 1]. Indeed, sample size is still rarely considered in practice, with little justification of the sample size used for developing a clinical prediction model [23, 24] despite widely available guidance [75, 76, 77, 78].

Model evaluation practices are also poor, often with an inappropriate focus on discrimination performance alone and little consideration of model calibration [1], despite model calibration's direct relevance for shared decision-making and patient counselling [79, 80]. For example, a review of prediction modelling studies in oncology found that 78% of papers used only discrimination measures to make comparisons between developed models (32/36 studies reporting comparisons in their discussion) [81].

In addition to these poor methodological practices, findings from such studies are vulnerable to overinterpretation [82, 81], with authors often making unfounded recommendations for a model's use. For example, a 2023 review of prediction model studies in oncology found that more than half (74/133, 55.6%) of included studies made recommendations for clinical use without any external validation [81]. Similarly, a 2023 review of prediction models using supervised machine learning techniques found that 95.2% (20/21) abstracts that recommended model use in daily practice did so without external validation of the developed models [82]. Given the regularity of this issue in the prediction modelling literature, and the concern of resultant harm to patients if poor models were implemented, recent guidelines have been published to help identify and evaluate spin practices in studies on prediction models [83].

Guidelines for the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) [84, 85] were introduced in 2015 to promote the complete, accurate, and transparent reporting of studies that develop a prediction model or evaluate its performance. Recent extensions have since been published, with tailoring specific to clustered data [86], systematic reviews and meta-analysis [87], and machine learning (or regression-based) methods [88], the latter of which supersedes the 2015 checklist. Despite this, both the use of and adherence to reporting guidelines has also been shown to be lacking [25, 26, 89, 90], hindering assessment of methodological quality.

Overall, there is still an urgent need to improve the quality of conduct and reporting of prediction modelling research.

1.8 Aims and overview of thesis

The broad objective of this thesis is to apply, evaluate and extend statistical methods for research involving clinical prediction models, particularly in the context of primary care applications. Specific aims are to:

1. identify methodological shortcomings in existing prediction research involving diagnostic and prognostic models of continuous outcomes;
2. ensure robust methods are used in current prediction-based projects funded at Keele University, including primary studies and Individual Participant Data meta-analyses;
3. extend existing statistical approaches for prediction research that could be used by other researchers in the prediction modelling field, to improve the quality of their research.

This thesis contains seven chapters, exploring a range of methodological issues and demonstrating statistical techniques through various clinical applied examples. An outline of the chapters is given below.

Chapter 2 contains a methodology review of recently published models for predicting Fetal Growth Restriction (FGR), either through prediction of birthweight on its continuous scale or through prediction of dichotomised birthweight, with or without the inclusion of specific birth complications (such as stillbirth, or neonatal death). The main aims of this review were to identify (a) how FGR and birthweight are being modelled in the current medical literature, (b) the justification given for any outcome dichotomisation in this setting, and (c) whether the treatment of this continuous outcome is recognised by study authors as a strength or a limitation of the model building process.

Chapter 3 develops illustrative prediction models for pain outcomes in individuals consulting their GP with non-specific Neck and/or Low Back Pain (NLBP). Pain score outcomes at 6-months were modelled as both continuous and when dichotomised (to define 'poor' prognosis), with models then compared for predictive accuracy on internal and external validation. This chapter further demonstrates how to use the model with continuous pain to make subsequent predictions of the probability of high pain, and compares these to predictions from the model where high pain was modelled directly. The applied research described in this chapter has been published in *Physical Therapy* [91].

Chapter 4 develops and proposes closed-form solutions to calculate a sample size that ensures sufficient precision around key performance statistics when externally validating a clinical prediction model with a continuous outcome. The approach is demonstrated in a case-study predicting fat-free mass in children and adolescents. The methods proposed in this chapter were subsequently published in *Statistics in Medicine* [92], and has lead to related research as a part of a wider research team into sample size calculations for external validation of a clinical prediction models for different outcome types [93, 94, 95, 96].

Chapter 5 describes the external validation of published prognostic models to predict continuous birthweight or the risk of delivering a growth restricted baby (FGR defined as birthweight $<$ 10th centile adjusted for gestational age, with severe complications: stillbirth, neonatal death or delivery before 32 weeks), using IPD meta-analysis methods to combine model performance estimates across multiple cohorts, allowing for a larger sample for external validation of these existing models. This clinical application discussed in this chapter forms part of a Health Technology Assessment (HTA) report, which has been accepted by the *HTA* for publication.

Chapter 6 discusses the use of EHR data to externally validate a model to predict the risk of a serious fall (resulting in hospitalisation or death), in those eligible for antihypertensive treatment in a primary care population. IPD meta-analysis methods included in Chapters 5 are further implemented to demonstrate the variability in model performance across different settings, investigating where precise overall estimates of model performance (given the large sample size for external validation) may have masked poor performance in GP practices with differing case-mix. This applied work has been published in *BMJ* [97], and has led to further research investigating other adverse events in those with an indication for antihypertensives [98], and for predicting falls risk in a wider population [99]

Chapter 7 summarises key findings and recommendations arising from the thesis, explains how this body of work adds value to the prediction research field, and outlines the limitations of this work. It concludes with a discussion of the next steps regarding further research in prediction modelling.

CHAPTER 2

Dichotomisation of continuous outcomes in prediction model research: a review of current practice in the prediction of Fetal Growth Restriction (FGR)

2 Chapter 2: Dichotomisation of continuous outcomes in prediction model research: a review of current practice in the prediction of Fetal Growth Restriction (FGR)

2.1 Introduction and objectives

Chapter 1 introduced the general concepts of prediction modelling, as well as the different modelling methods often employed in the development of prediction models with outcomes of different types. Prediction model research in primary care often involves the prediction of outcomes such as future blood pressure, pain scores, or depression levels, that are measured on a continuous scale. When developing a prediction model for such outcomes, researchers must decide on how best to treat them in their analyses. Historically, many researchers have opted to dichotomise their continuous outcome variable to form a binary outcome, which could then be modelled using a logistic regression framework. This allows for a model that produces predicted probabilities of an outcome for an individual, with clinical action then taken accordingly.

An example of this would be in the prediction of future pain intensity, where a pain score (on a scale of 0 to 100) is deemed "high" for a value exceeding 50 points, or some other clinically relevant cut off. A prediction model could be developed to give the probability that a patient will be experiencing high pain (above 50/100) at follow up, with such patients being referred for more intensive physiotherapy at first presentation in an attempt to prevent the anticipated negative outcome.

Dichotomisation of continuous outcome variables is often conducted in an attempt to facilitate model interpretation, to support the model's uptake and use in practice [33]. Modelling on a binary

scale allows for calculation of predicted probabilities for a particular outcome, which may be easier to communicate to patients, although in a shared decision-making context a continuous estimate is more useful than risk stratification [100]. The choice of threshold is often arbitrary, without consultation of patients to inform the appropriate threshold value. Furthermore, by dichotomising prior to modelling, researchers lose the opportunity to choose alternative, context-dependent threshold values after the analysis [100].

Dichotomising continuous outcome values risks losing valuable information needed to accurately estimate predictor-outcome relationships [101, 34, 102, 103]. Given that current studies developing a clinical prediction model often suffer from small sample sizes [104], this loss of information could have a detrimental impact on model performance and usefulness in practice, by affecting both the model's generalisability and its transportability to a new population [46]. As an example, a 2001 randomised-controlled trial comparing the efficacy of antidepressant drugs and counselling needed a sample size of only 88 participants to analyse their outcome (Beck depression inventory score) on its continuous scale, while 800 were needed to assess the impact of different treatments on the dichotomised outcome [105, 106].

Categorisation of continuous predictors prior to model development (regardless of outcome form) is widely regarded as unnecessary and inefficient, leading to a loss of information and a reduction in statistical power to identify any true relationship between the predictor variables and the outcome [34, 33]. Thus far, however, limited quantitative assessment has been conducted into the impact that categorisation of the *outcome* variable might have on a model's predictive performance. Of particular importance for a prediction model is how well it performs in external data, separate to that used for model development, to give an indication of how well the model will predict

outcomes for new patients in clinical practice [62]. It is therefore vital that predictor-outcome relationships are modelled appropriately and accurately to improve the chances of the model performing sufficiently well in data from new populations.

2.1.1 Clinical scenario

Before further research into the implications of varying modelling approaches, it is important to first investigate how continuous outcomes are currently treated in practice. The aim of this chapter is, therefore, to discuss the justification offered by authors (of prediction model research with continuous outcomes) for their choice of modelling approach, to identify clinical motivations for such choices to assess areas where outcome dichotomisation may be warranted. In particular, this review focuses on the prediction of birthweight (a continuous outcome) as a proxy for fetal growth restriction (FGR).

Babies who are born in the smallest 10% by birthweight are classified as being small-for-gestational-age (SGA), and are at an increased risk of perinatal complications as a result of restricted growth within the womb [107]. In these babies, the number of stillbirths and neonatal deaths is much higher than in those of normal growth [108]. Increased monitoring of growth restricted babies, gives the opportunity to identify and intervene early to improve outcomes. Thus efforts to anticipate adverse outcomes using risk prediction modelling techniques are commonplace in the literature.

Birthweight is often used as a proxy for FGR in prediction research, although may be modelled in a number of ways: on its continuous scale; using an arbitrary cut-point for dichotomisation;

using an informed cut-point for dichotomisation; or combining dichotomised birthweight with the occurrence of complications, to form a composite FGR outcome. Birthweight prediction, therefore, is a good clinical area in which to investigate current attitudes to outcome dichotomisation in prediction modelling. Throughout this chapter "birthweight" will be used to refer to the continuous outcome, while FGR will refer to a binary alternative (however defined).

2.1.2 Objectives

The primary objective of this chapter is to review published journal articles developing prediction models for birthweight, and thus to obtain and discuss:

1. how continuous outcomes are being modelled in the development of prediction models in practice;
2. what justification is given by researchers for dichotomisation of outcome variables, where such dichotomisation has taken place.
3. to assess whether dichotomisation of continuous outcomes is recognised as a strength or a limitation in articles' discussion sections;

The findings of this review will reveal current practice around handling of continuous outcomes in prediction of birthweight (as an indicator of FGR), as an example of prediction modelling research across all clinical areas, and will help to identify areas of good practice or where clearer recommendations may be necessary. The review will further motivate subsequent chapters in this thesis where continuous outcomes are modelled and recommendations are made for their inclusion in research.

2.2 Methods

2.2.1 Inclusion and exclusion criteria

There were no restrictions on study design, as long as the study was aiming to develop a clinical prediction model. Studies solely validating an existing model were excluded, as these were not relevant to the review question. Any primary studies identified during the searches or through links with collaborative groups were included if they provided details on the development of a prediction model for either FGR (binary) or birthweight (continuous). A summary of inclusion and exclusion criteria is given in Table 2.1.

Table 2.1: Inclusion and exclusion criteria

Inclusion criteria
Development of a clinical prediction model for either FGR (binary) or birthweight (continuous)
Population of pregnant women
Predictors including maternal clinical characteristics, biomarkers, and ultrasound measures
Prediction model including three or more predictors
Exclusion criteria
Only one predictor included in the "model" (univariable analyses)
Fewer than three variables in the prediction model
Full model not reported (e.g., missing intercept or baseline risk terms)
Reports only on external validation, without model updating
FGR or birthweight modelled only as a part of a composite outcome

Here, the term 'development' has been used to refer to the derivation of a prediction model, whether newly created or as an update of an existing model. All studies that reported the development of a multivariable model (containing at least three variables) to predict birthweight or the risk of FGR for use at any time in pregnancy, regardless of modelling approach. Studies that predicted FGR or birthweight only as part of a composite adverse outcome were excluded.

2.2.2 Search methods

Electronic searches

Published journal articles for this methodology review were identified through a systematic search for clinical prediction models for either FGR or birthweight, as a part of the International Prediction of Pregnancy Complications collaboration [109].

The following databases were searched, between July 2012 and December 2017, to identify articles containing relevant prediction models, for either FGR or birthweight in low risk pregnancies:

- MEDLINE
- EMBASE
- BIOSIS
- LILACS
- Pascal
- Science Citation Index
- Cochrane Database of Systematic Reviews (CDSR)

- Cochrane Central Register of Controlled Trials (CENTRAL)
- National Institute of Child and Human Development Data and Specimen Hub (NICHD-DASH)
- Database of Abstracts of Reviews of Effects (DARE)
- Health Technology Assessment Database (HTA)

MEDLINE and EMBASE were further searched with the same search strategy between December 2017 and January 2020 to update the original search. No language restriction was applied to the electronic searches.

Other searches

Prediction models reported solely in the grey literature were sought by searching relevant databases including Inside Conferences, Systems for Information in Grey Literature (SIGLE), dissertation abstracts and clinicaltrials.gov. Internet searches were further conducted in specialist gateways (JISC), general search engines (Google) and meta-search engines (Copernic).

Selection of studies

Titles and abstracts of all studies identified through searches were screened in duplicate by members of the IPPIC collaboration team, with discrepancies resolved through discussion and mutual consent.

2.2.3 Data collection and analysis

Data extraction and management

Data extraction was conducted using a pre-designed and piloted extraction form, with data stored in a spreadsheet in Microsoft Excel 2016. Extraction was completed by one reviewer. Where a paper presented multiple models for relevant outcomes (either binary or continuous), data was extracted for all models reported. The details of items to be extracted are summarised and described in Table 2.2. Information was extracted wherever reported.

Table 2.2: Description and brief explanation of items extracted from review papers

Item	Explanation
Author names	First author's surname, for study identification purposes
Title	Article title, for study identification purposes
Year	The year the model development was first published
Journal of publication	The name of the journal in which the model development was *first* published
Population at baseline	Brief description of recruitment criteria, including participant risk profile
Country of study	The country in which the research was conducted (not where recruitment took place)
Outcome definition	The outcome of interest being modelled, whether on the continuous or dichotomised scale
Study design	Brief description of study design. Was recruitment prospective, retrospective, or other?
Sample size	
Full dataset	The number of participants recruited prior to drop out and loss to follow-up
Data used in analysis	The number of participants who contributed to the prediction model development
Number of events	If a binary outcome, how many events were observed in the model development data?
Number of candidate predictors	How many candidate predictor parameters were considered at model development, prior to any statistical selection?
Review questions	
Outcome handling	How has the outcome been modelled: continuous, binary, time-to-event, ordinal
Categorisation basis	Where the outcome has been categorised, what is the basis used to define cut-points? e.g., standard clinical thresholds, previous study, derived in-sample

Item	Explanation
Method justification	The authors' description of any reasoning or justification for there choice of outcome handling choice, where such justification is given
Strength/limitation	Has the categorisation of continuous outcomes been discussed as a strength or a limitation in the discussion section of the study?

Assessment of risk of bias in included studies

As this review was concerned with a particular methodology and its use in practice, no effect estimates were extracted from the included studies. No formal risk of bias assessment was conducted, as biases assessed by tools such as PROBAST [110, 111] would not influence the use of the statistical methods this review concerns. Although no formal quality assessment took place, the narrative summary of review items gives some insight into the quality of the included research.

Assessment of heterogeneity

Data extracted from the included studies has been presented using descriptive methods and narrative summary. Data was also presented in a tabular format. No formal assessment of heterogeneity took place, as measures of effect were not extracted or combined.

2.3 Results

2.3.1 Identification of relevant articles

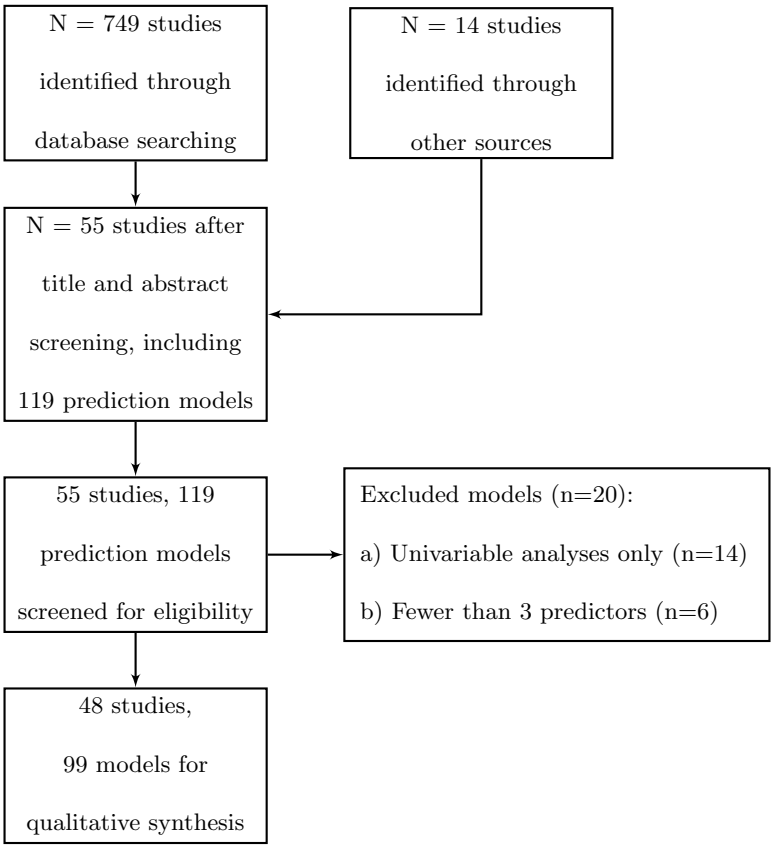


Figure 2.1: PRISMA flowchart showing numbers of studies identified, screened and included in the review

Searching of databases and other sources yielded a total of 55 studies, after duplicates were removed and title and abstract screening. These contained 119 different prediction models for outcomes of either FGR or birthweight. Twenty of these models were excluded as they contained fewer than three predictor variables, of which 14 referred to only univariable analyses rather than full models. Figure 2.1 shows the selection process for studies in this review.

2.3.2 Summary of articles included in the review

The majority of eligible publications were published in the past ten years, with 33 of the 48 publications (69%) published since 2012 (see Figure 2.2). Overall, the number of eligible studies appears to be increasing over time, while the number of eligible publications featuring the prediction of a continuous birthweight outcome was consistently low.

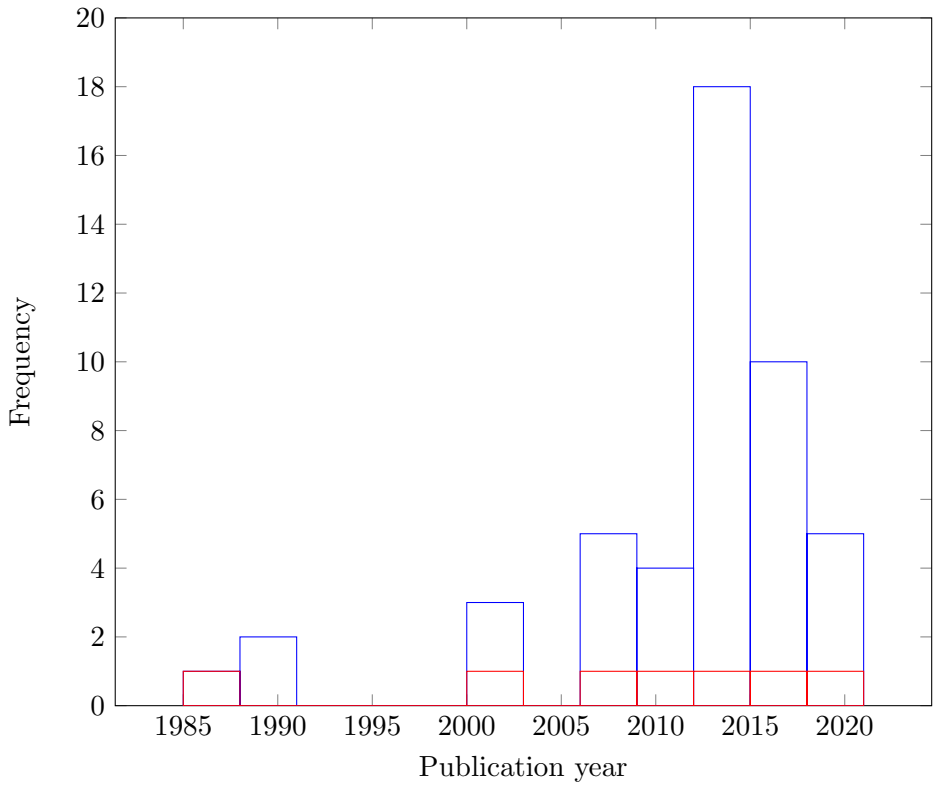


Figure 2.2: Stacked histogram showing the number of eligible publications over time. Red bars show studies developing models to predict continuous birthweight, while blue bars show studies predicting binary FGR.

The included studies represented research from across the globe, with Africa and Antarctica being the only continents not represented, as shown in Figure 2.3. Most studies arose from European,

Asian, or North American institutions. Almost half the total number of studies were conducted in the UK (23, 48%), predominantly arising from within the same few research groups, thus outcome handling decisions within these studies is likely to be more consistent than across other research teams.

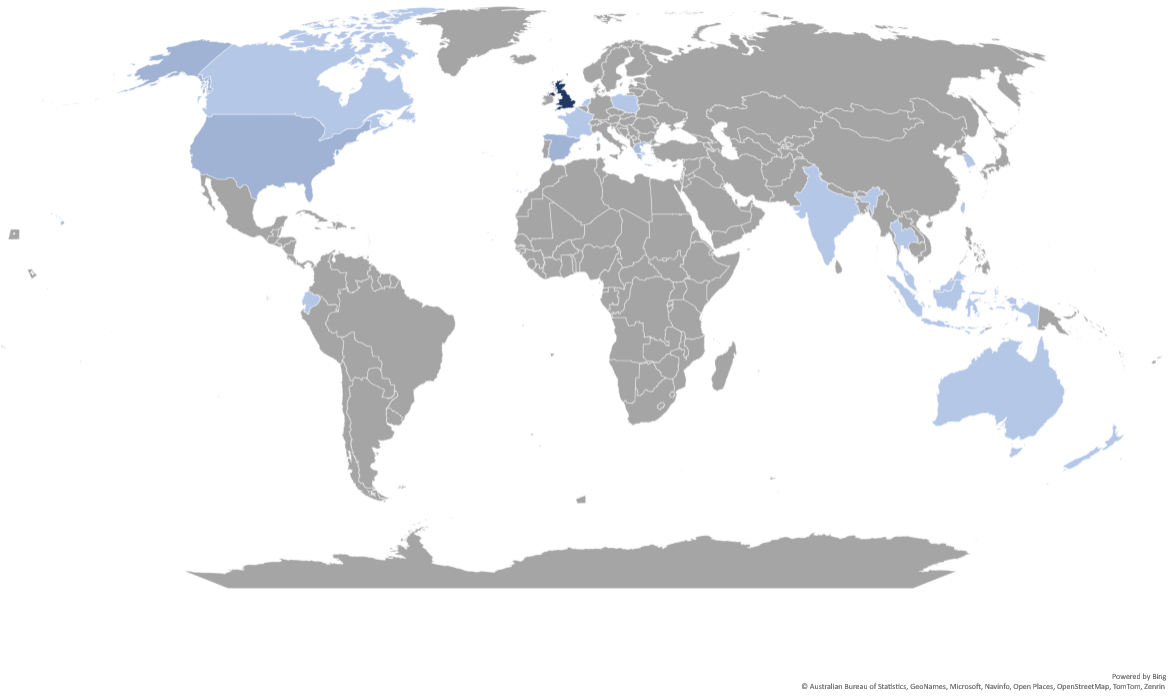


Figure 2.3: World map showing countries where studies included in this review were conducted. Darker blue indicates a higher number of publications from research teams in that country.

Prospective recruitment was seen in 37 (77%) studies, with seven (15%) developing models in retrospectively recruited cohorts. Other study designs represented were cross-sectional [112] and RCTs [113, 114]. Recruitment processes in the remaining study were unclear.

Sample sizes for model development tended to be reasonable, with a median total sample size

across studies of 2,125 (LQ to UQ: 761 to 12,190). Most studies saw reasonable numbers of events, although these were only defined in those studies modelling FGR as a binary outcome. Numbers of FGR events ranged from 22 [115] up to 5003 [116], with a median of 219 (LQ to UQ: 72 to 665). Correspondingly, numbers of predictors included in each model were relatively low, with a median of 5 (LQ to UQ: 4 to 8).

Table 2.4 gives a summary of characteristics across included studies, with a breakdown of features by outcome handling, while Table 2.5 summarises key characteristics across models, with models ordered by publication year. The validation option "External validation" refers to external validations included in the same report as the model development study. Further details on included studies and models is given in the appendix to this chapter.

Table 2.4: Table of study demographics for included studies. Numbers are n (%) or median [UQ to LQ], depending on data type.

	Studies *(N=48)*	Continuous *(N=7)*	Binary *(N=41)*
Country of study			
UK	23 (48)	2 (29)	21 (51)
Europe (excluding UK)	10 (21)	2 (29)	8 (20)
Asia	6 (13)	2 (29)	4 (10)
North America	6 (13)	1 (14)	5 (12)
Oceania	2 (4)	0 (0)	2 (5)
South America	1 (2)	0 (0)	1 (2)
Africa	0 (0)	0 (0)	0 (0)
Study design			
Prospective cohort	37 (77)	3 (43)	34 (83)
Retrospective cohort	7 (15)	4 (57)	3 (7)
Other	4 (8)	0 (0)	4 (10)
Outcome			
Birthweight (*continuous*)	7 (15)	7 (100)	-
Intrauterine Growth Restriction (*binary*)	8 (17)	-	8 (20)
Small-for-gestational-age (*binary*)			
Any definition	28 (58)	-	28 (68)
≤3rd percentile	1 (2)	-	1 (2)
≤5rd percentile	5 (10)	-	5 (12)
≤10rd percentile	4 (8)	-	4 (10)
Other (*binary*)	3 (6)	-	3 (7)
Risk profile			
High risk	13 (27)	2 (29)	11 (27)
Low risk	28 (58)	3 (43)	25 (61)
Unselected	7 (15)	2 (29)	5 (12)
Total sample size	2,125 [761 to 12,190]	1,298 [105 to 32,850]	3172 [772 to 9,850]
Number of events	219 [72 to 665]	-	219 [72 to 665]
Number of candidate predictors	5 [4 to 8]	5 [4 to 10]	5 [4 to 7]

Table 2.5: Summary of key dataset and analysis characteristics for each of the included models.

	Publication		Sample size		Predictors				Population				Outcome							Type of model			Validation			
Model reference	Paper reference	Year of publication	Number of pregnancies	Number of events	Number of predictors	Maternal characteristics	Biomarkers	Ultrasound	Low risk	High risk	Unselected	Unclear	Birthweight	SGA	BW $\leq$ 10th centile	BW $\leq$ 5th centile	BW $\leq$ 3rd centile	Early detection	Late detection	Linear regression	Logistic regression	Other	Internal validation	External validation*	Calibration assessed	Discrimination assessed
1	Ciobanu 2019a [117]	2019	19209	1803	5		✓	✓	✓					✓	✓						✓		✓		✓	✓
2	Ciobanu 2019a [117]	2019	19209	-	5		✓	✓	✓					✓			✓				✓		✓		✓	✓
3	Ciobanu 2019a [117]	2019	124443	15641	5		✓	✓		✓				✓	✓						✓		✓	✓	✓	
4	Ciobanu 2019b [118]	2019	44043	-	3	✓				✓				✓					✓		✓					✓
5	Ciobanu 2019b [118]	2019	44043	-	3	✓				✓				✓					✓		✓					✓
6	Ciobanu 2019b [118]	2019	44043	-	3	✓				✓				✓	✓						✓					✓
7	Ciobanu 2019b [118]	2019	44043	-	3	✓				✓				✓			✓				✓					✓
8	Sharp 2019 [114]	2019	105	-	4	✓		✓				✓						✓		✓			✓		✓	✓
9	Sotiriadis 2019 [119]	2019	105	-	5	✓	✓	✓				✓	✓							✓			✓		✓	✓
10	Sotiriadis 2019 [120]	2019	3250	109	6	✓	✓	✓	✓										✓		✓					✓
11	Sotiriadis 2019 [113]	2019	3250	56	6	✓	✓	✓	✓										✓		✓					✓
12	Sotiriadis 2019 [121]	2019	3250	292	6	✓	✓	✓	✓					✓							✓					✓
13	Sotiriadis 2019 [115]	2019	3250	109	7	✓	✓	✓	✓										✓		✓					✓
14	Sotiriadis 2019 [122]	2019	3250	56	7	✓	✓	✓	✓									✓	✓		✓					✓
15	Sharp 2019 [114]	2019	3250	292	7	✓	✓	✓	✓					✓							✓					✓
16	Hendrix 2018 [123]	2018	1094	45	8		✓	✓	✓					✓								✓	✓			✓
17	Anggraini 2018 [124]	2018	127	-	8	✓		✓			✓		✓							✓			✓	✓	✓	
18	Miranda 2017 [125]	2017	1590	175	8	✓	✓	✓		✓				✓	✓						✓		✓			✓
19	Miranda 2017 [125]	2017	?	-	5	✓	✓	✓				✓		✓	✓						✓					✓
20	Allen 2017 [126]	2017	1045	64	5		✓	✓			✓			✓		✓					✓					✓
21	Allen 2017 [126]	2017	1045	118	5		✓	✓			✓			✓							✓					✓
22	McCowan 2017 [127]	2017	5606	633	3	✓	✓	✓	✓					✓	✓						✓		✓		✓	✓
23	McCowan 2017 [127]	2017	5606	465	3	✓	✓	✓	✓												✓		✓		✓	✓
24	McCowan 2017 [127]	2017	5606	168	3	✓	✓	✓	✓												✓		✓		✓	✓
25	Kim 2017 [128]	2017	300	100	3		✓		✓					✓								✓				✓
26	Litwinska 2017 [129]	2017	?	43	3	✓	✓	✓			✓										✓					
27	González González 2017 [130]	2017	988	193	5			✓	✓											✓			✓			✓
28	González González 2017 [130]	2017	988	193	5			✓	✓												✓		✓			✓
29	Crovetto 2016 [131]	2016	?	979	9				✓					✓							✓					✓
30	Crovetto 2016a [132]	2016	?	462	9				✓												✓					✓
31	Crovetto 2016a [132]	2016	?	462	7				✓									✓			✓					
32	Crovetto 2016a [132]	2016	?	462	9				✓										✓		✓					✓
33	Lesmes 2015a [133]	2015	9715	-	10	✓	✓	✓		✓				✓							✓					
34	Lesmes 2015a [133]	2015	9715	481	10	✓	✓	✓		✓				✓	✓			✓			✓					✓
35	Lesmes 2015a [133]	2015	9715	-	10	✓	✓	✓		✓				✓	✓					?	✓					✓
36	Lesmes 2015a [133]	2015	9715	-	10	✓	✓	✓		✓				✓	✓					?	✓					✓
37	Lesmes 2015a [133]	2015	9715	46	10	✓	✓	✓		✓				✓		✓		✓			✓					✓
38	Lesmes 2015a [133]	2015	9715	-	10	✓	✓	✓		✓				✓		✓				?	✓					✓
39	Lesmes 2015a [133]	2015	9715	435	10	✓	✓	✓		✓				✓		✓				?	✓					✓
40	Lesmes 2015a [133]	2015	9715	-	10	✓	✓	✓		✓				✓			✓	✓			✓					✓
41	Lesmes 2015a [133]	2015	9715	-	10	✓	✓	✓		✓				✓			✓			?	✓					✓
42	Lesmes 2015a [133]	2015	9715	-	10	✓	✓	✓		✓				✓			✓			?	✓					✓
43	Lesmes 2015b [134]	2015	63975	3702	6	✓		✓		✓				✓							✓		✓		✓	
44	Lesmes 2015b [134]	2015	63975	-	6	✓		✓		✓				✓	✓			✓			✓		✓		✓	✓
45	Lesmes 2015b [134]	2015	63975	-	6	✓		✓		✓				✓	✓					?	✓		✓		✓	✓
46	Lesmes 2015b [134]	2015	63975	-	6	✓		✓		✓				✓	✓					?	✓		✓		✓	✓

	Publication		Sample size		Predictors			Population			Outcome						Type of model			Validation			
Model reference	Paper reference	Year of publication	Number of pregnancies	Number of events	Number of predictors	Maternal characteristics	Biomarkers	Ultrasound	Low risk	High risk	Unselected	Unclear	Birthweight	SGA	BW ≤ 10th centile	BW ≤ 5th centile	BW ≤ 3rd centile	Early detection	Late detection	Linear regression	Logistic regression	Other	Internal validation	External validation*	Calibration assessed	Discrimination assessed
47	Lesmes 2015b [134]	2015	63975	447	6	✓		✓		✓				✓		✓		✓			✓		✓		✓	✓
48	Lesmes 2015b [134]	2015	63975	-	6	✓		✓		✓				✓		✓			?		✓		✓		✓	✓
49	Lesmes 2015b [134]	2015	63975	3255	6	✓		✓		✓				✓			✓		?		✓		✓		✓	✓
50	Lesmes 2015b [134]	2015	63975	-	6	✓		✓		✓				✓			✓		?		✓		✓		✓	✓
51	Lesmes 2015b [134]	2015	63975	-	6	✓		✓		✓				✓			✓		?		✓		✓		✓	✓
52	Lesmes 2015b [134]	2015	63975	-	6	✓		✓		✓				✓			✓		?		✓		✓		✓	✓
53	Fadigas 2015a [135]	2015	5515	278	5	✓		✓		✓				✓		✓					✓		✓		✓	✓
54	Fadigas 2015c [136]	2015	5121	245	5	✓		✓		✓				✓		✓					✓		✓		✓	✓
55	Bakalis 2015a [137]	2015	9472	469	6	✓		✓		✓				✓							✓		✓			✓
56	Bakalis 2015b [138]	2015	9850	490	5		✓			✓				✓	✓						✓					
57	Bakalis 2015b [138]	2015	9850	-	5		✓			✓				✓		✓					✓					✓
58	Bakalis 2015b [138]	2015	9850	-	5		✓			✓				✓			✓				✓					✓
59	Bakalis 2015d [139]	2015	30849	1727					✓										✓		✓					
60	Kienast 2015 [122]	2015	346	26	3				✓	✓													✓			✓
61	Lesmes 2015c [116]	2015	88187	5003	4	?	?	?	✓	✓				✓		✓					✓		✓			✓
62	MacdonaldWallis 2015 [140]	2015	?	?								✓														
63	Papastefanou 2015 [141]	2015	1298	73	4	✓			✓	✓			✓							✓			✓		✓	✓
64	Seravalli 2014a [142]	2014	2267	191	4	✓		✓	✓	✓				✓							✓		✓		✓	✓
65	Seravalli 2014b [143]	2014	1982	36	6	✓	✓	✓	✓	✓				✓			✓				✓		✓		✓	✓
66	Seravalli 2014b [143]	2014	1982	136	6	✓	✓	✓	✓	✓				✓			✓				✓		✓		✓	✓
67	Schwartz 2014 [144]	2014	?	56	5				✓	✓				✓							✓		✓			✓
68	Boucoiran 2013 [145]	2013	772	72	3		✓		✓					✓								✓				
69	Boucoiran 2013 [145]	2013	772	-	3		✓		✓					✓		✓						✓			✓	✓
70	Boucoiran 2013 [145]	2013	772	-	3		✓		✓					✓		✓						✓			✓	✓
71	Poon 2013 [146]	2013	62052	3168	8	✓	✓	✓		✓										✓						
72	Poon 2013 [146]	2013	62052	397	8	✓	✓	✓		✓				✓						✓						
73	Poon 2013 [146]	2013	62052	2771	8	✓	✓	✓		✓				✓						✓						
74	Singh 2013 [147]	2013	100	65	4			✓		✓												✓				
75	Yadav 2013 [112]	2013	666	93	7	✓			✓					✓							✓					
76	Seed 2011 [148]	2011	1121	255	6					✓				✓							✓		✓		✓	✓
77	Seed 2011 [148]	2011	1121	104	4					✓				✓							✓		✓		✓	✓
78	Poon 2011 [149]	2011	32850	1536	-				✓												✓		✓			
79	Karagiannis 2011 [150]	2011	32850	163	-				✓					✓							✓					
80	Karagiannis 2011 [150]	2011	32850	1373	-				✓				✓								✓					
81	Karagiannis 2011 [150]	2011	32850	1536	10	✓			✓				✓				✓					✓				
82	Poon 2011 [149]	2011	32850	1536	11	✓		✓	✓					✓							✓					
83	Poon 2011 [149]	2011	32850	1536	11				✓												✓		✓		✓	✓
84	De Paco 2008 [151]	2008	4376	532	5				✓					✓							✓		✓		✓	✓
85	Onwudiwe 2008 [152]	2008	3172	366	5				✓					✓							✓		✓		✓	✓
86	Liu 2008 [153]	2008	1322	-	11	✓				✓			✓											✓		
87	Plasencia 2007 [154]	2007	6015	760	7				✓					✓							✓				✓	✓
88	Pilalis 2007 [155]	2007	878	94	4				✓					✓												
89	Bachman 2003 [115]	2003	260	22	3				✓												✓		✓			✓
90	Doherty 2002 [113]	2002	114	86	11						✓												✓			
91	Mamelle 2001 [156]	2001	71778	-	5	✓			✓				✓							✓						
92	de Caunes 1990 [121]	1990	746	-	12				✓														✓		✓	
93	de Caunes 1990 [121]	1990	746	-	10				✓														✓		✓	
94	Snidvongs 1989 [120]	1989	766	71	6				✓													✓				
95	Weiner 1985 [157]	1985	33	-	3			✓		✓			✓							✓				✓	✓	
96	Weiner 1985 [157]	1985	33	-	3			✓		✓			✓							✓				✓	✓	
97	Weiner 1985 [157]	1985	33	-	3			✓		✓			✓							✓				✓	✓	
98	Weiner 1985 [157]	1985	33	-	3			✓		✓			✓							✓				✓	✓	
99	Weiner 1985 [157]	1985	33	-	3			✓		✓			✓							✓				✓	✓	
						54	43	58	46	44	5	4	12	65	15	11	11	11	8	17	70	12	19	8	20	69

2.3.3 Model development methods

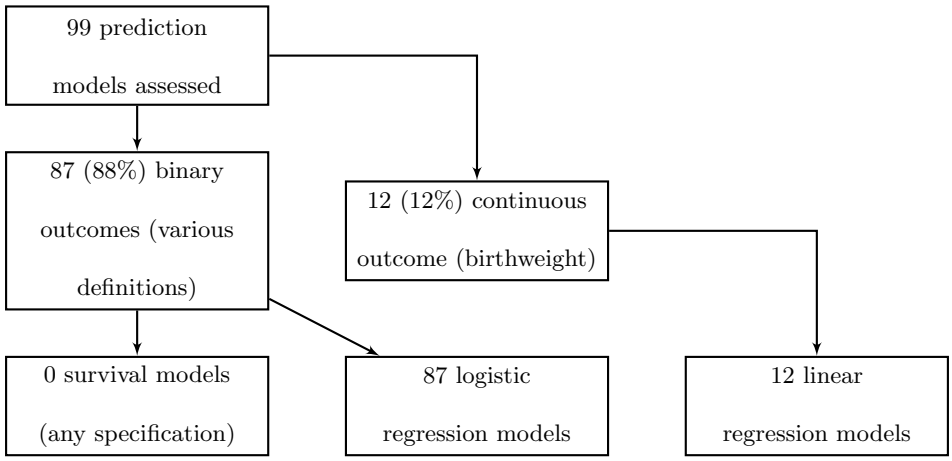


Figure 2.4: Summary of outcome handling and analysis discussed for included models

Of the 99 prediction models identified, 87 (88%) were to predict the probability of FGR as a binary outcome, while only 12 (12%) aimed to predict birthweight on a continuous scale (see Figure 2.4). All binary outcome models used logistic regression, while birthweight was solely modelled using linear regression. In three cases, birthweight was transformed to the log scale prior to modelling. No studies modelled the time-to-FGR as a survival outcome.

2.3.4 Binary FGR definitions

Of the 41 papers reporting on models for binary FGR outcomes, 11 (27%) gave no information about the source of the cut-point used for dichotomisation, as shown in Figure 2.5. A further four papers (10%) gave some information, but the source for their cut-point was unclear. Where sufficient detail was given, cut-points were derived from previously published information (14, 34%), from local standards (9, 22%), or in sample (3, 7%).

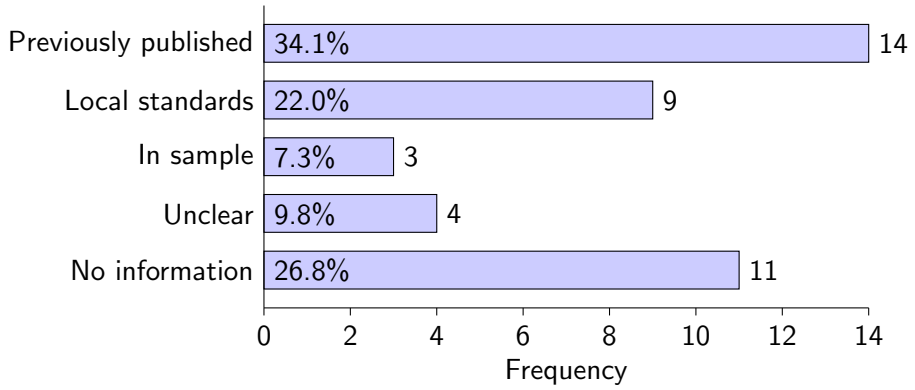


Figure 2.5: Bar chart showing the frequency of each source of cut-point for dichotomising birthweight to define FGR, as reported in the literature

Figure 2.6 and Table 2.7 summarise information given about the FGR outcome definition, across studies and by study respectively, for those that gave any information on their dichotomy source. Most often, birthweight was dichotomised at the tenth percentile, with 14 of the 30 studies (47%) reporting this cut-point. In four of these studies, the cut-point value was adjusted for gestational age at delivery (GA) and in one case was adjusted for both GA and maternal characteristics.

The second most common cut-point was at the fifth percentile value (11, 37%), a choice that was referenced to previously published information in all-but-one cases. Fifth percentile values were adjusted for GA in nine of the 11 studies using this cut-point. Other definitions of FGR were formed using: composite measures, including birthweight percentile values with other outcome components [125, 131, 132]; single cut-point value of 2,500g (referencing the previous literature) [112]; and derived in-sample, after scaling birthweight by crown-to-heel length [115].

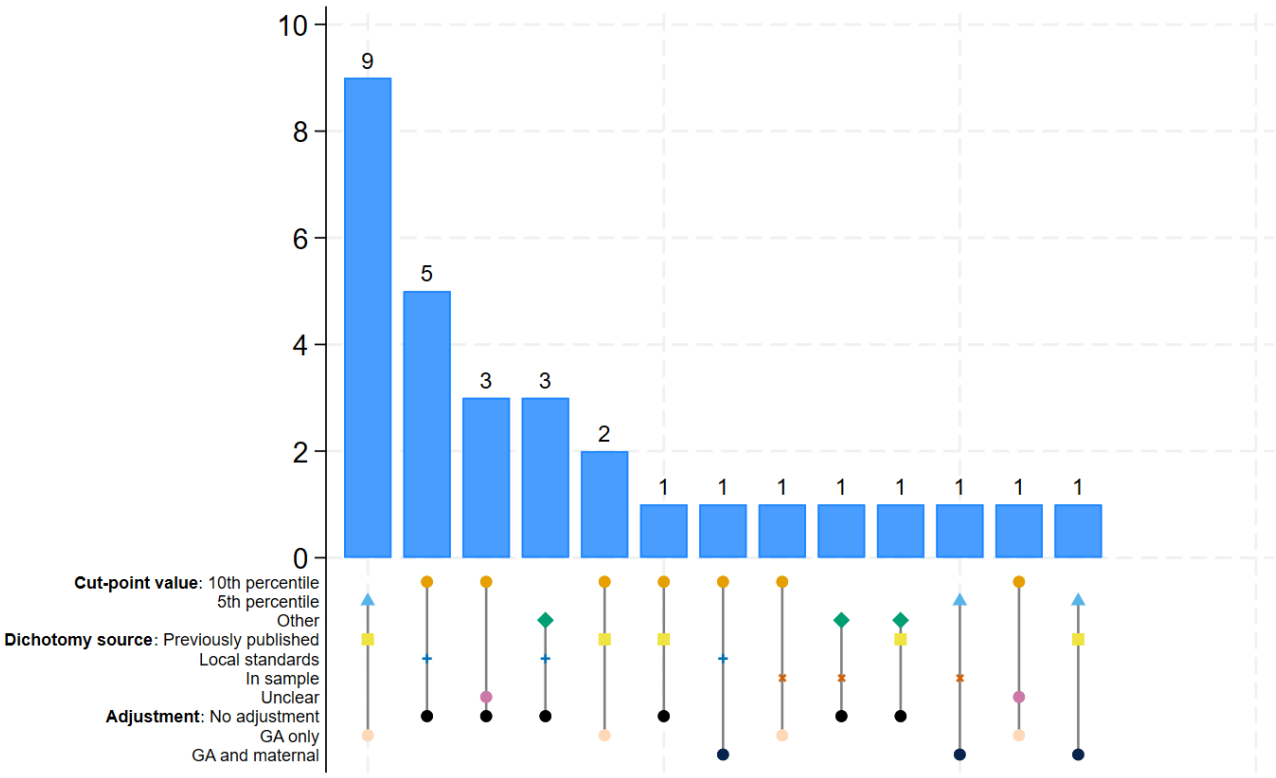


Figure 2.6: Bar chart showing the frequency of different combinations of cut-point value, dichotomy source, and adjustment factors in the dichotomy

The most common FGR definition seen was a cut-point at the fifth percentile, adjusting only for GA, referenced to previously published literature. This combination was seen in 9 of the 30 studies (30%), all of which were published from the same research group (based in the UK) over a 13 year range.

The second most common combination, seen in five (17%) of the publications, was a cut-point at the tenth percentile with no adjustment, referenced to local standards. Unlike the above, this definition was seen across independent research groups, with researchers based in five different countries (Australia, Greece, Netherlands, Spain, and United States) between 2002 and 2019.

Table 2.7: Summary of cut-point value, dichotomy source, and any adjustment factors in the dichotomy, for publications that gave information on how they dichotomised their birthweight outcome.

	Cut-point value			Dichotomy source				Adjustment		
Publication	10th percentile	5th percentile	Other	Previously published	Local standards	In sample	Unclear	No adjustment	GA only	GA and maternal
Ciobanu 2019a [117]	✓			✓					✓	
Ciobanu 2019b [118]	✓						✓		✓	
Sotiriadis 2019 [119]	✓				✓			✓		
Sharp 2019 [114]	✓						✓	✓		
Hendrix 2018 [123]	✓				✓			✓		
Kim 2017 [128]	✓			✓				✓		
Miranda 2017 [125]			✓		✓			✓		
McCowan 2017 [127]	✓				✓					✓
González González 2017 [130]	✓				✓			✓		
Crovetto 2016 [131]			✓		✓			✓		
Crovetto 2016a [132]			✓		✓			✓		
Lesmes 2015a [133]		✓		✓					✓	
Lesmes 2015b [134]		✓		✓					✓	
Fadigas 2015a [135]		✓		✓					✓	
Fadigas 2015c [136]		✓		✓					✓	
Bakalis 2015a [137]		✓		✓					✓	
Bakalis 2015b [138]		✓		✓					✓	
MacdonaldWallis 2015 [140]	✓					✓			✓	
Lesmes 2015c [116]		✓		✓					✓	
Bakalis 2015d [139]		✓		✓					✓	
Schwartz 2014 [144]	✓						✓	✓		
Seravalli 2014a [142]	✓				✓			✓		
Seravalli 2014b [143]	✓						✓	✓		
Yadav 2013 [112]			✓	✓				✓		
Poon 2011 [149]		✓		✓					✓	
Onwudiwe 2008 [152]		✓				✓				✓
Plasencia 2007 [154]		✓		✓						✓
Pilalis 2007 [155]	✓			✓					✓	
Bachman 2003 [115]			✓			✓		✓		
Doherty 2002 [113]	✓				✓			✓		
	14	11	5	14	9	3	4	14	13	3

2.3.5 Justification for dichotomisation

Only two papers gave any justification for their choice in cut-point value. The first, Bachman et al 2003 [115], derived their dichotomisation definition in-sample, using the 25th percentile of birthweight scaled by the crown-to-heel length to define neonates as having FGR. In their methods section, their decision is described thus:

> "These cut-off values were chosen because they identify only a small proportion of the low-risk population as being growth-restricted and as such are likely to represent true growth restriction." (Bachman 2003 [115])

In their discussion, Bachman et al further state that their inclusion of crown-to-heel length in their outcome definition as an outcome over birthweight alone was due to previously shown associations between neonatal anthropometry and subsequent infant outcomes. They did not reflect on their decision to dichotomise their scaled birthweight measure.

Sotiriadis et al 2019 [119] used a composite outcome definition, where FGR was considered present for those with either a birthweight (or estimated fetal weight (EFW) prior to birth) below the 10th percentile if accompanied by ultrasound abnormalities, or below the 3rd percentile, regardless of abnormalities. In their methods section, Sotiriadis et al justified their FGR definition as follows:

> "This is a modification of the recently published consensus definition of FGR" (Sotiriadis 2019 [119])

On discussion, they confirmed that this definition identified a clinically relevant high-risk group, with reference to Neonatal Intensive Care Unit (NICU) admission rates. They further commented on the heterogeneity seen in published definitions for (late-presenting) FGR. They did not discuss their dichotomisation choice as either a strength or a limitation of their approach.

No report included any further consideration of dichotomisation of the continuous birthweight value, as either a positive or negative aspect of their study design.

2.4 Discussion

2.4.1 Summary of main findings

This review has found that outcome dichotomisation is common in the prediction of birthweight as a proxy for FGR, with 87% of the identified models (87, from a total of 99) predicting some binary version of the birthweight outcome. No studies were identified modelling time-to-FGR as a survival outcome, thus the most common statistical modelling method used was logistic regression.

Birthweight cut-points were most often referenced to previously published literature or local standards, a far more appropriate approach than derivation in sample, but still lacked consistency across studies. A high proportion of the studies predicting a binary outcome (11, 27%) gave no information on their definition or cut-point choice. Where details were given, varying definitions were used across the identified studies, with cut-points based on tenth (14, 47%) or fifth (11, 37%) percentile values being most common.

Justification of definitions was rare in either the methods or discussion sections of reports. Where discussions of the outcome definition were included, these were invariably referenced to the outcome's association with subsequent infant outcomes (such as developmental delays, or NICU admission) rather than statistical or clinical benefits regarding maximum use of available information, facilitating interpretation, or ease of use. Thus, categorisation of birthweight outcomes is generally not considered as a limitation or a strength in the discussion section of studies. This may be due to a current lack of evidence around the impact of outcome dichotomisation in a clinical setting, thus this is not an issue currently being considered important by researchers.

2.4.2 Strengths and weaknesses of the review

Identification of prediction modelling studies for systematic reviews in prognosis research is known to be difficult, generally due to a lack of standard indexing terms and paper descriptors [158, 159, 160]. The search strategies used here were intended to have high sensitivity in identifying prediction model studies, focussing on ruling out clearly irrelevant papers prior to screening. It is still possible that some relevant publications were missed by these strategies. Given this review involved an overview of statistical methods, and not a comprehensive summary of models in the FGR field, it is unlikely that missed papers would have a substantial affect on the conclusions drawn. The search used in this review is therefore likely to have given a representative view of the current methods being used in the FGR prediction modelling literature.

Extraction of items for this methods review was conducted by a single researcher, thus some items may have been missed that could have been picked up by second review. Given the presence or absence of reporting is objective, it is unlikely that any systematic bias in extraction was present. Thus, where no justification for outcome handling has been reported here, this can be interpreted as no *clear* justification was given. Regardless, extraction and discussion with a second reviewer would have improved the validity of the conclusions from this review, and may have identified some reported items that were missed on single assessment.

Multiple papers within this review were conducted and reported from within the same research teams, thus outcome choice and reporting levels were not independent between observed studies and models. This can be seen, with some authors being lead on up to three different papers under review. Further, research published from the *Harris Birthright Research Centre for Fetal Medicine* (King's College Hospital, UK) contributed 15 of the papers identified for this review,

accounting for 37% of papers reporting a binary outcome prediction model. Consistent modelling and reporting choices within this team's work may not be representative of the wider literature, and may have given more extreme results than if only independent publications were considered. Similarly, if one such research group had discussed their decisions well, and published prolifically, this would have an extreme positive impact on the literature in the clinical area. Thus, whether a limitation of the review or of the literature itself, it should be noted that conclusions here are reported with the caveat of publications and their approaches not being entirely independent from one another.

2.4.3 Applicability of findings to the review question

Given this review of methods was confined to only one clinical area, that of birthweight and FGR risk, it is possible that findings here may not be generalisable to wider clinical areas. The area under review was chosen as it gave a somewhat unique example where dichotomisation of the continuous birthweight outcome might be clinically useful when combined with pregnancy and birth complications to give a composite outcome definition. This is relevant as FGR is not simply a case of low birthweight, but some restriction in growth. As such, it is likely that this review gives an optimistic view of the current level of justification for outcome dichotomisation in the literature. This is an area where justification and explanation was particularly warranted, and yet was still lacking.

2.4.4 Conclusions and next steps

Outcome dichotomisation is common in the prediction of birthweight as a proxy for FGR, with a lack of consistent cut-points and definitions across studies. Reporting the reasoning behind dichotomisation decision justification is rare, with no discussion of possible strengths or weaknesses of the authors' chosen approach. Further consideration should be given to whether outcome dichotomisation is necessary in individual studies.

In this case, predictions on the continuous scale could give a clear indication of the expected extent of any growth restriction, allowing identification of clinically meaningful, smaller changes in expected birthweight that would be missed by a solely binary outcome. Predicting birthweight outcomes on the continuous scale also maximises the available information for detecting predictor-outcome associations during model development. In particular, prediction models with continuous birthweight outcomes would have greater flexibility for use in different contexts or geographical locations, where a different birthweight dichotomy might better reflect restricted growth. This is especially relevant for those studies basing their dichotomisation cut-point on local birthweight standards.

The following chapters will expand on the concepts of outcome handling in prediction modelling research, delving further into the impact and implications of dichotomising continuous outcome variables, with applied examples in both model development and validation. There will be further discussion of prediction modelling applications in FGR prediction in Chapter 5. Chapter 3 will now demonstrate a more straightforward clinical example, in the prediction of future pain intensity for those with neck and/or lower back pain, illustrating differences in predictive performance for models developed with or without outcome dichotomisation.

CHAPTER 3

Development, evaluation and comparison of continuous and binary outcome prediction models for pain outcomes following primary care consultation for neck and/or low back pain

3 Chapter 3: Development, internal and external validation of prediction models for pain outcomes following primary care consultation for neck and/or low back pain

3.1 Introduction and objectives

The literature review in the previous chapter highlighted how the continuous outcome of birthweight is commonly dichotomised in practice, and demonstrated the lack of any justification offered by authors for their choice of outcome type. The clinical example discussed in Chapter 2 was a scenario where there was a possible clinical benefit to a binary outcome model for FGR over-and-above a simple dichotomisation of continuous birthweight, in that binary definitions could be used to include important clinical complications in a composite outcome definition. Thus, dichotomisation might have been reasonably justified in this case.

Chapter 2 focussed purely on how often birthweight was dichotomised in prediction model development studies, along with authors' justification of their outcome handling choice. It did not discuss the impact of this choice on the resulting prediction models. To delve further, this chapter now investigates how outcome dichotomisation might influence the performance of a prediction model, with reference to prediction distributions and statistical measures of predictive performance. Such measures contribute towards assessment of a model's appropriateness for use in practice, and so are highly relevant to the clinical utility of a prediction model.

The focus will now be in a more straightforward clinical setting, where the binary outcome is purely a dichotomised version of the continuous outcome. Thus, the binary outcome here has no clinical benefit above a continuous outcome at implementation. This chapter, therefore, aims

to give a fair comparison of both model performance and usefulness for continuous and binary outcome models.

The methods section of this chapter will be broken into two parts. Part (i) will introduce the methods used to calculate predictions from models for both outcome types. It will further discuss a proposed approach to gain predicted probabilities from the output of a linear regression model, in order to assess the risk of a dichotomised outcome variable, after modelling on the continuous scale. Part (ii) will describe a clinical example in which this approach is applied. The results section then demonstrates these proposed methods and compares the results of modelling on the different scales in terms of predictive performance. In the included clinical setting, the outcome was discrete pain intensity, a score ranging from zero to ten, where an increased value implies a more severe pain outcome. In practice, such scores are commonly modelled after dichotomisation (into high and low pain) to facilitate interpretation for both doctors and patients. Pain intensity equally could be modelled on a continuous scale to retain the maximum information available for the analyses.

3.1.1 Clinical scenario

The aim of the clinical research project associated with Chapter 3 is to develop prediction models for patients' pain intensity outcomes following initial consultation in primary care, to identify patients presenting with neck and/or low back pain (NLBP) who might benefit from an altered treatment pathway [91]. This project builds on previous research that showed an increase in the number of people progressing to disabling NLBP internationally [161], and demonstrated how the transition from acute to persistent pain can be predicted [162, 163].

This research forms a part of a body of work, aiming to develop varied digital health technologies to support first-contact decision making for consulters with NLBP [164]. Thus, predictions in both the continuous and binary form are desired to facilitate communication of expected disease trajectory, with intended side-by-side visualisations of both outcome predictions for an individual.

3.1.2 Objectives

To further investigate the impact of outcome dichotomisation on model performance and clinical usefulness, this chapter focuses on the development and, importantly, the validation of prediction models for pain intensity at six months, for those consulting in primary care with an NLBP problem. The particular focus was on comparing the models' predictive performance when modelling pain intensity as a continuous versus a binary outcome variable. The objectives of this chapter, therefore, were to:

1. Demonstrate methods of calculating predicted values and probabilities from a linear regression model, and probabilities from a logistic regression model;
2. Develop prediction models to estimate six-month pain intensity outcomes, on both the continuous and dichotomised scales;
3. Evaluate the predictive performance of the developed prediction models on internal validation, in the model development data; and to
4. Evaluate the predictive performance of the developed prediction models on external validation, in new data.

3.2 Methods, part (i): proposal for calculating predicted probabilities from a linear regression model

Following development, clinical prediction models need to be applied to calculate individual-level predictions, whether this be an expected outcome value or expected probability of an outcome event. Methods to generate these predictions have common elements across model types (for example, calculation of a linear predictor value), but fundamentally predictions are in the form of the outcome being modelled.

Methods to calculate predictions from both linear and logistic regression models, for example for predicted pain intensity scores and probabilities of high pain, respectively, are shown below. Also demonstrated is a proposed transformation of the output from the linear regression model, to further gain a predicted probability of an individual having a high risk of the dichotomised pain outcome, after having modelled pain intensity in its continuous form.

3.2.1 Generating predicted values from a linear regression model

Following model development by linear regression, the resulting model equation can be applied to participants to generate individual-level predictions of their six-month pain intensity score, as follows.

$$Y_{PRED_i} = \alpha_{Cont} + \boldsymbol{\beta_{Cont}} \boldsymbol{X_i} = \alpha_{Cont} + \beta_{Cont1} * X_{1i} + \beta_{Cont2} * X_{2i} + ...$$

Here Y_{PRED_i} denotes the predicted pain intensity score for individual i, α_{Cont} gives the intercept

term from the linear regression model, and β_{Contj} gives the coefficient (predictor effect estimate) from the linear regression model for predictor variable X_j.

The predicted outcome value Y_{PRED_i} is then obtained for an individual (i) by applying the right-hand side of the equation, which utilises the intercept value (α_{Cont}), the individual's reported values for the included predictors (X_{1i}, X_{2i}, X_{3i}, etc.) and their corresponding predictor effects from the linear regression model (β_{Cont1}, β_{Cont2}, β_{Cont3}, etc.).

3.2.2 Generating predicted probabilities from a linear regression model

While the performance of linear and logistic regression models can be compared to a certain extent, being able to gain estimates of predicted probabilities using the above equation from a linear regression would allow a much fairer comparison, as predictions would then be on the same scale. If this were possible, it would allow contrast of predictions stemming from dichotomisation before or after the modelling stage. Thus, methods were investigated to calculate a predicted probability of high pain using only the output from the linear regression.

For any given clinical prediction scenario, given some cut-point value C, the probability $p_i = P(Y_i < C)$ or $P(Y_i > C)$, where Y_i denotes the observed outcome value on the continuous scale, may be of particular interest to help inform treatment decisions. This C might be context dependent, so may be different in different locations or for different people. For example, in the applied example included in this chaper, predicted probabilities of pain intensity scores greater than or equal to five were of particular interest to the clinical team (see below), though in a different setting a higher or lower cut-point, anywhere on the 0-10 scale, might be preferable.

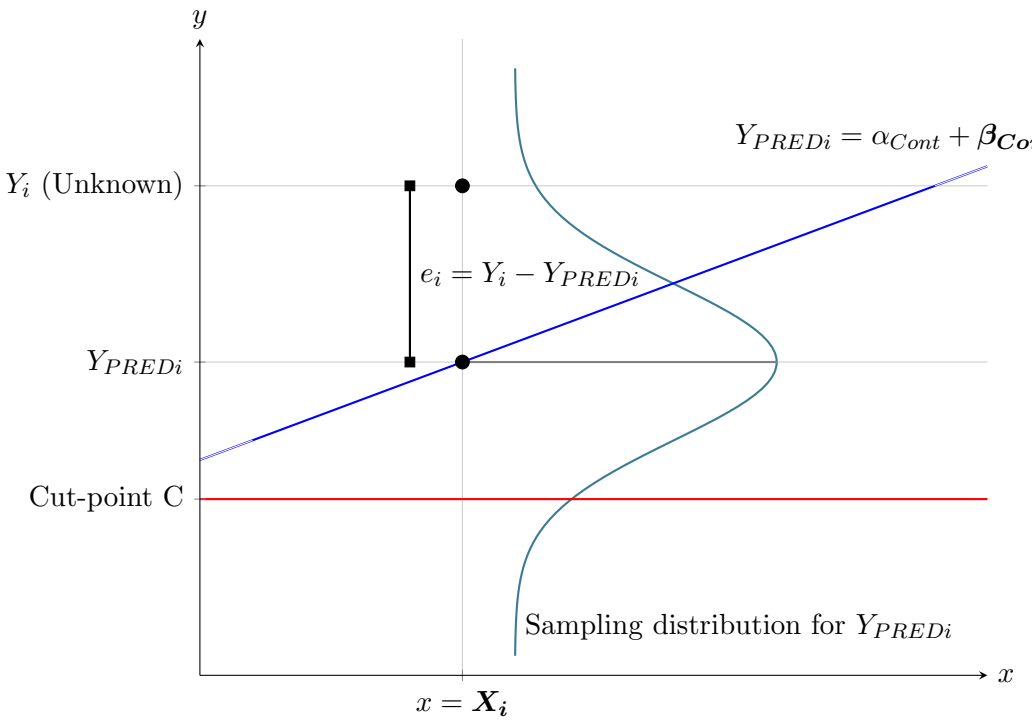


Figure 3.1: Demonstration of the distribution of a predicted outcome from the linear regression model $Y_{PREDi} = \alpha_{Cont} + \boldsymbol{\beta_{Cont}} \boldsymbol{X_i}$, where $\boldsymbol{X_i} = (X_{1i}, X_{2i}, X_{3i}, ...)^T$ are the values of the predictor variables for individual i, and $\boldsymbol{\beta_{Cont}} = (\beta_{Cont1}, \beta_{Cont2}, \beta_{Cont3}, ...)^T$ are the corresponding coefficients from the linear regression model

Predicted outcomes from a linear regression model are subject to some uncertainty, the magnitude of which depends on the sample size used for model development, among other things. Thus, while the prediction itself is on the continuous scale, we can also consider the sampling distribution for Y_{PREDi} given the particular combination of attributes $\boldsymbol{X_i}$ for individual i. Figure 3.1 shows such a distribution for the predicted outcome value (Y_{PREDi}) from a hypothetical linear regression model. The proportion of the distribution to the left (or right) of the cut-point equals the estimated probability that the individual's observed value is above (or below) the cut-point.

Considering the distribution of the predicted value, $\mathbf{Y_{PREDi}}$

We consider the distribution of the predicted outcome value from the linear regression model.

$$Y_{PREDi} = E(Y_i \mid x = \mathbf{X_i}) = \mu_{Y_i|x=\mathbf{X_i}}$$

where $\mu_{Y_i|x=\mathbf{X_i}}$ is a normally distributed random variable, as it is a linear combination of the observations in the model development data. The value of Y_{PREDi} is, therefore, dependant on a sample of the population (those in which the model was developed) rather than the population itself. As such, Y_{PREDi} follows a sampling distribution, in the same way that the values for α_{Cont} and the β_{Contj} do.

The predicted value, Y_{PREDi}, is, in fact, known to follow a student's t distribution with $n - p - 1$ degrees of freedom, where p is the number of predictor parameters, and n is the number of participants contributing to the model development. The t distribution is generally used to estimate population-level parameters when the population variance itself is unknown, or the sample size is too small for a normal approximation to be appropriate. Values of the t distribution (or "t-statistics") are given by:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

where $\bar{x}$ is a sample mean, μ is the population mean, s is the standard deviation in the sample, and n is the sample size.

When applying this to our context, the probability of interest is that of the unknown value (observed outcome value) Y_i being as extreme as (or more extreme than) the desired cut-off value, C. Assuming the variance of the individual Y_{PREDi} is equal to the residual variance of the linear

regression model (${\sigma^2}_{model}$), the probability $P(Y_i < C)$ can be calculated from the probability tables from the sampling distribution with:

$$t = \frac{C - Y_{PREDi}}{\sigma_{model}}$$

This t-statistic is associated with a specific cumulative probability, representing the likelihood of gaining a sample (predicted outcome) value less than or equal to $\bar{x} = C$, with a population value assumed to be $\mu = \mu_{Y_i|x=\mathbf{X_i}} = Y_{PREDi}$.

Thus the desired probabilities can be estimated from the t-distribution as:

$$P(Y_i < C) = P\left(Z < \frac{C - Y_{PREDi}}{\sigma_{model}}\right)$$

Given a sufficiently large sample size ($n \geq 30$, say), this sampling distribution will converge with the normal distribution, thus either distribution would be appropriate for estimating the required probability.

Note that the above discusses a situation where the question is of Y_i lying below the cut-off, $p_i = P(Y_i < C)$. The above extends logically to $p_i = P(Y_i > C)$, given the symmetrical nature of the t and normal distributions, and to $p_i = P(Y_i \leq C)$, given $P(Y_i = C) = 0$ (by definition).

Thus, in this clinical example, the predicted probability of individual i being in high pain six months after their consultation can be derived from the linear regression model as

$$p_i = 1 - P\left(Z < \frac{5 - Y_{PREDi}}{\sigma_{model}}\right)$$

.

3.2.3 Generating predicted probabilities from a logistic regression model

Given an outcome that has been dichotomised prior to modelling, the individual-level outcome probabilities $p_i = P(Y_i < C)$, or similarly $P(Y_i > C)$, can be gained from the logistic regression equation. The dependant variable in the logistic regression equation is the logit transformation of the event probability for each individual i, as follows:

$$logit\,(p_i) = LP_i = \alpha_{Bin} + \boldsymbol{\beta_{Bin}} \boldsymbol{X_i}$$

where $p_i = P(Y_i < C)$ for the pre-defined cut-point of interest, α_{Bin} is the estimated intercept from the logistic regression model, $\boldsymbol{X_i} = (X_{1i}, X_{2i}, X_{3i}, ...)^T$ are the values of the predictor variables for individual i, and $\boldsymbol{\beta_{Bin}} = (\beta_{Bin1}, \beta_{Bin2}, \beta_{Bin3}, ...)^T$ are the corresponding coefficients from the logistic regression model.

The probability of individual i being in high pain at six months, then, is given by

$$p_i = \frac{\exp LP_i}{1 + \exp LP_i}$$

.

3.3 Methods, part (ii): development and validation of clinical prediction models for pain outcomes

3.3.1 Data source

Data were available from two existing datasets for model development and internal validation: the Keele Aches and Pains Study (KAPS) [165] and the STarT-MSK pilot study (STarT-MSK-pilot) [166]. External validation was conducted in data from a third dataset, containing participants from a similar population as the model development data, that was not available at the time of development: the STarT-MSK main trial (STarT-MSK-MT) [167]. Eligible patients for these analyses were defined the same way in all three datasets: adults (aged 18 and over) consulting at a participating GP practice with NLBP.

Model development data

Keele Aches and Pains Study (KAPS) [165] A prospective cohort study, recruiting between 2014 and 2016, in which patients who consulted one of 14 participating GP practices with common MSK pain presentations (back, knee, shoulder, neck or multi-site), were invited to participate. Patients were sent an invitation letter and survey pack following their initial GP consultation. Return of the completed questionnaire, which was typically 3-6 weeks after the GP consultation, included consent to participate, and patient-reported information on demographics and candidate predictor values. Follow-up questionnaires were mailed to participants six months after baseline.

The STarT-MSK trial pilot study (STarT-MSK-pilot) [166] A two-parallel-arm, cluster randomised controlled trial (cRCT) in 8 general practices, pilot-testing the feasibility of using the Keele STarT-MSK Tool [165] to stratify care, through matching treatment options to risk groups

for adults with one of the above five MSK pain presentations. Recruitment took place between November 2016 and May 2017. Four practices were randomly assigned to continue usual care, while the remaining four stratified care based on the Keele STarT-MSK Tool [165]. Following eligibility screening, consulters were invited to take part in monthly data collection over 6 months, regardless of practice allocation.

Only patients from the KAPS and STarT-MSK-pilot studies who consulted specifically for NLBP (reported as their primary pain site) were included in the present analyses. These patients were defined in slightly different ways for each study. Within KAPS, pain site was a self-reported patient response to the question "*When you recently visited your GP practice, which part of your body did you consult about?*" (patients who responded with "neck" or "back" were included). In the STarT-MSK-pilot, GPs were asked at the point of consultation via an automated electronic pop-up template: "*Please confirm the primary pain site the patient is consulting with today*" (patients were included if the GP confirmed "neck" or "back" pain).

External validation data

The STarT-MSK main trial (STarT-MSK-MT) [167] A two-armed cRCT aiming to assess whether stratified care, based on Keele STarT-MSK Tool [165] risk group allocation, resulted in improved pain outcomes for adults consulting their GP for MSK pain. Recruitment began in May 2018 and ended in July 2019. Trial results showed no significant differences in pain outcomes between treatment arms. Data collection was identical to the STarT-MSK-pilot methods, described above, with only patients who's GP confirmed "neck" or "back" pain included in the external validation data.

Predictors in the STarT-MSK-MT data were available for two distinct time-points: (i) at consultation, and (ii) by postal questionnaire two weeks after consultation. Information from the consultation was only available for participants in the intervention arm of the trial, due to implementation of the Keele STarT MSK Tool at consultation being the intervention under investigation. For this chapter, only data recorded through the postal questionnaires will be discussed, due to the similarity in recording method to the data used for model development. In the published paper relating to this chapter (Appendix IIa), predictor information collected at the time of GP consultation in patients in the intervention arm was used for additional validation analyses.

3.3.2 Outcome definition for continuous and binary outcomes

The outcome of interest was the patient's pain intensity at six-month follow-up, defined through participants' self-reported response to the question "*How intense was your pain, on average, over the last 2 weeks? [Responses on a 0-10 scale, where 0 is "no pain" and 10 is "worst pain ever"]*".

The binary pain intensity outcome was formed from a dichotomisation at a pre-defined cut-point of five on the 0-10 pain scale. This cut-point has been previously reported as corresponding to at least moderate pain [168, 169]), and was considered clinically meaningful by the physiotherapists in the research team. Thus, binary pain intensity was modelled with scores of 0-4 corresponding to "low pain" (a good prognosis), and 5-10 being "high pain" (a poor prognosis).

3.3.3 Candidate predictors

The ten items of the Keele STarT MSK Tool were chosen *a priori* to be predictors in both models, due to their expected clinical importance in predicting MSK pain intensity outcomes. No statistical variable selection was performed. Predictor variables were baseline pain intensity (on a scale from 0-10), self-management of pain condition, pain impact, walking short distances only, pain elsewhere, long-term expectations, other important health problems, emotional well-being, fear of harm and pain duration [165]. The additional predictor of primary pain site (back or neck) was also included in both models for face validity, and to allow for differences in expected prognosis for these clinically distinct groups. Further detail on all predictors included in the models is given in Table 3.1.

Predictor information for all cohorts was collected through a postal questionnaire sent to patients within a few days of their GP consultation.

Table 3.1: Question phrasing and possible response values for predictors used in models to predict 6-month pain intensity outcomes

Predictor	Phrasing	Coding
Primary pain site	When you recently visited your GP practice, which part of your body did you consult about?/Please confirm the primary pain site the patient is consulting with today	1 = back, 0 = neck
Pain intensity	On average, how intense was your pain [where 0 is "no pain" and 10 is "pain as bad as it could be"]?	0-10
Pain self-management	Do you often feel unsure about how to manage your pain condition?	1 = yes, 0 = no
Pain impact	Over the last two weeks, have you been bothered a lot by your pain?	1 = yes, 0 = no
Walking short distances only	Have you only been able to walk short distances because of your pain?	1 = yes, 0 = no
Pain elsewhere	Have you had troublesome joint or muscle pain in more than one part of your body?	1 = yes, 0 = no
Long-term expectations	Do you think your condition will last a long time?	1 = yes, 0 = no
Other important health problems	Do you have other important health problems?	1 = yes, 0 = no
Emotional well-being	Has pain made you feel down or depressed in the last two weeks?	1 = yes, 0 = no
Fear of harm	Do you feel it is unsafe for a person with a condition like yours to be physically active?	1 = yes, 0 = no
Pain duration	Have you had your current pain problem for 6 months or more?	1 = yes, 0 = no

3.3.4 Sample size

The sample size for all analyses was fixed due to the size of the available datasets. Therefore, comparisons were performed between the number of available participants and current minimum sample size recommendations (see Table 3.2) to assess whether the available data were likely to have been sufficient for developing [170, 76, 78] and externally validating [92, 94] these prediction models. These calculations did not account for the clustering of participants within the two datasets used for model development, or within individual GP practices, as (at the time of writing) there was no guidance for sample size requirements for model development or validation in clustered data settings.

Table 3.2: Numbers of participants and events required (per sample size recommendations) and available (complete outcome data) for each analysis. Values are number, or number (percentage)

	Available		Required	
	Participants	Events (%)	Participants	Events* (%)
Model development				
Continuous outcome	545	-	311	-
Binary outcome	545	240 (44.0%)	824	412 (50%)
External validation				
Continuous outcome	586	-	892	-
Binary outcome	485	275 (56.7%)	1946	1071 (55%)

*Based on the expected prevalence before data analysis

The sample size for model development should ensure precise estimates of the mean outcome value or overall risk, whilst minimising overfitting in the model's linear predictor and overall fit [78]. Based on the inclusion of 11 pre-defined predictor parameters (one continuous predictor, modelled linearly, and 10 binary predictors), 311 participants were required for the development of the

model for continuous pain score (assumed $R^2 = 0.22$ [165], mean pain score 5.3 with standard deviation 2.2 [171]), and at least 824 participants (with 412 "high pain" events, assuming an outcome prevalence of 50% and a default Nagelkerke's $R^2 = 0.15$ [76]) for binary pain outcomes. Thus, the available data exceeded the requirements for the continuous pain score model but was not sufficient for developing the model with the binary pain outcome.

Minimum sample size recommendations for precise estimation of model performance for prediction models with continuous outcomes were not available at the time of analysis, and will be discussed further in Chapter 4. Calculations will not be included in detail here, but, basing estimates on the model performance on internal validation, the minimum sample size required to meet criteria for the continuous outcome model was 892 (assuming an $R^2 = 0.39$) [92].

To meet the Collins et al rule-of-thumb recommendations [172] a minimum of 200 events (defined here as high pain intensity) and non-events were required to externally validate the binary outcome model. To meet the Riley et al tailored sample size criteria for external validation of the binary outcome model [94], at least 1946 (1071 events) were required. This requirement was driven by the criterion to precisely estimate the calibration slope, though also ensured precise estimation of the Observed to Expected ratio (O/E) and the c-statistic, and involved the following assumptions, taken from each model's performance on internal validation: high pain proportion of 55%, C-statistic of 0.81, linear predictor following a skew-normal distribution with a mean of –0.45, a variance of 2.17, a skewness parameter of –0.5, and a kurtosis parameter of 3.

Notably, according to current recommendations, the binary outcome model required more participants to minimise overfitting to the development data, and more participants to ensure

precise estimation of model performance measures on external validation, than the continuous outcome model.

3.3.5 Missing data

Missing data were seen in both predictor and outcome measurements for all three cohorts. Preliminary checks for associations between missingness and predictor or outcome values were conducted to test the validity of the assumption that data were missing at random (MAR) [173]. No obvious associations were seen, so multiple imputation by chained equations was used to account for missing data in both predictor and outcome variables, under the MAR assumption [174]. Imputation was conducted separately by cohort, rather than for the combined individual participant data (IPD), to allow for clustering of patients within each dataset in the imputed values. The imputed datasets for model development were combined by imputation prior to analysis ($KAPS_i + STarTMSKpilot_i = Imputation_i$, for each imputation i).

Forty-six imputed datasets were generated in both cohorts of the development data, equal to the maximum percentage of incomplete cases across the two [173]. Sixty-seven imputations were generated in the external validation data, by the same reasoning.

All candidate predictors were included in the imputation model, along with a number of auxiliary variables that were expected (based on clinical input from the wider research team) to be highly correlated with predictor variables with high proportions of missing values. The aim of including these extra variables when estimating the imputed values was to increase precision and decrease bias in the resulting prediction model estimates [175]. Auxiliary variables included: self-rated

health, intensity of least painful pain, EuroQol 5 Dimension (EQ5D) mobility domain, EQ5D anxiety/depression domain, co-morbidities (diabetes, breathing problems, heart problems, chronic fatigue, anxiety/depression and other), how often help was needed to read written materials, and kinesiophobia (fear of movement). These variables were included in the imputation models only, and were not considered as candidate predictors in the prediction models.

Continuous pain intensity outcome values were included in the imputation of predictor variables, being included in their continuous forms to maximise the available information included in the imputation model. Outcome values were imputed for individuals with missing outcome measurements as a part of the imputation process, with participants who had originally (prior to imputation) had missing data for the outcome being dropped prior to model development analyses [176].

Imputed values for all variables were checked through visual inspection of histograms (continuous variables) and frequency tables (categorical variables) to examine whether values were realistic and consistent across imputed datasets.

Results of analyses involving multiply imputed data were combined across imputations using Rubin's rules where appropriate [177, 174], and described through summary statistics (median, lower quartile (LQ), upper quartile (UQ), range) where Rubin's rules did not hold [178]. Calibration plots were checked for consistency across imputations and, where appropriate, a single representative example is shown.

3.3.6 Model development

Outcomes were modelled using multilevel mixed-effects regression to account for the combination of two distinct cohorts forming the development data [60]. The continuous outcome, pain intensity score, was modelled using linear regression (using the *mixed* command in Stata 16). The score was modelled as if it were a truly continuous variable, as is often the case with pain score data, for computational efficiency and ease of interpretation over an 11-category ordinal regression. The binary outcome of high pain was modelled using logistic regression (using the *xtmelogit* command in Stata 16). Models were fitted using restricted maximum likelihood (REML) and maximum likelihood estimation for the continuous and binary outcome models respectively, with an unstructured variance-covariance matrix for a random effect on the intercept term to account for clustering in the two model development datasets [64]. The continuous predictor of baseline pain score was modelled linearly and all predictors were forced into both models (with no statistical selection), as previously stated [34, 179].

3.3.7 Predictive performance measures and apparent performance

Following model development, outcome predictions were calculated for all individuals in the model development data, as described below. The predictive performance of the models was assessed through calibration and overall model fit for the continuous pain intensity outcome, and through calibration, overall model fit, and discrimination for the binary outcome [57]. The apparent performance of each model was calculated as the performance of the prediction models when applied directly in model development data.

Calibration was assessed using the calibration slope, calibration-in-the-large (CITL), and the

ratio of Observed to Expected cases (O/E, for the binary outcome predictions only). Calibration plots were produced for visual assessment of calibration performance, comparing predicted to observed pain intensity score for the continuous outcome, and proportion of high pain events for the binary outcome. Calibration curves were overlaid on calibration plots, produced using a loess non-parametric smoother. Discrimination was assessed using the c-statistic for the binary outcome predictions only. Overall model fit was assessed as the proportion of the variance in the outcome explained by the model predictions, using the adjusted R^2, for continuous pain intensity score. Pseudo R^2 values were calculated for binary outcome predictions, using Nagelkerke and Cox-Snell approaches.

3.3.8 Internal validation and shrinkage

Internal validation was conducted using bootstrapping with 1,000 samples, sampling with replacement, from the original data [60]. The full modelling process was repeated within each bootstrap sample, including multiple imputation [174]. The predictive performance of the models developed within each bootstrap sample was evaluated within the (imputed) bootstrap sample itself, as well as in the original (imputed) data. In both cases, performance measures were pooled across imputations using Rubin's Rules, on the log-scale for O/E, the logit-scale for the c-statistic, and their usual scales for the c-slope and CITL [72, 178].

The optimism in each performance measure was calculated as the average of the differences between the performance (in the original data) of the bootstrap and original models. Each optimism estimate was subtracted from the associated apparent performance measure, to provide optimism-adjusted estimates of predictive performance [180].

The optimism-adjusted calibration slope was also used as the estimate of the uniform shrinkage factor for each regression model. The regression coefficients were multiplied by this shrinkage factor to correct for overfitting to the model development data (a consequence of having a low number of participants or outcomes relative to the number of predictor parameters) [60, 54]. After shrinkage was applied to the coefficients, the intercept term (with random effect) was re-estimated for each model whilst holding the shrunken predictor effects fixed, to ensure predictions remained correct on average. The models with shrunken coefficients and re-estimated intercepts give the final prediction models, which were taken forward to external validation [180].

A visual representation of the sequence of analysis steps for the internal validation is given in Figure 3.2.

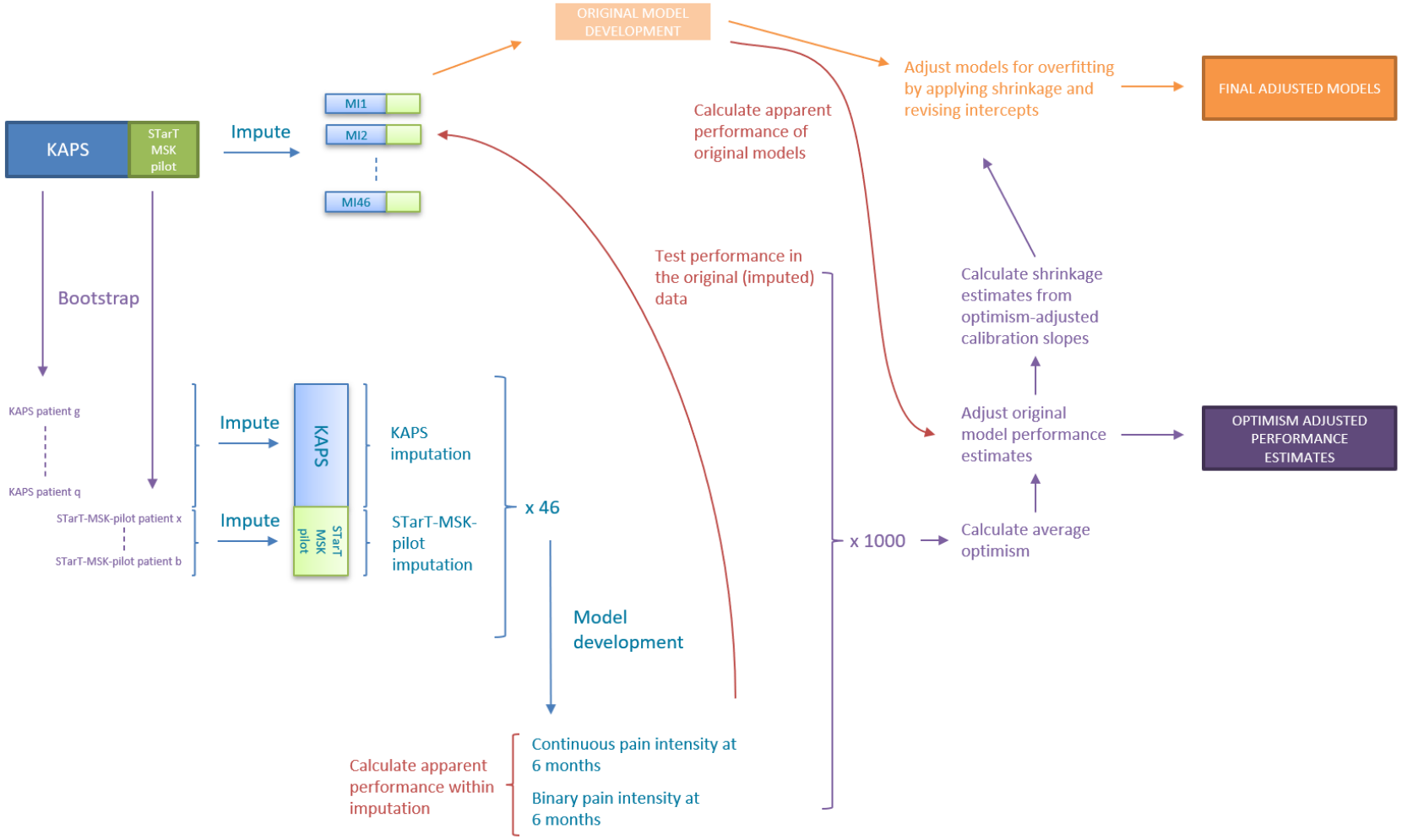


Figure 3.2: Flow chart showing the stages of analysis in the internal validation of models to predict six-month pain intensity

3.3.9 Model stability checks

Separate to the internal validation process, the prediction models were also checked for stability in the developed model across different development samples of the same size from the same population [181, 182]. Modelling procedures were repeated, as specified above, in 200 new samples of participants obtained by bootstrapping (sampling, with replacement, from the original dataset). Stability in individual-level predicted risks from the two modelling approaches was assessed and compared using prediction instability plots (showing the spread of predicted risks across bootstrap models, plotted against predicted risk values from the original model), instability indices (the mean absolute difference between the predicted risks for an individual, as calculated from the original and bootstrap models, plotted and summarised across all bootstrap samples), and classification indices (the proportion of bootstrap models that give a different outcome classification than the original model, given a pre-defined, clinically relevant threshold probability (e.g., 10% outcome risk), plotted for each individual). Consistency in smoothed calibration curves across bootstrap models was also assessed.

3.3.10 External validation

Model equations for both models were applied to the participants in the STarT-MSK-MT data to calculate individual-level predictions for each included patient. Predictive performance measures were calculated as described for the apparent and internal validation, including measures of calibration (calibration slope, CITL, O/E ratio, calibration plots with calibration curves) discrimination (c-statistic), and measures of overall model fit (R^2 or Nagelkerke's pseudo R^2 for binary outcomes), in each imputed dataset separately, and combined across imputations as described above.

3.4 Results

3.4.1 Study populations

Table 3.3 summarises the numbers of participants available for each analysis, being those with complete outcome data recorded at six months. In the model development data, pain intensity score was recorded at six months in a total of 545 participants, with 240 (44%) in high pain at that time. The external validation data included 586 NLBP patients overall, 485 of whom had complete outcome data, with just over half of these experiencing high pain at six months, at 56.7% (n=275).

Table 3.3: Numbers of participants and events available for each analysis, for participants with complete outcome data. Values are number (percentage) unless otherwise stated.

	Participants	Events (%)
Model development		
STarT-MSK-pilot	196	78 (39.8%)
KAPS	349	162 (46.4%)
Total (outcomes measured)	545	240 (44.0%)
External validation		
STarT-MSK-MT	485	275 (56.7%)

Development data

The KAPS cohort contained 465 eligible NLBP participants. A further 214 participants were available from the STarT-MSK-pilot data. This gave a total of 679 individuals across the two datasets, for the model development and internal validation analysis. Of these, 545 (80.3%) had pain outcome measurements at 6 months. The majority of people presented with back pain (n=563, 83%), had troublesome MSK pain in more than one part of their bodies (n=451,

67.2%), and expressed the expectation that their condition would be long-lasting ("*Do you think your condition will last a long time?*", n=469, 71.1%). The median pain intensity score at first assessment was 7 (LQ to UQ: 5 to 8) out of 10. Six months after initial assessment, median pain intensity had reduced to 4 (LQ to UQ: 2 to 7).

Few predictor variables showed substantial differences (larger than 10%) in distribution between the STarT-MSK-pilot and KAPS datasets, as can be seen in Table 3.4. Notable differences in predictor variable distributions include those for pain self-management (STarT-MSK-pilot having 67.3% "*unsure about how to manage [their] pain condition*"; KAPS only 50%); and for pain impact (TAPS with 53.1% "*bothered a lot by [their] pain*" in the preceding 2 weeks, KAPS with 72.7%).

External validation data

Data to generate predictions for pain score and probability of high pain were available for 586 patients in the external validation data, with outcome data recorded in 485 (82.8%) of these. Patients predominantly presented with back pain (78%) and had a median baseline pain score of 7 (IQR 5 to 8) out of 10. A comparison of summary values for predictor information in the external validation data and the model development data is given in Table 3.4.

Though median baseline pain intensity across the development and validation data matched at 7 (IQR 5 to 8) out of 10, pain impact ("*Over the last two weeks, have you been bothered a lot by your pain?*") was more severe in the external validation data, affecting 79.7%, compared to 65.2% of the development population. Items regarding long-term expectations ("*Do you think your condition will last a long time?*") and fear of harm ("*Do you feel it is unsafe for a person*

with a condition like yours to be physically active?") were both more commonly answered as "yes" in STarT-MSK-MT, at 82.1% and 55.4% respectively, than in the model development data (69.1% and 27.8%).

Table 3.4: Predictor measurements and outcome summaries for participants in each of the model development datasets, across both development datasets combined, and in the external validation data. Numbers are n (%) responding "Yes" unless otherwise stated)

	Model development			External validation
	STarT-MSK -pilot (n=214)	KAPS (n=465)	Total (n=679)	STarT-MSK-MT (n=586)
Baseline				
Primary pain site: "Neck"	59 (27.6)	57 (12.3)	116 (17.1)	129 (22.0)
Pain intensity, median (LQ-UQ)	7 (5 to 8)	7 (4 to 8)	7 (5 to 8)	7 (5 to 8)
Pain self-management	144 (67.3)	225 (48.4)	369 (54.3)	279 (47.6)
Pain impact	113 (52.8)	330 (71.0)	443 (65.2)	462 (78.8)
Walking short distances only	117 (54.7)	248 (53.3)	365 (53.8)	344 (58.7)
Pain elsewhere	147 (68.7)	304 (65.4)	451 (66.4)	398 (67.9)
Long-term expectations	154 (72.0)	315 (67.7)	469 (69.1)	477 (81.4)
Other important health problems	81 (37.9)	183 (39.4)	264 (38.9)	242 (41.3)
Emotional well-being	132 (61.7)	284 (61.1)	416 (61.3)	388 (66.2)
Fear of harm	56 (26.2)	133 (28.6)	189 (27.8)	322 (55.0)
Pain duration	114 (53.3)	219 (47.1)	333 (49.0)	345 (58.9)
Outcome				
Missing	18 (8.4)	116 (24.9)	134 (19.7)	101 (17.2)
Complete	196 (91.6)	349 (75.1)	545 (80.3)	485 (82.8)
Pain intensity, median (LQ-UQ)	3 (1 to 6)	4 (2 to 7)	4 (2 to 7)	4 (1 to 7)
Event (high pain)	78 (39.8)	162 (46.4)	240 (44.0)	275 (56.7)
No event	118 (60.2)	187 (53.6)	305 (56)	210 (43.3)

LQ - Lower Quartile, UQ - Upper Quartile.

3.4.2 Prediction models equations for continuous and binary outcomes

Table 3.5: Final prediction models for 6-month pain, after optimism adjustment. Numbers are intercepts (α) and coefficients (β) and for continuous outcome models, intercepts (α) and odds ratios ($exp(\beta)$) and for binary outcome models. Uniform shrinkage factors for each model were obtained through bootstrapping with 1000 replications.

	Linear regression: Pain score (coefficient)	Logistic regression: High pain (odds ratio)
Pain intensity	0.269	1.26
Pain self-management	0.212	1.20
Pain impact	0.632	1.40
Walking short distances only	0.934	2.37
Pain elsewhere	0.278	1.33
Long-term expectations	1.673	3.61
Other important health problems	0.578	1.38
Emotional well-being	-0.002	1.28
Fear of harm	-0.434	0.63
Pain duration	1.129	2.19
Pain site	0.515	1.02
Constant	-1.153	-4.324
Var(constant)	0.132	0.041
Shrinkage factor	0.982	0.938

Model coefficients after shrinkage and re-estimated intercept terms for each pain intensity model are presented in Table 3.5.

Conditional on other variables in the model, a patient's baseline pain intensity and their long-term expectations (thinking their condition would last a long time) contributed most to predictions of pain intensity for both the continuous and binary outcome models, with a higher baseline pain

intensity score and expecting their condition to last a long time being associated with both higher expected pain intensity scores and higher predicted probabilities of high pain at six months.

The direction of effect was generally consistent across the two model types, with fear of harm being the only predictor associated with a protective effect for both the continuous and binary outcomes, after adjusting for other covariates. Direction of effect differed between the two models for emotional wellbeing, although the estimated effect size in the linear regression was negligible.

Within the logistic regression model, after adjusting for other variables, the effect of primary pain site (back vs neck) was negligible, while the linear model predicted six-month pain scores that were around half a unit higher for those with the back as their primary pain site. The importance of this difference in expected pain score is likely to be patient and clinician dependant, but would be relevant to patients across the full range of pain intensities. Information on the effect of primary pain site is apparently lost with the early dichotomisation of the pain intensity outcome variable.

Generating predictions for an individual

Example (a): Demonstration of using linear regression equation to predict six-month pain score in an individual patient

For a back pain patient X with a pain intensity of 7, pain elsewhere, who thinks their condition will last a long time and has pain that has lasted for more than 6 months, pain score in 6 months time would be estimated as follows.

Predicted pain score = – 1.153 + 0.269 x (Pain intensity) + 0.212 x (Pain self-efficacy)
+ 0.632 x (Pain impact) + 0.934 x (Walking short distances only)
+ 0.278 x (Pain elsewhere) + 1.673 x (Thinking their condition will last a long time)
+ 0.578 x (Other important health problems) – 0.002 x (Emotional well-being)
– 0.434 x (fear of pain-related movement) + 1.129 x (Pain duration)
+ 0.515 x (Primary pain site)

Where:
Pain intensity is scored from 0 to 10, where 0 is "no pain" and 10 is "pain as bad as it could be"
Primary pain site is scored as 1 for patients with pain in their back, and 0 for neck
Other variables are scored 1 if the patient answered "yes" to that question, and 0 otherwise

So for patient X,

Predicted pain score = – 1.153 + 0.269 x (7) + 0.212 x (0) + 0.632 x (0) + 0.934 x (0)
+ 0.278 x (1) +1.673 x (1) + 0.578 x (0) – 0.002 x (0) – 0.434 x (0)
+ 1.129 x (1) + 0.515 x (1)

= – 1.153 + (0.269 x 7) + 0.278 + 1.673 + 1.129 + 0.515

= 5.5 out of 10

Example (b): Demonstration of using linear regression equation to predict the probability of high pain at six months in an individual patient

For a back pain patient X with a pain intensity of 7, pain elsewhere, who thinks their condition will last a long time and has pain that has lasted for more than 6 months, pain score in 6 months time would be estimated as follows.

Predicted pain score = 5.5 out of 10, as seen in the previous box, therefore:

Probability of high pain in 6 months $= 1 - P\left(Z < \frac{C - Y_{PREDi}}{\sigma_{model}}\right)$

$= 1 - P\left(Z < \frac{5 - 5.5}{2.293}\right)$

$= 1 - P\left(Z < -0.218\right)$

$= 1 - 0.414$

Probability of high pain in 6 months = 59%

Example (c): Demonstration of using logistic regression equation to predict the probability of high pain at six months in an individual patient

For a back pain patient X with a pain intensity of 7, pain elsewhere, who thinks their condition will last a long time and has pain that has lasted for more than 6 months, the probability of high pain in 6 months time would be estimated as follows.

Probability of high pain in 6 months = $\frac{\exp(LP)}{1+\exp(LP)}$

LP = -4.324 + 0.231 x (Pain intensity) + 0.182 x (Pain self-efficacy)
+ 0.336 x (Pain impact) + 0.863 x (Walking short distances only)
+ 0.285 x (Pain elsewhere) + 1.284 x (Thinking their condition will last a long time)
+ 0.322 x (Other important health problems) + 0.247 x (Emotional well-being)
– 0.462 x (fear of pain-related movement) + 0.784 x (Pain duration)
+ 0.020 x (Primary pain site)

Where:
exp is the exponential function
Pain intensity is scored from 0 to 10, where 0 is "no pain" and 10 is "pain as bad as it could be"
Primary pain site is scored as 1 for patients with pain in their back, and 0 for neck
Other variables are scored 1 if the patient answered "yes" to that question, and 0 otherwise

So for patient X,

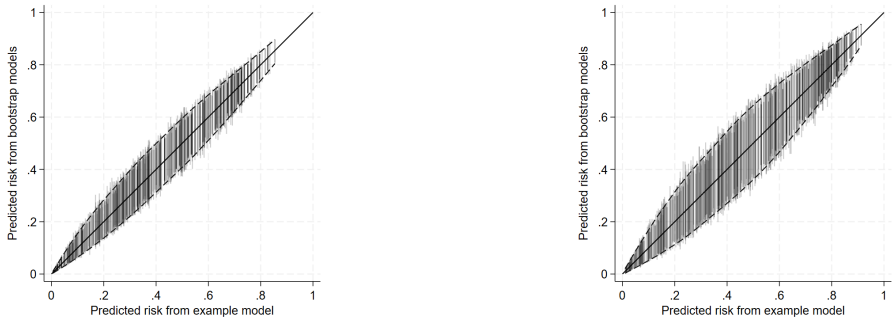
LP = -4.324 + 0.231 x (7) + 0.182 x (0) + 0.336 x (0) + 0.863 x (0)
+ 0.285 x (1) + 1.284 x (1) + 0.322 x (0) + 0.247 x (0)
– 0.462 x (0) + 0.784 x (1) + 0.020 x (1)

= -4.324 + (0.231 x 7) + 0.285 + 1.284 + 0.784 + 0.020
= - 0.33

Probability of high pain in 6 months = $\frac{\exp(-0.33)}{1+\exp(-0.33)} = \underline{42\%}$

3.4.3 Model stability checks

Prediction instability plots, plotting the predicted risk of being in high pain at six months based on the example model (developed in the original sample) against risk estimates for that same individual from models developed across 200 bootstrap samples, are shown for both the linear and logistic regression models in Figure 3.3. Dashed lines denote the region where 95% of the predicted risks across bootstrap models lie.



(a) Linear model

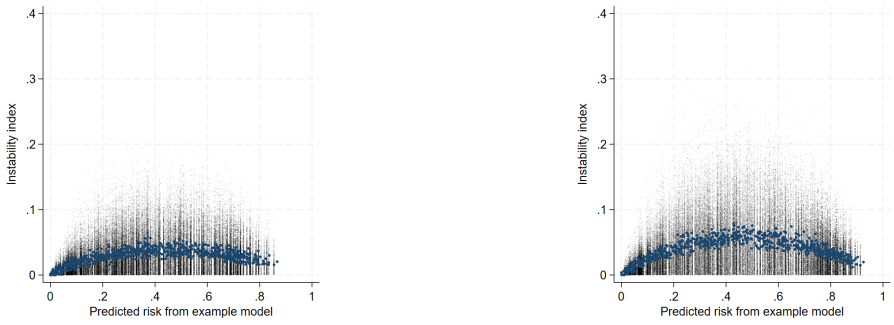
(b) Logistic model

Figure 3.3: Prediction instability plots, for expected risks of high pain at six months generated with pain score dichotomised before or after modelling

Predicted outcome risks from the linear regression approach were relatively stable, with a narrow spread of values across bootstrap models for each individual. This was evident across the full range of possible predicted risks (from 0 to 1). The probability estimates from the linear regression model were notably more stable than those from the logistic regression, where the spread of predicted probabilities across bootstrap samples is slightly wider. This can be seen both through the spread of individual markers, and by the width of the 95% uncertainty bands (dashed lines).

While the logistic model was relatively stable at the extremes, suggesting more certainty in predictions for those at especially high or low risk of high pain, for those with less extreme predicted risks had higher instability in their predictions across bootstrap models. For example, given a predicted risk of around 0.5 from the original logistic model, 95% of the bootstrap models gave predictions between 0.36 and 0.65. The corresponding bootstrap interval for probabilities from the linear regression model spans from 0.39 to 0.60.

Figure 3.4 shows the absolute difference between predicted risks from the original model and each of the bootstrap models, as a scatter against the value of the predicted risk from the original model. Highlighted points (in navy) show the mean difference between these risk estimates across all bootstrap models, for each individual (known as the instability index).



(a) Linear model

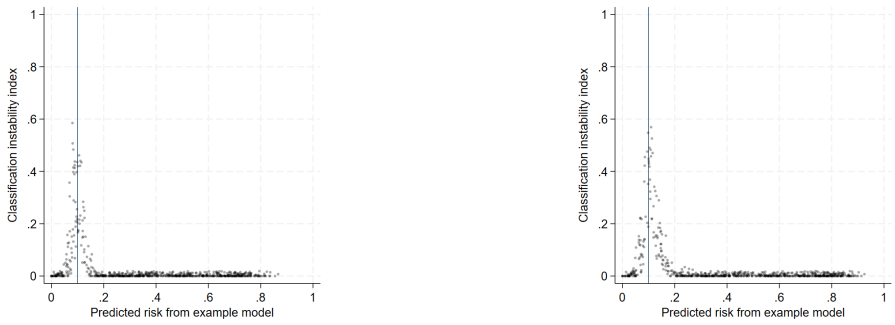
(b) Logistic model

Figure 3.4: Instability index plots, for expected risks of high pain at six months generated with pain score dichotomised before or after modelling

Mean absolute differences between predicted risks from the original and bootstrap models was lower for the linear model when summarised across all bootstrap models, with a median instability index across all individuals of 0.026 (0.020 to 0.031), compared to 0.036 (LQ to UQ: 0.025 to

0.045) for the logistic model. This is demonstrated in Figure 3.4, where the mean values for individuals are lower for the linear model, and the spread of values across bootstrap models is visibly narrower.

The predicted risks from a model may be used to classify patients as high or low risk of high pain at six months, to then inform an alteration in treatment course for those who are most likely to benefit from it. Where a patient's predicted risk of high pain varies across bootstrap models, so might their classification to either high or low risk, especially where their predicted risk from the original model lies close to the threshold being used to assign groups. Figure 3.5 shows the proportion of bootstrap models from which an individual is assigned a different classification that was given by the original model, plotted against their predicted risk from the original model, for an example threshold probability of 0.1 (i.e., when using a 10% threshold for categorising "high risk").

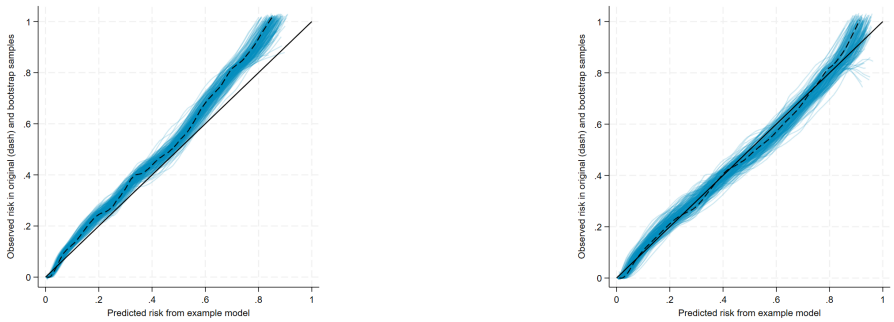


(a) Linear model (b) Logistic model

Figure 3.5: Classification instability plots, for expected risks of high pain at six months generated with pain score dichotomised before or after modelling

When considering a threshold of 0.1, consistency in classification was similar across bootstrap models for both the linear and logistic regression approaches. Those very close to the threshold (those most vulnerable to changes in classification) based the linear model had slightly higher probabilities of a different classification across bootstrap models, peaking at 58% (compared to 57.5% from the logistic model), though the range of individuals affected by classification changes was slightly narrower for the linear model. Figure 3.5 demonstrates how classification changes were seen for those with predicted risks close to the threshold (0.1) for both the linear and the logistic model.

While individual-level predictions were slightly more stable across bootstrap models for predicted probabilities from the linear regression model (Figure 3.3), this was not reflected in increased stability in the model calibration curve (Figure 3.6). Smoothed calibration curves are similarly consistent across bootstrap models for both the linear and logistic regression approaches.



(a) Linear model

(b) Logistic model

Figure 3.6: Calibration instability plots, for expected risks of high pain at six months generated with pain score dichotomised before or after modelling

As was seen in the calibration curve for the original model (Figure 3.10), some miscalibration was

evident in the predicted risks from the linear model across bootstrap models. In contrast, the logistic model appeared to be well-calibrated across the full range of predicted risks, with stability in calibration across bootstrap models similar to that of the less well calibrated linear model.

3.4.4 Comparing prediction distributions

Distributions of the predictions from the two model types are summarised in Table 3.6, for values when the models were applied to individuals in the development and external validation datasets separately. Histograms of predicted pain scores and probabilities of high pain (calculated from both the linear and logistic regression models) are shown in Figure 3.7, with consistent distributions for each prediction type across the model development and validation datasets on visual inspection.

Notably, there is a tendency towards lower predicted probabilities from the linear regression model on average, when compared to those from the logistic regression, as can be seen when comparing the mean and median values, and the minimum and maximum prediction values, shown in Table 3.6. However, on visual inspection (Figure 3.7) the shapes of the distributions were similar.

Probabilities of high pain calculated from the linear and logistic regression models were highly correlated across individuals, as can be seen in Figure 3.8. Probabilities across the two modelling types are most consistent at the extremes (for those with particularly high or particularly low probabilities of high pain), and are notably less consistent around the centre of the distributions.

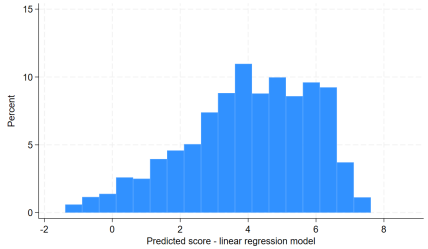
Where differences occurred, predicted probabilities from the linear regression model were generally lower than those from the logistic regression, suggesting this modelling method gives more

conservative estimates of risk. For example, for those with a predicted probability of around 0.5 from the logistic regression model, linear regression estimates varied between 0.3 and 0.5.

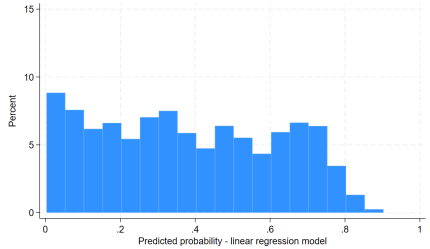
Table 3.6: Prediction distribution summary for models predicting six-month pain outcomes, after application of shrinkage, when applied in the model development and external validation data.

	Linear regression model		Logistic regression model
	Pain score	Probability of high pain	Probability of high pain
Model development data			
Mean	4.02	0.38	0.435
Standard deviation	1.868	0.239	0.261
Median	4.15	0.355	0.421
LQ to UQ	2.835 to 5.495	0.173 to 0.585	0.212 to 0.678
Minimum to maximum	-1.387* to 7.486	0.003 to 0.861	0.007 to 0.91
Skew	-0.524	0.118	0.02
Kurtosis	2.665	1.809	1.74
External validation data			
Mean	4.404	0.427	0.483
Standard deviation	1.687	0.226	0.246
Median	4.586	0.428	0.512
LQ to UQ	3.383 to 5.658	0.24 to 0.613	0.285 to 0.697
Minimum to maximum	-0.672* to 8.119	0.007 to 0.913	0.017 to 0.946
Skew	-0.734	-0.133	-0.255
Kurtosis	3.159	1.959	1.917

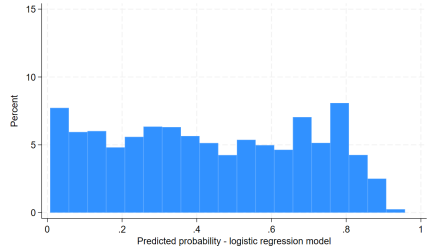
*note: outside the valid range of pain intensity scores (0 to 10)



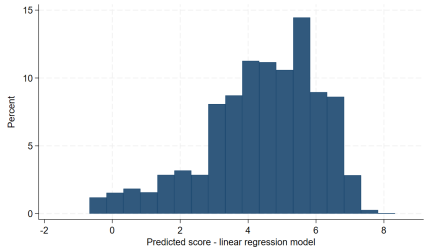
(a) Linear model, score - development data



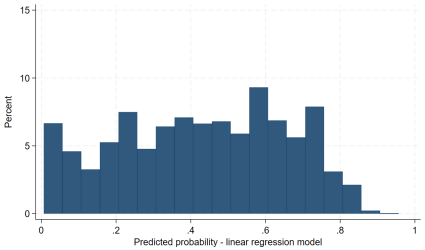
(b) Linear model, probability - development data



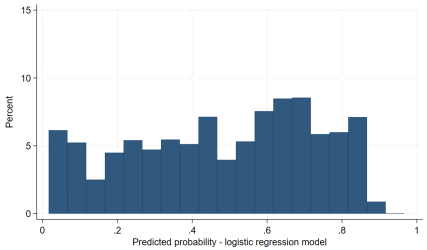
(c) Logistic model, probability - development data



(d) Linear model, pain score - validation data

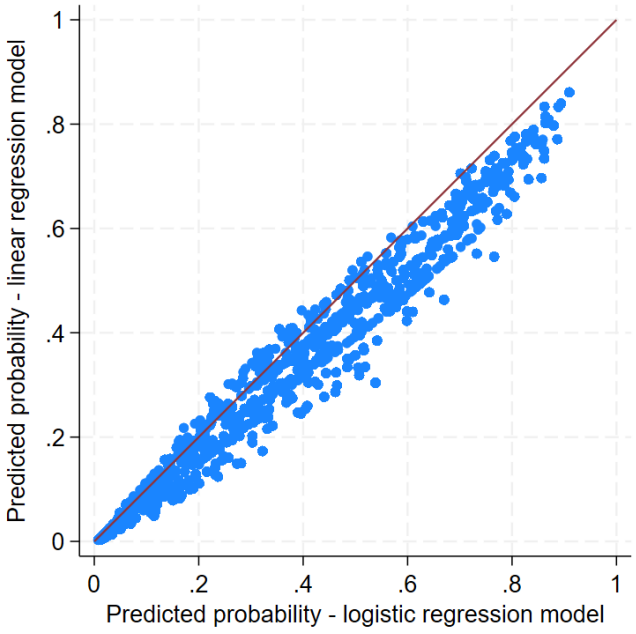


(e) Linear model, probability - validation data

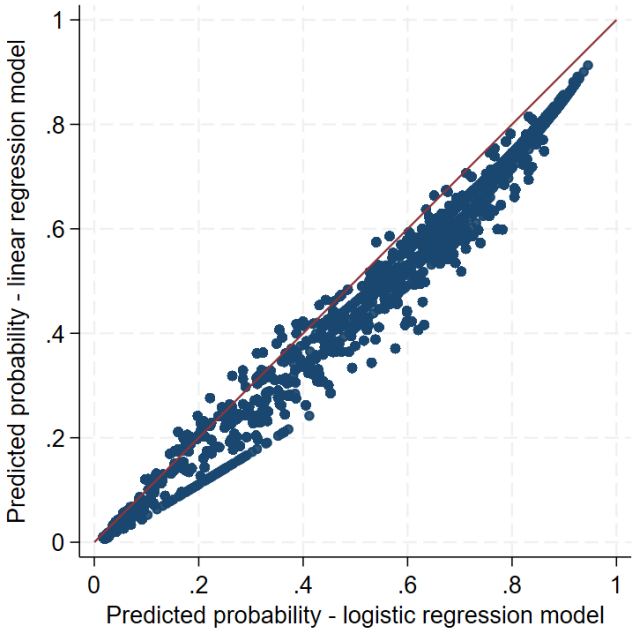


(f) Logistic model, probability - validation data

Figure 3.7: Histograms of prediction distributions in (a-c) model development and (d-f) external validation data



(a) Model development data



(b) External validation data

Figure 3.8: A comparison of individuals' probabilities of high pain when calculated using the logistic and linear regression models in the (a) model development and (b) external validation data

3.4.5 Predictive performance on internal validation

Predictive performance measures from the internal validation of the linear regression model for pain score and the logistic regression model for high pain, before and after optimism adjustment, are presented in Table 3.7.

On internal validation through bootstrapping, models for predicting pain outcomes showed reasonable calibration, as would be expected. The optimism-adjusted calibration slope for the continuous and binary outcome models were 0.982 and 0.944 respectively, indicating overfitting was more of a concern for the model predicting high pain as a binary outcome, though was small in both models. For the continuous outcome model, the shrinkage factor close to one suggests minimal overfitting to the development data, while slightly more adjustment for overfitting is required for the binary outcome model - potentially a result of the information lost on outcome dichotomisation and thus the smaller effective sample size available for analysis.

For predicted probabilities of high pain from the logistic model, discriminative ability was reasonable. The optimism-adjusted C-statistic of 0.810 indicates approximately an 81% probability that a randomly selected individual with high pain at six months would receive a higher predicted probability from the model than a randomly selected individual without high pain would.

Table 3.7: Predictive performance of prediction models on internal validation using bootstrapping, before and after optimism adjustment

Measure		Linear regression model	Logistic regression model
		Continuous outcome	*Binary outcome*
Calibration slope	Apparent	1.000 (0.895 to 1.105)	1.000 (0.811 to 1.177)
	Average optimism	0.018 (0.017 to 0.018)	0.056 (0.055 to 0.057)
	Optimism adjusted	0.982	0.944
CITL	Apparent	0.000 (-0.193 to 0.193)	0.000 (-0.165 to 0.237)
	Average optimism	-0.547 (-0.551 to -0.544)	0.329 (0.323 to 0.334)
	Optimism adjusted	0.547	-0.329
O/E	Apparent	1.000 (1.000 to 1.000))	1.000 (0.997 to 1.003)
	Average optimism	0.133 (0.132 to 0.134)	0.014 (0.012 to 0.016)
	Optimism adjusted	1.133	1.014
C-statistic	Apparent	-	0.811 (0.775 to 0.847)
	Average optimism	-	0.01 (0.01 to 0.01)
	Optimism adjusted	-	0.810
R^2/Pseudo R^2	Apparent	39.1% (38.8% to 39.2%)	37.5% (37.3% to 37.7%)
	Optimism adjusted	0.37	0.33

Good calibration would be concluded for the linear regression model to predict the continuous outcome when considering only the calibration slope value, though the calibration plot (Figure 3.9) shows that model calibration varied considerably across individuals for the continuous outcome predictions. For example, for people with a predicted six-month pain intensity score of 4 (on the 0-10 scale), observed scores ranged from 0 to 10.

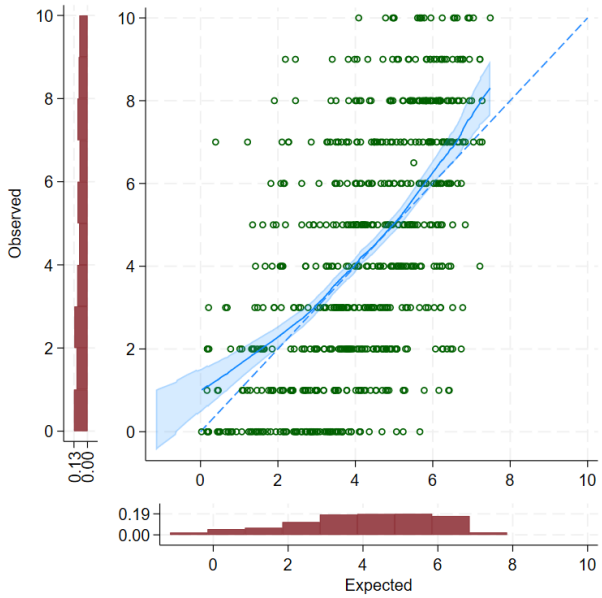
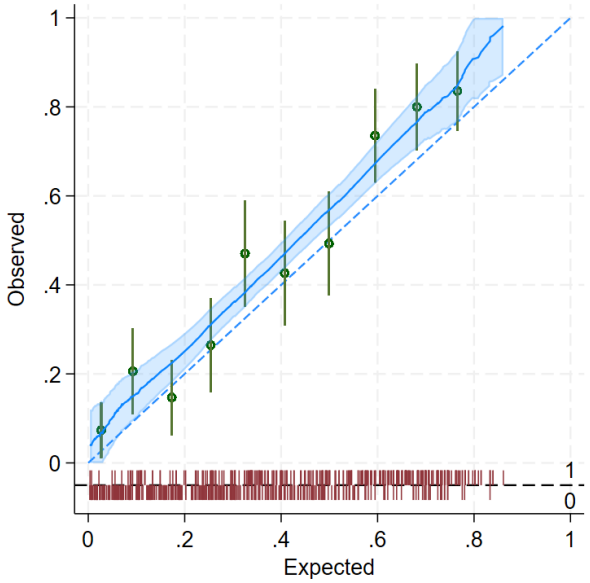
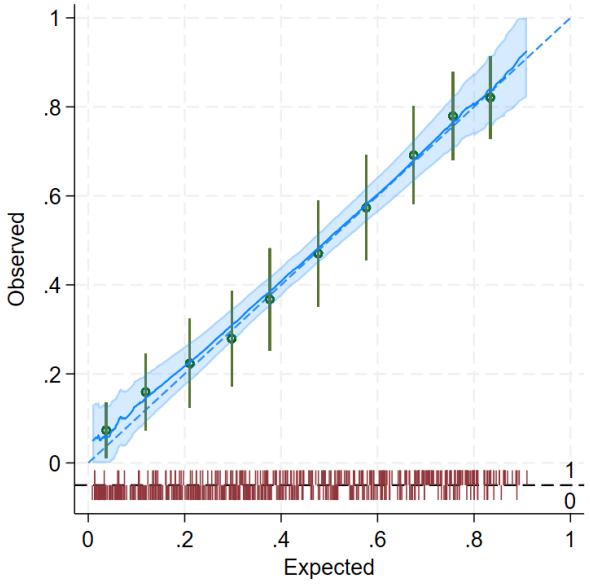


Figure 3.9: Apparent calibration of the model to predict six-month pain intensity score, after shrinkage

When assessing predicted probabilities generated from the linear model (after shrinkage) some miscalibration was evident, with expected risks consistently too low (when compared to observed risks) across the whole range of predicted probabilities, as shown in Figure 3.10. This miscalibration was not present in the logistic model, where predicted probabilities were generally higher than their corresponding values from the linear model (see Figure 3.8, above).



(a) Linear model - probability



(b) Logistic model - probability

Figure 3.10: Apparent calibration of predicted probabilities for high pain at six months, after shrinkage, with predictions generated for pain score dichotomised before or after modelling

3.4.6 Predictive performance on external validation

Table 3.8: Performance of prediction models for continuous and binary pain intensity outcomes on external validation

Measure	Linear regression model *Continuous outcome*	Linear regression model *Binary outcome*	Logistic regression model *Binary outcome*
Calibration			
Calibration slope	0.735 (0.656 to 0.815)	0.854 (0.665 to 1.043)	0.710 (0.598 to 0.823)
CITL	-1.262 (-1.408 to -1.116)	0.027 (-0.157 to 0.211)	-0.307 (-0.438 to -0.176)
O/E*	-	1.012 (0.983 to 1.043)	0.880 (0.854 to 0.908)
Discrimination			
C-statistic*	-	0.734 (0.691 to 0.771)	0.721 (0.692 to 0.749)
Overall model fit			
R^2/Pseudo R^2**	20.8% (20.2% to 21.3%)	23.3% (22.0% to 24.2%)	22.1% (21.1% to 23.1%)

*For binary outcome assessments only

**Not summarised using Rubin's rules: values are median (LQ to UQ) across imputations

Pain intensity score at six months

The predictions for pain intensity score on its continuous scale showed poor calibration on average across the validation population, with the calibration plot (shown in Figure 3.11) suggesting predicted scores were too high on average, with a lot of individual-level variation in accuracy. The CITL value confirmed that the continuous outcome model systematically over-predicted pain intensity score at six months by an average of 1.2 points. The calibration slope value of 0.74 (0.66 to 0.82) further suggests that the model does not fit the external validation data well, perhaps due to population or temporal differences, despite the application of shrinkage at the model development stage.

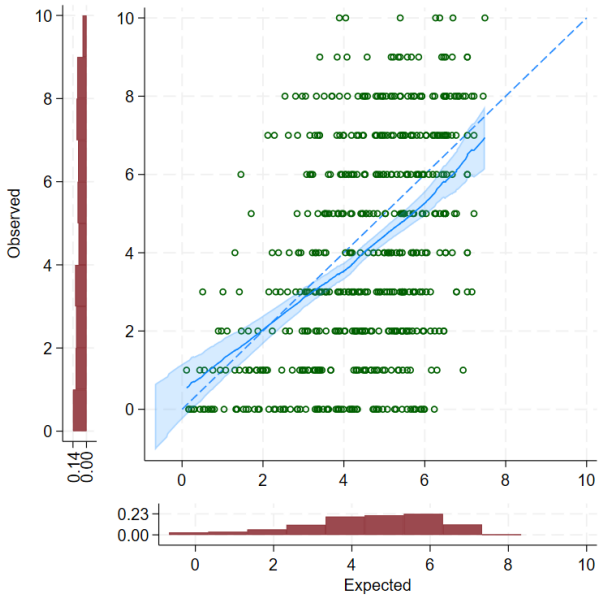


Figure 3.11: Calibration plot for the model to predict six-month pain intensity score on external validation

Probability of high pain at six months

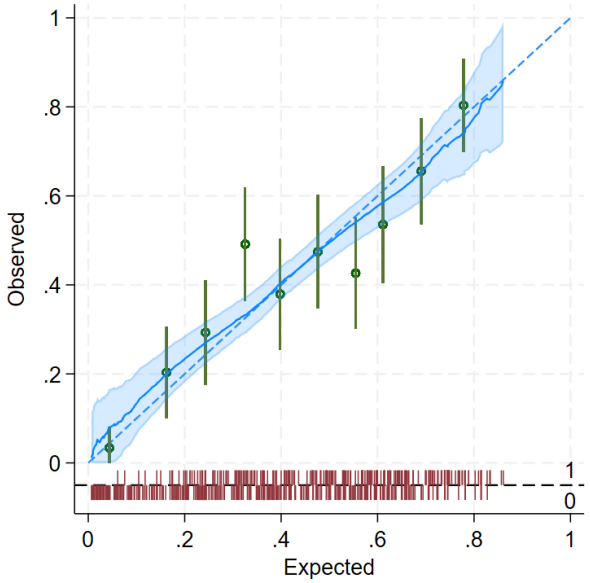
Calibration performance was reasonable in both the linear and logistic models to predict the binary outcome of high pain at six months, as can be seen in Figure 3.12, where calibration curves indicate good calibration on average across the external validation population. The calibration slope values of 0.854 (0.665 to 1.043) and 0.710 (0.598 to 0.823) for predicted probabilities from the linear and logistic regression models respectively suggest both models gave predictions that were slightly too low for those at low risk, and slightly too high for those at high risk.

The observed/expected ratio for the linear model, 1.012 (0.983 to 1.043), implied minimal miscalibration-in-the-large, while the corresponding value for the logistic model, 0.880 (0.854 to 0.908), was indicative of an over-prediction of risk, on average, in the external calibration data.

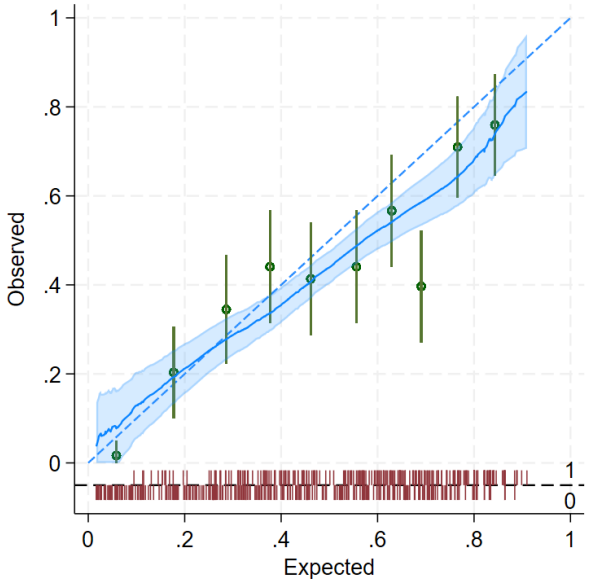
Given the under-prediction of risk from the linear model in the model development data, it is unclear whether this better performance on external validation is due to the tendency of the linear model to predict lower risks than the logistic model, and an external validation in a population that coincidently had a lower baseline falls risk that the development population.

Discrimination performance was very similar across model types (see Table 3.8). C-statistics of 0.73 (0.69 to 0.77) and 0.72 (0.69 to 0.75) from the linear and logistic regression models respectively suggest that around 73% and 72% of concordant pairs were correctly identified by these models. Confidence intervals for these values were also highly consistent across model types, with a very slightly wider confidence interval for the value from the linear regression model (an artefact of the use of the delta method in the estimation of the standard errors for the C-statistic).

A slightly higher value of Nagelkerke's R^2_N for predicted probabilities generated using the linear regression model (median 23.3%, LQ to UQ 22.0% to 24.2%) compared to the logistic regression model (22.1%, 21.1% to 23.1%) indicates a higher proportion of the variance in the outcome is explained by the model when applied in the external validation data. While Nagelkerke's R^2_N values for the predicted probabilities are on the same 0-1 scale as the R^2 for the continuous model, caution should be taken when comparing the two, as their calculation methods clearly differ. Nevertheless, a tentative comparison suggests that the overall model performance of the binary outcome models was reasonably consistent with, if not slightly better than, that of the continuous outcome model.



(a) Linear model - probability



(b) Logistic model - probability

Figure 3.12: Calibration performance on external validation for models to predict the probability of high pain at six months, when generated from linear and logistic regression models

3.5 Discussion

This chapter discussed the impact of outcome dichotomisation on the development and external validation of clinical prediction models. The comparison focussed on an applied example aiming to improve communication and treatment matching for patients consulting with NLBP, and aimed to develop models based on pre-defined, clinically important predictor variables. These prediction models were to estimate an individual's predicted pain intensity score, on a continuous scale, and the probability that they will be experiencing high pain (defined by a simple dichotomisation of this same pain score), six months after their initial consultation in primary care. The development and external validation of these models were published in *Physical Therapy* [91].

3.5.1 Summary of key findings

The first part of the methods for the chapter demonstrated a proposal for how predicted probabilities, if desired to facilitate clinical decision making, can easily be generated based on the output from a linear regression model. These calculations can be applied post-modelling, meaning there is no need to dichotomise a continuous outcome variable beforehand. In the applied NLBP example, predicted probabilities generated from a linear regression model, using the methods demonstrated here, were highly correlated with those stemming from a logistic regression model for the same outcome, though the weighting and importance of predictors varied slightly across model types.

Assessments of stability at model development suggested that modelling the continuous outcomes on its continuous scale and dichotomising after the modelling stage results in more stable predictions than a logistic model developed on the pre-dichotomised outcome. This is likely due

to the more efficient use of the data when modelling on the continuous scale, resulting in a larger effective sample size for model development. This increase in stability was seen in individual-level predictions, with lower instability index values for risk predictions from the linear regression model. Increased stability was also seen for risk group classification, when using a probability threshold of 10% to determine those at high risk.

Crucially, internal validation analyses suggested a higher level of overfitting in the logistic regression model, as shown by an optimism-adjusted calibration slope that was further from the ideal value of one. This is likely due to the loss of information when the continuous pain score outcome was dichotomised prior to modelling. Some miscalibration was evident on visual inspection of calibration plots for the predicted risks from the linear model, both when considering the apparent performance of the shrunken model and on calibration stability assessment. Though methods to recalibrate predicted probabilities from the linear regression model might result in better calibration in the model development data, this would require fixing the cut-off value to a single point, as with the logistic model on the pre-dichotomised outcome, thus isn't ideal. Instability in the calibration curve for predicted probabilities from the linear regression model was similar to that of the better calibrated logistic model, where predicted risks better matched the observed outcomes in the model development data.

Predictive performance of the continuous and binary outcome models was reasonably consistent on external validation, though the continuous outcome model gave values closest to the ideal for all measures of calibration. Predicted probabilities from the linear regression model were better calibrated in the external validation data than the logistic regression model, with some over-prediction of falls risk evident in predictions from the logistic model. However, this may be

related to a combination of more conservative estimates of risk from the linear model seen in the model development population, and a slightly lower risk population used for external validation. Further assessment of the two modelling types over a wide range of scenarios, in different clinical contexts and in simulated data, are needed before firm conclusions can be made surrounding the improved calibration performance on external validation.

A tentative comparison of R^2 values from the continuous model to the Nagelkerke's R^2 values from corresponding binary model suggested similar overall model fit for both outcome types, with the binary outcome model showing a slightly higher (pseudo) R^2. These measures, however, are calculated differently for the two modelling types, so caution should be taken in the comparison and interpretation of these results. When making the more valid comparison of Nagelkerke's R^2 values for predictions from the linear and logistic regression models to predict high pain probabilities, the linear regression model performed better.

3.5.2 Strengths and limitations

Despite differences in outcome definitions, predictors of poor prognosis in both linear and logistic regression models were consistent with what has previously been reported. In particular, baseline pain intensity [183] and long-term expectations ("*Do you think your condition will last a long time?*") [184] were the strongest predictors in both of the models developed in this chapter, after adjusting for other covariates, and have previously been reported as important for the prediction of NLBP progression. The differing direction of effect seen for emotional wellbeing ("*Has pain made you feel down or depressed in the last two weeks?*") between the linear and logistic models, along with the negligible effect size in the linear model, is in contrast to previously reported associations

of of both depression [183] and low mood [185] with NLBP prognosis. This difference, however, may be related to adjustment for other covariates, which were consistent between the two models developed here, as no statistical selection of predictors was conducted. Thus the impact of mood may have been absorbed into the coefficients of other predictors, such as bothersomeness or pain intensity.

The available data used in the applied example for this chapter provided a sufficient sample size for the model development analyses, per current guidance. To predict the continuous outcome of pain intensity score, data comfortably exceeded the recommended 354 participants recommended for model development [75]. For the binary outcome analyses, patients with high pain were insufficiently common to exceed the requirement for 412 events and non-events based on an assumed prevalence of 50% (the observed high pain proportion was 44% at six months) [76]. The demonstrated differences in model stability and performance have not been assessed in data that was also sufficient for the binary outcome model, though the requirement for more data to minimise overfitting (and possibly increase stability) in the logistic regression model is in itself of interest.

The sample size available for external validation was far below what is currently recommended for the external validation of a prediction model with a binary outcome. To meet rule-of-thumb recommendations [172] at the time of analysis, a minimum of 200 high pain events (and non-events) were required to externally validate the binary outcome model. Though this was exceeded in the external validation sample, rules of thumb are now not recommended in sample size calculations for the validation or clinical prediction models. When considering tailored sample size calculations for the external validation in this clinical example, at least 1946 (1071 events) were required [94].

The external validation in a UK population, one very similar to that used for model development, showed systematic over-prediction of pain intensity score at six months. Given the small sample size and poor calibration on external validation, albeit with reasonable discrimination performance for prediction of the dichotomised outcome, this chapter can not demonstrate how well conclusions might extend to situations where the external validation sample is large, and continuous outcome is well predicted in the external population.

In this example, the modelling methods used were relatively uncomplicated, thus findings may not extend to scenarios involving more complicated analysis methods, such as variable selection, assessment of non-linear trends, or differing outcome distributions. Future work, including more generalisable simulation studies, will investigate stability and performance of clinical prediction models for dichotomised continuous outcomes when using these more complex modelling strategies, though this research will not form a part of this thesis.

3.5.3 Conclusions and next steps

This chapter has illustrated the development of prediction models when modelling discrete outcomes in continuous and binary forms, and how this choice of outcome can effect predictive performance statistics and model usability on external validation. Prediction models that fail to keep continuous outcomes on their continuous scale may suffer from loss of information and reduced power to detect predictor effects, with this applied example demonstrating a larger degree of overfitting in a logistic model compared to the linear regression model for the same outcome. This difference in overfitting to the development data is possibly due to the loss of information that arose when the continuous outcome was split.

Further, this chapter demonstrated how predicted probabilities, if desired, can be generated post-modelling from the outputs of a linear regression model. Thus, in most cases, dichotomisation prior to modelling is unlikely to be necessary. However, the methods used to gain predicted probabilities from the linear regression model were highly dependent on this model's underlying assumptions: namely the requirement for normally distributed residuals, and for constant error terms across different values of observed outcome (homoscedasticity). Further research should investigate methods to generate such probabilities from a continuous outcome prediction, where data do not meet these modelling assumptions and thus linear regression may not be appropriate. This extension is beyond the scope of this thesis.

As mentioned above, both model development and external validation analyses in this chapter took place in pre-existing datasets of fixed size. While published recommendations were available to assess whether the model development data contained enough participants (continuous outcome models) and enough high pain and poor function events (binary event models), only rules of thumb were available to assess the appropriateness of the size the external validation sample for accurate assessment of model predictive performance. The literature was lacking in close-form or simulation-based methods to calculate the minimum sample size required to ensure precise estimation of the model calibration (calibration slope, calibration-in-the-large), discrimination (c-statistic), and overall model fit (R^2, pseudo R^2) on external validation. Without sufficient external participants on whom to test the model performance, validation may lead to imprecise estimates with wide confidence intervals, giving researchers very little confidence in the generalisability or transportability of the prediction model.

The next chapter, therefore, discusses the requirements of a suitably sized sample to externally validate a clinical prediction model with a continuous outcome. Sample size guidance is developed to ensure sufficiently precise estimation of those key performance statistics when predicting a continuous outcome, including proposed new, closed-form calculation methods to facilitate easy application. These sample size recommendations are applicable, regardless of clinical area, whenever researchers might need to identify the minimal sample size to ensure a desired precision in predictive performance estimates.

CHAPTER 4

Minimum sample size for external validation of a clinical prediction model with a continuous outcome

4 Chapter 4: Minimum sample size for external validation of a clinical prediction model with a continuous outcome

4.1 Introduction and objectives

Research involving clinical prediction models uses data from a sample of the population to develop or validate methods to provide individualised outcome predictions to help inform clinical decision making and patient counselling [186]. Recent methodological work has focussed on estimating the necessary size of the sample needed to develop a prediction model, to achieve a minimum level of precision in the estimation of model parameters and performance metrics [78]. Chapter 3 demonstrated an example where pre-existing datasets contained information on sufficient participants to meet minimum recommendations for the development of models to predict the outcome of pain intensity score on both a continuous [75] and dichotomised [76] scale.

Once a model has been developed, evaluation of its predictive performance in new data is often crucial, in a process known as external validation [57, 58]. A model's predictive performance upon external validation can indicate how well the model performs in new individuals from the target population for use of the model in practice. The sample size and techniques used to develop the model are of little importance if it is shown that the model performs well in new data [186]. Despite being widely encouraged, with clear evidence of its importance, external validation of published prediction models is rare in practice [58, 5]. Even where external validation is performed, the sample size is often too small to provide reliable conclusions and key measures of predictive performance, such as calibration, are often neglected [187].

The sample size required to suitably evaluate a prediction model during external validation is not

yet known. The previous chapter demonstrated an external validation of the developed models, conducted as a secondary analysis of a pre-existing dataset recruited for a different purpose. At the time of the Chapter 3 analysis, no methods existed to determine *a priori* whether the number of participants in this dataset was sufficient to ensure precise estimation of performance measures, though confidence intervals for these measures were notably wide in some cases.

Therefore, in this chapter, I discuss criteria that could be used to inform the minimum sample size needed for external validation of a clinical prediction model, to gain precise estimates of key performance measures. The focus is on evaluating predictions of continuous outcomes (such as birthweight or pain score, as discussed in Chapters 2 and 3 respectively), for which the modelling approach typically follows a linear regression framework. Such models provide an equation to predict the continuous outcome value, either on its original scale or following some transformation, conditional on the values of one or more predictor variables. The outcome may relate to something current, such as current levels of fat mass (an example that will be discussed further throughout this chapter) or in the future, such as pain scores six months after a consultation for NLBP (as in the previous chapter).

In the case of a continuous outcome prediction, the sample size needs to be large enough to precisely estimate calibration and overall model fit. Three key measures of predictive performance are targeted: calibration slope (agreement between predicted and observed values across the range of predicted values), calibration-in-the-large (agreement between predicted and observed outcome values on average), and R^2 (the proportion of variance explained). These performance measures are first introduced, and then closed-form solutions are derived for the sample size required to estimate each one of them precisely. As all of these solutions depend on an estimate of the variance

of observed outcome values, a fourth criterion is also suggested, aiming to ensure this variance is estimated precisely.

Thus, the final sample size calculation proposed in this chapter comprises checking four criteria, and concluding the minimum required for the external validation (to meet all four criteria) as the largest sample size calculated across the four approaches. Demonstrations of the calculation approach are shown in an applied example using a model to predict fat-mass in children and adolescents. This sample size proposal, along with the applied example in predicting fat-free mass, have been published in *Statistics in Medicine* [92].

4.2 Key measures of predictive performance

Suppose that a clinical prediction model for a continuous outcome (Y_i) has been developed, presented as a linear regression equation:

$$Y_{PREDi} = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots . \quad (1)$$

A model of this form allows predicted values to be generated for new individuals, outside of the development dataset. That is, the predicted (expected) outcome value Y_{PREDi} can be obtained for a new individual, i, by calculating the right-hand side of the equation, which utilises an intercept term (α), individual i's values for the included predictors (X_{1i}, X_{2i}, X_{3i}, etc.), and the corresponding predictor effect estimates ($\beta_1, \beta_2, \beta_3$, ...). For example, if X_{1i} were a continuous predictor (such as age or blood pressure) with a simple linear association with the outcome, then β_1 represents the change in Y_{PREDi} for each 1-unit (year, mmHg) increase in the value of X_{1i}, after adjusted for other predictors in the model. Equally, if X_{2i} were a binary predictor (such as treatment group), denoted by 0 for category A (placebo) and 1 for category B (drug), then β_2 represents the expected change in Y_{PREDi} for those in category B compared to category A (drug compared to placebo), having adjusted for other included predictors.

This chapter focuses on a situation where a prediction model such as that given in Equation 1 is already fully defined, based on a previous model development study, and an external validation is required to evaluate the accuracy of the model's predicted values (Y_{PREDi}) in new individuals. External validation studies must obtain a dataset of new individuals from the population of interest, with each individual providing values for all necessary predictors ($X_{1i}, X_{2i}, X_{3i}, \dots$) included in the model. The new data must also contain each individual's observed outcome value (Y_i). The former allows Y_{PREDi} to be calculated for each individual by applying the prediction

model equation, and the latter means that the model's predictive performance can be quantified, by comparing the predicted Y_{PREDi} to the observed outcome values, Y_i.

There are three key statistics to quantify the predictive performance of a model with a continuous outcome upon external validation, which focus on the model's calibration and overall fit.

R-squared R^2 is the proportion of variation in the observed Y_i values that is explained by the prediction model (the fitted linear regression model), and represents a measure of overall model fit. Let R^2_{val} denote the R^2 of an existing prediction model when examined in an external validation dataset. R^2_{val} gives a measure of the proportion of variation in Y_i values in the external validation dataset that is explained by the model's predictions (Y_{PREDi}).

Let $var(Y_i)$ denote the variance of Y_i values, and $var(Y_i - Y_{PREDi})$ denote the variance of the $Y_i - Y_{PREDi}$ values (i.e., the variance of e_i, the errors in the predictions). The proportion of outcome variation explained by the predicted values from the prediction model, R^2_{val}, is calculated by:

$$R^2_{val} = 1 - \frac{var(Y_i - Y_{PREDi})}{var(Y_i)}, \tag{2}$$

where values of R^2_{val} closer to 1 indicate a better fit of the Y_{PREDi} from the prediction model.

Calibration slope and calibration-in-the-large Calibration measures the agreement between predicted (Y_{PREDi}) and observed (Y_i) outcome values [80]. It is best shown graphically on a calibration plot, with Y_{PREDi} on the horizontal x-axis plotted against Y_i on the vertical y-axis, as shown in the previous chapters. For a continuous outcome, every individual provides a single data

point, comparing their personal prediction value to their observed outcome. A loess-smoothed calibration curve can be fitted through all of these individual points and, when presented on the calibration plot, gives a summary of the calibration performance on average across the validation population [30, 79, 188]. Ideally, the predicted outcome values should not be systematically under- or over-estimated across the range of predicted values. Points should be scattered randomly around the 45 degree line of ideal calibration (corresponding to a calibration slope of 1, and a CITL of 0), with little variation around the line and with close agreement between predicted and observed values across the entire range of predicted values.

To formally quantify calibration performance in an external validation dataset, a calibration model can be fitted of the form,

$$\begin{aligned} Y_i &= \alpha_{cal} + \lambda_{cal}(Y_{PREDi}) + e_{cali} \\ e_{cali} &\sim \mathcal{N}(0, \sigma^2_{cal}), \end{aligned} \tag{3}$$

where *cal* is used to denote that parameters are from the calibration model, rather than the prediction model itself.

The calibration model is fitted using the standard estimation methods for a linear regression, for example using restricted maximum likelihood estimation. The parameter λ_{cal} represents the calibration slope, which measures agreement between predicted and observed outcomes across the whole range of predicted values.[30, 60] As mentioned, the ideal λ_{cal} value is 1. A $\lambda_{cal} < 1$ indicates that some predictions are too extreme (predictions above the mean are too high, and predictions below the mean are too low) and a slope > 1 indicates that the range of predictions is too narrow. A calibration slope < 1 is often observed in external validation studies, as clinical

prediction models tend to be developed in small datasets without adjustment for overfitting to that development data. This leads to extreme predictions, or miscalibration, in new individuals [55, 50, 189, 190]. The term σ^2_{cal} refers to the residual variance in the calibration model.

Note that the calibration slope can also be expressed as, [191]

$$\lambda_{cal} = \sqrt{\frac{R^2_{cal} var(Y_i)}{var(Y_{PREDi})}}, \tag{4}$$

where R^2_{cal} is the proportion of variance of Y_i values explained by the calibration model 3.

Systematic under- or over-prediction is still possible even when the calibration slope is equal to 1. Thus the calibration slope should always be considered alongside calibration-in-the-large (CITL), which measures the agreement between mean predicted ($\overline{Y_{PRED}}$) and mean observed ($\bar{Y}$) outcome values:

$$CITL_{val} = \bar{Y} - \overline{Y_{PRED}}. \tag{5}$$

Estimating $CITL_{val}$ from applying equation 5 in an external validation dataset is equivalent to estimating α_{cal} by fitting model 3 with the addition of a constraint that λ_{cal} should be equal to 1 (see Section 4.3.2).

4.3 Sample size to target precise estimates of predictive performance

This section introduces the four proposed criteria, mentioned above, that could be used by researchers to determine the minimum sample size required for an external validation of an existing prediction model. The first three criteria aim to ensure the sample size is large enough to estimate $R^2_{val}, CITL_{val}, and \lambda_{cal}$ with a small margin of error. Closed-form solutions are presented for this purpose, suitable for calculation without requiring access to specialist software. As these expressions depend on an estimate of the residual variance of the prediction model in the validation population, a fourth criterion also aims to ensure precise estimation of this variance.

4.3.1 Criterion (i): Precise estimate of R^2_{val}

The first criterion targets a precise estimate for R^2_{val}, such that the confidence interval for R^2_{val} is sufficiently narrow. There are many suggestions for deriving confidence intervals for R^2 [192], and the following utilises the approach suggested by Wishart [193], which approximates the standard error (SE) of $\hat{R}^2_{val}$ as:

$$SE_{\hat{R}^2_{val}} = \sqrt{(\frac{(4R^2_{val}(1-R^2_{val})^2)}{n}}. \tag{6}$$

This approximation works well when the sample size (n) is reasonably large (> 50) [192], which is likely to be the case when externally validating a clinical prediction model, given the requirements of the other criteria (for example, see criterion (iv)). Rearranging equation (6) gives a closed-form calculation for the required minimum sample size of:

$$n = \frac{4R^2_{val}(1-R^2_{val})^2}{SE^2_{\hat{R}^2_{val}}}. \tag{7}$$

Thus, equation 7 can be used to calculate the sample size (n) required to meet criterion (i), by

specifying a desired precision through the value for $SE^2_{\hat{R}^2_{val}}$ and by setting R^2_{val} at the anticipated value for R^2 in the external validation population.

For example, consider an existing prediction model with an adjusted R^2 value of 0.5 in the model development data, with this adjusted R^2 giving an unbiased estimate of expected performance in new data from the same population (as opposed to the apparent R^2, which would be overly optimistic). Then, if we assume the validation sample is from a similar target population to the development sample, such that the proportion of variance explained is likely to be similar, a simple starting point would be to anticipate R^2_{val} upon external validation is the same as the adjusted $\hat{R}^2$ reported in the model development study. To target a 95% confidence interval for R^2_{val} that has a narrow width of about 0.1, a $SE_{\hat{R}^2_{val}}$ of 0.0255 is needed. This stems from the assumption that the 95% confidence interval for R^2_{val} can be derived as approximately $\hat{R}^2_{val} \pm (1.96 * SE_{\hat{R}^2_{val}})$. Applying equation 7 gives

$$n = \frac{4R^2_{val}(1 - R^2_{val})^2}{SE^2_{\hat{R}^2_{val}}} = \frac{4 * 0.5 * (1 - 0.5)^2}{0.0255^2} = 768.9,$$

and so 769 participants would be the minimum required to meet criterion (i).

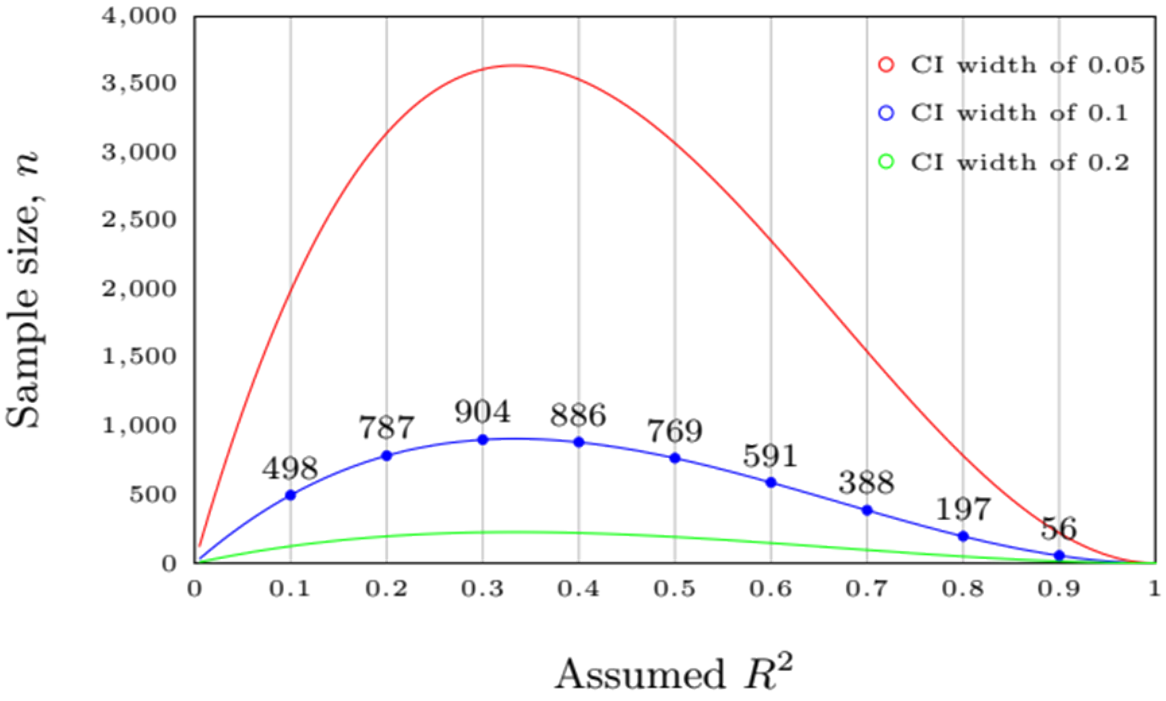
To achieve the same margin of error, with the same confidence interval width, 905 participants would be required to externally validate the model, assuming a lower R^2_{val} of 0.3, and 197 participants would be required when assuming a better overall model fit in the validation data with a R^2_{val} of 0.8. These values are close to examples gained using more exact (but not closed-form) approaches to confidence interval derivation for R^2, such as intervals based on a scaled, non-central F approximation [194] The *ss.aipe.R2* function within Kelley's MBESS package for R software identifies the sample size required to ensure this approximate confidence interval for R^2_{val} is

sufficiently narrow,[195, 196] and so is an alternative to using equation 7, though requires sufficient knowledge of R to implement.

Figure 4.1 shows how the required sample size changes for R^2_{val} values between 0.1 and 0.9 based on equation 7 and assuming $SE^2_{\hat{R}^2_{val}}$ is 0.0255, to target a confidence interval width of 0.1 as shown above. The required sample size to achieve the desired precision is lower when allowing for wider target confidence intervals (less precise), and higher when aiming for narrower target confidence intervals (more precise), as can be seen in Figure 4.1. Targetting a $SE^2_{\hat{R}^2_{val}} \leq 0.0225$ is likely a sensible compromise, as it aims for a precise estimate (with a margin of error of 0.05 or less, compared to the true value) and still gives a required sample size that will be realistic to obtain in practice.

It is worth noting that the observed R^2_{val} upon external validation may be lower or higher than the adjusted $\hat{R}^2$ reported for model development. Therefore, although the adjusted $\hat{R}^2$ from the development study is a useful starting point, it is beneficial to also calculate the sample size that would be required for external validation when assuming a range of different values for the true R^2_{val}. For example, researchers might apply equation 7 assuming R^2_{val} values ± 0.1 of the adjusted $\hat{R}^2$ reported from the development study, and note the largest sample size across this range as being their minimum.

Figure 4.1: Sample size (number of participants, n) needed in an external validation dataset to target a confidence interval for R^2_{val} of a particular width (either 0.05, 0.1, or 0.2) for different assumed R^2_{val} values between 0.1 and 0.9. Sample size calculated using equation 7.



4.3.2 Criterion (ii): Precise estimate of CITL

In order to target an accurate assessment of the level of systematic under- or over-prediction in the external validation set, this second criterion targets a precise estimate of $CITL_{val}$. Here $CITL_{val}$ is estimated using $\bar{Y} - \bar{Y}_{PRED}$ (from equation 5), which is equivalent to estimating the intercept term when fitting model 3 in the external validation dataset, with the predicted values used as an offset term:

$$\begin{aligned} Y_i &= CITL_{val} + 1(Y_{PREDi}) + e_{CITLi} \\ e_{CITLi} &\sim \mathcal{N}(0, \sigma^2_{CITL}) \end{aligned} \tag{8}$$

Therefore the standard error (SE) of this $\hat{CITL}$ is estimated from the residual errors when the prediction model is applied in the validation population, as follows:

$$SE^2_{\hat{CITL}} = var(\bar{Y} - \overline{Y_{PRED}}) = var\left(\frac{\sum_{i=1}^{n}(Y_i - Y_{PREDi})}{n}\right) = \frac{\sigma^2_{CITL}}{n} = \frac{var(Y_i)(1 - R^2_{CITL})}{n} \tag{9}$$

Rearranging equation 9 gives an expression in terms of n, for the required sample size subject to a desired precision in CITL, $SE_{\hat{CITL}}$:

$$n = \frac{var(Y_i)(1 - R^2_{CITL})}{SE^2_{\hat{CITL}}} \tag{10}$$

Hence, the sample size required to meet criterion (ii) can be derived using equation 10, for which the researcher must pre-specify R^2_{CITL} (the anticipated proportion of variance explained by the predictions in the external validation population), along with $var(Y_i)$ (the anticipated variance of Y_i in the target population), and the desired $SE_{\hat{CITL}}$ to achieve their target precision in CITL.

A sensible starting point is to assume CITL is zero. In this case, R^2_{CITL} is simply equal to R^2_{val} (the anticipated proportion of variance explained by the predictions upon validation), and so

$$n = \frac{var(Y_i)(1 - R^2_{val})}{SE^2_{\hat{CITL}}} \tag{11}$$

with R^2_{val} assumed to be the same as the adjusted $\hat{R}^2$ reported from the development study.

If CITL is non-zero then R^2_{CITL} will not equal R^2_{val}. Therefore, it is also sensible to consider a realistic range of values for R^2_{CITL} when applying equation 10, such as ± 0.1 of the adjusted $\hat{R}^2$ reported from the development study, and to note the largest sample size across this range, as was the case for criterion (i).

A value to define a suitably precise $SE_{\hat{CITL}}$ is clearly context specific, as it depends on the scale of the continuous outcome values. For example, for a model predicting systolic blood pressure (SBP), a standard error of about 2.5mmHg for CITL may indicate an appropriately high precision. If considering the pain intensity score example from Chapter 3, a $SE_{\hat{CITL}}$ of 2.5 units on the 0-10 pain scale would be extremely imprecise, thus a much smaller target standard error would be required.

Consider further the external validation of a prediction model for SBP. Suppose this model had an

adjusted R^2 of 0.5 in the development study, and the variance of the observed Y_i values in the target population for the validation study is anticipated to be around 400mmHg. Targetting a $SE_{\hat{CITL}}$ of 2.5mmHg gives a 95% confidence interval for $CITL_{val}$ with a narrow width of about 10mmHg, when deriving an approximate 95% confidence interval for $CITL_{val}$ as $\hat{CITL} \pm (1.96 * SE_{\hat{R}^2_{val}})$. Assuming $R^2_{CITL} = R^2_{val} = 0.5$, then applying equation 11 gives,

$$n = \frac{var(Y_i)(1 - R^2_{val})}{SE^2_{\hat{CITL}}} = \frac{400(1 - 0.5)}{2.55^2} = 30.76$$

and so a minimum of only 31 participants would be required for the external validation to meet criterion (ii).

More cautiously assuming that $R^2_{CITL} = 0.4$, the required sample size would be

$$n = \frac{var(Y_i)(1 - R^2_{val})}{SE^2_{\hat{CITL}}} = \frac{400(1 - 0.4)}{2.55^2} = 36.91$$

and so 37 participants would instead be required.

Clearly, these minimum values to precisely estimate the CITL on external validation are very low. It is likely that the sample size to precisely estimate CITL is smaller than that required to precisely estimate the measures outlined in criteria (i), (iii) and (iv) in most cases, and so criterion (ii) is unlikely to drive the overall sample size requirement for the external validation study.

4.3.3 Criterion (iii): Precise estimate of calibration slope

The third suggested criterion also aims for accurate estimation of calibration performance, now by targetting a precise estimate of λ_{cal}, the calibration slope obtained from fitting a calibration model 3 in the external validation dataset. As $\hat{\lambda}_{cal}$ is essentially the slope from a simple linear regression model, the standard error of $\hat{\lambda}_{cal}$ can be estimated by, [197]

$$SE^2_{\hat{\lambda}_{cal}} = \frac{\sigma^2_{cal}}{\sum_{i=1}^{n}(Y_{PREDi} - \bar{Y}_{PRED})^2}$$

where σ^2_{cal} is the residual variance from model 3.

Recognising that $\sigma^2_{cal} = var(Y_i)(1-R^2_{cal})$, and given $\sum_{i=1}^{n}(Y_{PREDi} - \bar{Y}_{PRED})^2 = (n-1)var(Y_{PREDi})$, we can rewrite this as

$$\begin{aligned} SE^2_{\hat{\lambda}_{cal}} &= \frac{\sigma^2_{cal}}{\sum_{i=1}^{n}(Y_{PREDi} - \bar{Y}_{PRED})^2} \\ &= \frac{var(Y_i)(1-R^2_{cal})}{(n-1)var(Y_{PREDi})} \\ &= \frac{var(Y_i)}{(n-1)var(Y_{PREDi})} - \frac{var(Y_i)R^2_{cal}}{(n-1)var(Y_{PREDi})} \end{aligned} \tag{12}$$

Further, by utilising equation 4, we can write $SE^2_{\hat{\lambda}_{cal}}$ in terms of λ^2_{cal} and R^2_{cal} values [191], as follows:

$$
\begin{aligned}
SE^2_{\hat{\lambda}_{cal}} &= \frac{var(Y_i)}{(n-1)var(Y_{PREDi})} - \frac{var(Y_i)R^2_{cal}}{(n-1)var(Y_{PREDi})} \\
&= \frac{1}{(n-1)R^2_{cal}}\left(\sqrt{\frac{R^2_{cal}var(Y_i)}{var(Y_{PREDi})}}\right)^2 - \frac{1}{(n-1)}\left(\sqrt{\frac{R^2_{cal}var(Y_i)}{var(Y_{PREDi})}}\right)^2 \\
&= \frac{\lambda^2_{cal}}{(n-1)R^2_{cal}} - \frac{\lambda^2_{cal}}{(n-1)} \\
&= \frac{\lambda^2_{cal}}{(n-1)R^2_{cal}} - \frac{\lambda^2_{cal}R^2_{cal}}{(n-1)R^2_{cal}} \\
&= \frac{\lambda^2_{cal}(1-R^2_{cal})}{(n-1)R^2_{cal}}
\end{aligned} \tag{13}
$$

As with previous criteria, rearranging this gives the sample size, n, that corresponds with a given $SE^2_{\hat{\lambda}_{cal}}$:

$$
n = \frac{\lambda^2_{cal}(1-R^2_{cal})}{SE^2_{\hat{\lambda}_{cal}}R^2_{cal}} + 1 \tag{14}
$$

Equation 14 therefore allows calculation of the required sample size for a desired $SE_{\hat{\lambda}_{cal}}$, conditional on specifying the anticipated calibration slope across the range of predicted values, λ_{cal}, and the anticipated proportion of variance in observed Y_i values explained by the calibration model, R^2_{cal}.

When choosing a suitable level of precision, the value $SE_{\hat{\lambda}_{cal}} = 0.051$ is proposed, to target a 95% confidence interval for λ_{cal} with a narrow width (≤ 0.2). Using this value, if the calibration slope was 1 for example, the resulting confidence interval would be 0.9 to 1.1, assuming confidence intervals derived by $\hat{\lambda}_{cal} \pm 1.96SE_{\hat{\lambda}_{cal}}$. It is worth noting that λ_{cal} is known to follow a student's t-distribution [197], thus in small samples, critical values from this distribution would

replace 1.96 in the confidence interval calculation. Given sample sizes will not be small in practice, however, the normal approximation to the t-distribution is suitable (by the Central Limit Theorem).

As an appropriate value for λ_{cal}, a simple starting point is to assume good calibration, such that $\lambda_{cal} = 1$ and $\alpha_{cal} = 0$ in model 3. In this case, R^2_{cal} can be approximated by R^2_{val}, introduced in criteria (i). Thus R^2_{cal} can be assumed to be the same as the adjusted R^2 value in the model development study.

For example, in an external validation of a prediction model that had an estimated adjusted R^2 of 0.5 in the development dataset, a simple and logical starting point would be to anticipate the same value for R^2_{val}. Then, assuming the model's predictions will be well calibrated in the external validation dataset, such that fitting model 3 would give $\hat{\alpha}_{cal}$ of zero and an $\hat{\lambda}_{cal}$ of one, using equation 14 gives,

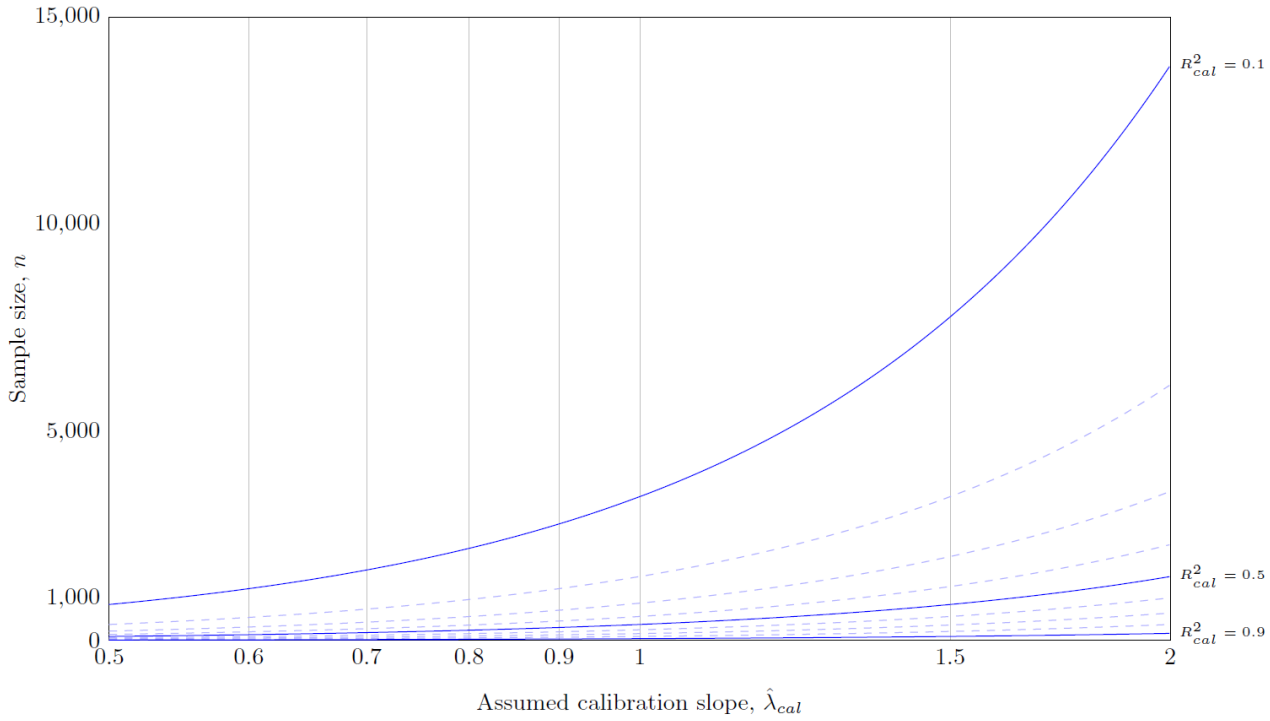
$$n = \frac{\lambda^2_{cal}(1 - R^2_{cal})}{SE^2_{\hat{\lambda}_{cal}} R^2_{cal}} + 1 = \frac{1(1 - 0.5)}{0.051 * 0.051 * 0.5} + 1 = 385.47$$

and thus 386 participants are required to target a confidence interval width of 0.1 for the calibration slope, under the assumption of good calibration.

In practice, the value of $\hat{\lambda}_{cal}$ is unlikely to exactly equal one, thus the sample size should also be sufficient to precisely estimate some miscalibration. Often during an external validation the calibration slope is less than one, with extreme predictions due to overfitting during model development. In such situations R^2_{cal} can still be assumed to be equal to the adjusted R^2 presented for model development, as this adjusted value will have allowed for optimism due to overfitting. When fixing R^2_{cal} and $SE^2_{\hat{\lambda}_{cal}}$ values, applying equation 14 for lower assumed $\hat{\lambda}_{cal}$ (values below

one) gives lower required sample sizes than when assuming the prediction model is well calibrated (see Figure 4.2). Thus, assuming $\hat{\lambda}_{cal}$ is equal to one gives a more conservative estimate of the minimum sample size required.

Figure 4.2: Sample size (number of participants, n) needed in an external validation dataset to target a confidence interval for $\hat{\lambda}_{cal}$ of width of 0.2, for different assumed $\hat{\lambda}_{cal}$ values between 0.5 and 2, and for R^2_{cal} values between 0.1 and 0.9. Sample size calculated using equation 14.



Further combinations of $\hat{\lambda}_{cal}$ and R^2_{cal} values could be tested if desired, such as those shown in Figure 4.2. Choosing appropriate combinations is likely to be more complex than with previous criteria, however, as the values of $\hat{\lambda}_{cal}$ and R^2_{cal} are not independent. Equation 4 shows that $\hat{\lambda}_{cal}$ depends on R^2_{cal} (along with $var(Y_i)$ and $var(Y_{PREDi})$), thus testing values while changing assumed $\hat{\lambda}_{cal}$ has implications for what the assumed value of R^2_{cal} should be. Allowing for this

interaction is likely too intricate for this sample size calculation, and is not necessary in cases where overfitting is expected to result in a calibration slope less than one.

It is possible for a prediction model to be underfit to the development data, resulting in a range of predictions that is too narrow on external validation and a $\hat{\lambda}_{cal} > 1$. Such situations result in considerably larger required sample sizes for precise estimation of $\hat{\lambda}_{cal}$, especially for lower anticipated values for R^2_{cal}. In practice, such situations are very rare, with overfitting being more common, thus it is unlikely to be necessary to consider this eventuality as a part of the sample size calculation. In general, applying equation 14 assuming good calibration ($\hat{\lambda}_{cal} = 1$) will be sufficient.

4.3.4 Criterion (iv): Precise estimates of the residual variance

This final criterion targets precise estimates of the residual variances of the calibration models, $\hat{\sigma}^2_{CITL}$ and $\hat{\sigma}^2_{cal}$. Although these residual variances are not direct measures of predictive performance themselves, precision is essential as their estimated values are used toward parameter estimates and, crucially, toward values of $SE_{\hat{CITL}_{val}}$ and $SE_{\hat{\lambda}_{cal}}$.

Given the format of the calibration model in equation 8, to ensure precision in $\hat{\sigma}^2_{CITL}$, we consider the residual variance in a linear regression model with only an intercept term (see model 8). In such situations, Harrell suggests calculating the sample size to ensure the lower and upper bounds of the 95% confidence interval for the residual variance has a small multiplicative margin of error (MMOE) around the true value [198], using

$$MMOE = \sqrt{max\left(\frac{\chi^2_{1-\frac{\alpha}{2},n-1}}{n-1}, \frac{n-1}{\chi^2_{\frac{\alpha}{2},n-1}}\right)} \tag{15}$$

where $\chi^2_{1-\frac{\alpha}{2},n-1}$ and $\chi^2_{\frac{\alpha}{2},n-1}$ are the critical values of the χ^2 distribution with n-1 degrees of freedom for which there are probabilities of $1-\frac{\alpha}{2}$ and $\frac{\alpha}{2}$ of being less than the critical value, respectively. The largest MMOE typically results from the second term in the bracket of equation 15.

A margin of error of within 10% of the true value ($1.0 \leq$ MMOE ≤ 1.1) is sufficient in practice. Using equation 15 to target this margin reveals that a sample size of at least 234 participants is needed to ensure a $MMOE \leq 1.1$ for $\hat{\sigma}^2_{CITL}$.

For precise estimation of $\hat{\sigma}^2_{cal}$, the sample size for precise estimation of $\hat{\sigma}^2_{CITL}$ simply needs to be adjusted for the additional slope parameter being estimated in model 3. As outlined by Riley et

al.,[75] this is achieved as just $234 + 1$, and thus 235 participants are required to ensure a MMOE ≤ 1.1 for $\hat{\sigma}^2_{cal}$. Thus, the minimum sample size required to meet criterion (iv) is driven by the required precision in $\hat{\sigma}^2_{cal}$, and at least 235 participants are needed for any external validation of a prediction model for a continuous outcome, regardless of context. This minimum is prior to the consideration of criteria (i), (ii) or (iii).

4.3.5 Summary of the proposed criteria

The above sample size criteria aim to ensure that the external validation dataset is sufficiently sized to precisely estimate key performance measures (R^2_{val}, CITL, calibration slope) and residual variances. The approach requires a separate sample size assessment for each criterion, with the largest required sample size across criteria providing the minimum needed to meet all requirements simultaneously. A step-by-step summary of the proposed sample size calculation follows in Figure 4.3.

Figure 4.3: Summary of the steps involved in the proposed sample size calculation, for the external validation of a clinical prediction model with a continuous outcome

STEP 1: Calculate the sample size needed to precisely estimate R^2_{val} (criterion (i))

Apply equation 7,

$$n = \frac{4R^2_{val}(1 - R^2_{val})^2}{SE^2_{\hat{R}^2_{val}}}$$

after specifying suitable values for $SE_{\hat{R}^2_{val}}$ and R^2_{val}. Using $SE_{\hat{R}^2_{val}} \leq 0.0255$ (to target a 0.1 confidence interval width) is recommended, and initially choosing R^2_{val} to equal the adjusted $\hat{R}^2$ reported for the model development study. Other values for R^2_{val} might also be considered, including values ± 0.1 the adjusted $\hat{R}^2$ reported from the development study.

STEP 2: Calculate the sample size needed to precisely estimate calibration-in-the-large (criterion (ii))

Apply equation 10,

$$n = \frac{var(Y_i)(1 - R^2_{CITL})}{SE^2_{\hat{CITL}}}$$

after specifying suitable values for R^2_{CITL} (akin to those used for R^2_{val} in step 1), $SE_{\hat{CITL}}$ and $var(Y_i)$. The latter represents the variance of outcome values in the population of interest, and should be based on other existing knowledge (e.g. from previous studies). The value of $SE_{\hat{CITL}}$ should aim to ensure that $(\bar{Y} - \bar{Y}_{PRED}) \pm (1.96 * SE_{\hat{CITL}})$ is sufficiently narrow, and so needs to be chosen in context of what constitutes a precise estimate of the mean prediction error in the clinical setting of interest.

STEP 3: Calculate the sample size needed to precisely estimate calibration slope (criterion (iii))

Apply equation 14,

$$n = \frac{\lambda^2_{cal}(1 - R^2_{cal})}{SE^2_{\hat{\lambda}_{cal}} R^2_{cal}} + 1$$

after specifying suitable values for λ_{cal}, R^2_{cal} and $SE_{\hat{\lambda}_{cal}}$. $SE_{\hat{\lambda}_{cal}} \leq 0.051$ is recommended, to target a confidence interval width ≤ 0.2, as is choosing R^2_{cal} to be the same as that chosen for R^2_{val} (i.e., the adjusted R^2 reported from the model development study; see step 1). Assuming $\lambda_{cal} = 1$ (good calibration) is also recommended.

STEP 4: Calculate the sample size for precisely estimating residual variances (criterion (iv))

To target residual variance estimates in the calibration models that have a margin of error of $\leq 10\%$, at least 235 participants are required, regardless of clinical area, based on equation 15

STEP 5: Calculate the final sample size

Identify the minimum sample size required as the maximum value from steps 1 to 4, to ensure that each of criteria (i) to (iv) are met.

4.4 Applied example: sample size required to externally validate a model for predicting fat-free mass in children

To demonstrate the proposed sample size calculation outlined in Figure 4.3, the suggested criteria are now applied in an illustrative example for a model to predict fat-free mass on the continuous kilogram (kg) scale [17]. In 2019, Hudda et al. developed a prediction model for the natural logarithm of fat-free mass in children and adolescents, aged 4 to 15 years, from five predictors (including ten predictor parameters): the child's height, weight, age, sex and ethnicity. The apparent calibration of the model in the development dataset was reported in the original publication and is shown in Figure 4.4a. In the development dataset, the estimated adjusted R^2 was reported to be very good, at 0.948. The published model equation is as follows:

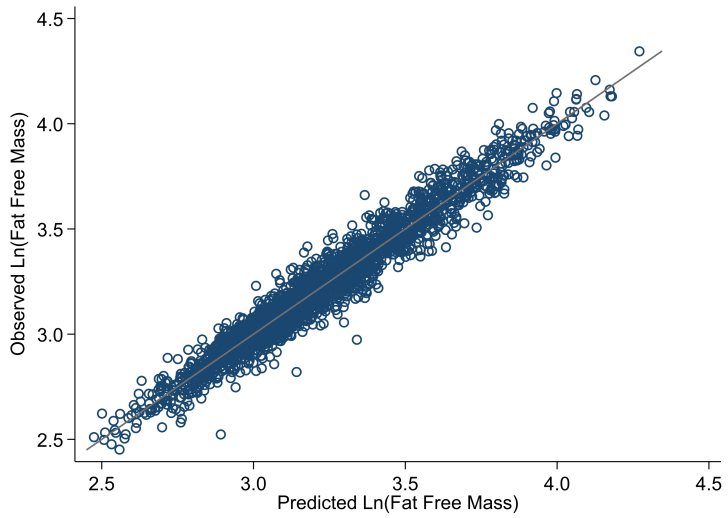
$$\begin{aligned}\ln \text{fat-free mass} =& 2.8055 + 0.3073(\text{height}^2) - 10.0155(\text{weight}^{-1}) + 0.004571(\text{weight}) \\ &+ 0.01408(\text{if Black ethnicity}) - 0.06509(\text{if South Asian ethnicity}) \\ &- 0.02624(\text{if other Asian ethnicty}) - 0.01745(\text{if other ethnicty}) \\ &- 0.9180(\ln(\text{age})) + 0.6488(\text{age}^{0.5}) + 0.04723(\text{if male})\end{aligned}$$

where predictor variables of Black, South Asian, other Asian, or other ethnic origins are binary, with value of 1 if individual has the particular origin and 0 otherwise. The child's height, weight and age are all continuous predictors, with height measured in metres, weight in kilograms, age in years.

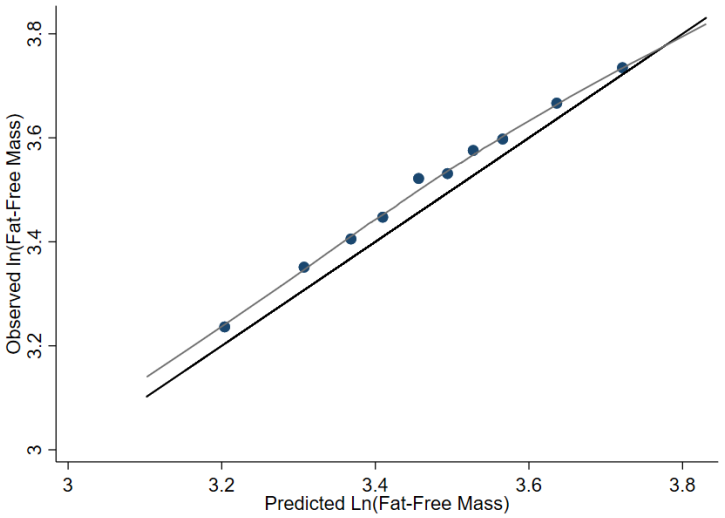
Hudda presented an initial external validation, undertaken in a sample of 176 children aged 11-12 years from the UK Avon Longitudinal Study of Parents and Children (ALSPAC) study [199, 200], where the model had an estimated R^2_{val} of 0.90 (Figure 4.4b). However, as acknowledged by the

study authors, further external validation was also warranted in a broader population, for example with a wider age range. A sample size calculation for a new external validation of this prediction model can be undertaken using the above proposed methods, assuming that the performance of the model in the validation population would be similar to that seen in the development and original external validation populations.

Figure 4.4: Calibration performance: panel A – in the development dataset; and panel B – on external validation of the prediction model for ln(fat-free mass) in children, as proposed by Hudda et al. The 45 degree line shows perfect calibration on both plots.



(a) Development dataset



(b) Validation dataset

4.4.1 STEP 1: Calculate the sample size needed to precisely estimate R^2_{val} (criterion (i))

As shown in Figure 4.3, the first stage of the calculation requires the application of equation 7. Based on assuming an $\hat{R}^2_{val} = 0.90$, as in the initial published external validation of the model, and using a $SE_{\hat{R}^2_{val}}$ of 0.0255 to target a confidence interval width of 0.1:

$$\begin{aligned} n &= \frac{4R^2_{val}(1 - R^2_{val})^2}{SE^2_{\hat{R}^2_{val}}} \\ &= \frac{4 * 0.9 * (1 - 0.9)^2}{0.0255^2} \\ &= 55.4 \end{aligned}$$

Thus, a sample size of only 56 children is required to target the desired precision around the expected $\hat{R}^2_{val}$.

As it is sensible to also consider the possibility that the model could perform worse upon further external validation, it would be beneficial to consider the sample size needed for similarly precise estimation of a lower $\hat{R}^2_{val}$ value. If we were to consider a 0.1 reduction in $\hat{R}^2_{val}$ to 0.80, then the required sample size to meet criteria (i) increases to 197 children, thus this value could be taken as a more conservative estimate of the required number of children needed in a new external validation sample.

The sample size values above were calculated through application of equation 7, but could equally have been obtained from a visual inspection of the curve for a target confidence interval width of 0.1, shown in Figure 4.1.

4.4.2 STEP 2: Calculate the sample size needed to precisely estimate calibration-in-the-large (criterion (ii))

The second step in the calculation requires the application of equation 10. Thus, this step needs specification of a value for $v\hat{a}r(Y_i)$, the anticipated variance of outcome values in the target population for external validation. In their paper, Hudda did not report the variance of ln(fat-free mass) in their population, thus to apply equation 10, we must derive an estimate from the published information.

Given their outcome of interest on its pre-transformed scale was skewed, Hudda summarise the distribution of fat-free mass in their development dataset using the median, lower quartile (LQ) and upper quartile (UQ) values. They report a LQ of 20.8kg and a UQ of 30.6kg, for fat-free mass on the kg scale. By transforming this to the ln(kg) scale, and assuming that, following the natural log transformation, ln(fat-free mass) values are approximately normally distributed, an estimate of the standard deviation of the ln(fat-free mass) in the development population can be derived using [201]:

$$\sqrt{v\hat{a}r(Y_i)} \approx \frac{\ln UQ - \ln LQ}{1.35} = \frac{\ln 30.6 - \ln 20.8}{1.35} = 0.29$$

Therefore, based on the published information $v\hat{a}r(Y_i) \approx 0.29^2 = 0.082$. Interestingly, when contacting the study authors directly for this information, they calculated the observed value to be similar, at $v\hat{a}r(Y_i) = 0.089$. The value given by the study authors will be used in the calculation going forward, though the concept would equally apply with the estimation from published information, if study authors we not contactable.

To apply the equation, the expected value for R^2_{CITL} must also be specified. As suggested above, this is assumed to be the same as the expected $\hat{R}^2_{val}$ in the previous step, such that $\hat{R}^2_{CITL} = \hat{R}^2_{val} = 0.90$, as in Hudda's initial external validation of the model.

The desired precision in calibration-in-the-large needs to be placed in context of the mean outcome value in the population if interest, which, in this case, is the mean fat-free mass in a population of children and adolescents. Hudda reported a median baseline fat-free mass of 24.8kg in their population, thus assuming that the median and mean values on the natural log scale are similar, we have:

$$\bar{Y} \approx \ln 24.8 = 3.21$$

Considering the original untransformed scale, an accuracy of approximately ±1kg around $\bar{Y}$ was considered to be reasonably precise. A 95% confidence interval from 23.8 to 25.8 on the kg scale would correspond to an interval of about 3.17 to 3.25 around $\bar{Y}$ on the ln(kg) scale, implying a target $SE_{\hat{CITL}}$ of about 0.02.

Incorporating the above information, equation 10 can be applied to obtain a required sample size of at least:

$$
\begin{aligned}
n &= \frac{var(Y_i)(1 - R^2_{CITL})}{SE^2_{C\hat{I}TL}} \\
&= \frac{0.089 * (1 - 0.9)}{0.02^2} \\
&= 22.3
\end{aligned}
$$

for a sufficiently precise estimation of CITL on external validation. Thus a minimum of 23 participants would be required to meet criterion (ii).

As in the previous step, it is sensible to also consider a situation with worse model performance on further validation. To be conservative, instead assume a decrease in $\hat{R}^2_{CITL}$ by 0.1 compared to the initial validation data, to 0.80. Then, the required sample size to meet criteria (ii) would increase to 45 children.

$$
\begin{aligned}
n &= \frac{var(Y_i)(1 - R^2_{CITL})}{SE^2_{C\hat{I}TL}} \\
&= \frac{0.089 * (1 - 0.8)}{0.02^2} \\
&= 44.5
\end{aligned}
$$

4.4.3 STEP 3: Calculate the sample size needed to precisely estimate calibration slope (criterion (iii))

This third step requires the application of equation 14, after choosing appropriate values for $SE_{\hat{\lambda}_{cal}}, \hat{R}^2_{cal}$ and $\hat{\lambda}^2_{cal}$. As mentioned above, assuming an $SE_{\hat{\lambda}_{cal}}$ of 0.051 targets a confidence interval width for the calibration slope of at most 0.2. Further, we can assume $\hat{R}^2_{cal} = \hat{R}^2_{val}$ and take the value of 0.90 as reported by the initial validation study of Hudda. With the further assumption of good calibration, such that $\hat{\lambda}^2_{cal}$ is one, equation 14 can be applied to give,

$$\begin{aligned} n &= \frac{\lambda^2_{cal}(1 - R^2_{cal})}{SE^2_{\hat{\lambda}_{cal}} R^2_{cal}} + 1 \\ &= \frac{1 * (1 - 0.9)}{0.051^2 * 0.9} \\ &= 43.72 \end{aligned}$$

and so a minimum of 44 participants would be required to gain a sufficiently precise estimate of the calibration slope in the new external validation population.

4.4.4 STEP 4: Calculate the sample size for precisely estimating residual variances (criterion (iv))

To fulfil the final criterion, to ensure a 10% margin of error in residual variance estimates from the calibration models, at least 235 participants are required regardless of clinical context, as discussed in Section 4.3.4 above. Thus, a minimum of 235 participants are needed to ensure precision in the residual variances of the calibration models for Hudda's fat-free mass example.

4.4.5 STEP 5: Calculate the final sample size

Assuming Hudda's fat-free mass model will be validated in a population where it performs to a similar level of accuracy as was seen in the initial external validation data, steps one to four have provided four sample size requirements to ensure that each of criterion (i) to (iv) are met. These requirements are summarised in Table 4.1, below. Based on the largest of these values, the minimum sample size required to meet all four criteria simultaneously is simply 235 participants. This is driven by criterion (iv), to target sufficient precision around $\hat{\sigma}^2_{CITL}$ and $\hat{\sigma}^2_{cal}$.

Table 4.1: Summary of the sample size calculation for external validation of the prediction model of Hudda et al.

Criterion	Target precision	Assumptions	Minimum sample size required
(i) Precise estimate of R^2_{val}	$SE_{\hat{R}^2_{val}} = 0.0255$	$R^2_{val} = 0.8$	197
		$R^2_{val} = 0.9$	56
(ii) Precise estimate of CITL	$SE_{\hat{CITL}} = 0.02$	$R^2_{CITL} = R^2_{val} = 0.8$, $var(Y_i) = 0.089$	45
		$R^2_{CITL} = R^2_{val} = 0.9$, $var(Y_i) = 0.089$	23
(iii) Precise estimate of λ_{cal}	$SE_{\hat{\lambda}_{cal}} = 0.051$	$R^2_{cal} = R^2_{val} = 0.8$	98
		$R^2_{cal} = R^2_{val} = 0.9$	44
(iv) Precise $\hat{\sigma}^2_{CITL}$ and $\hat{\sigma}^2_{cal}$	$1.0 \leq MMOE \leq 1.1$	–	235
Minimum sample size			**235**

4.5 Expected precision when sample size for external validation is fixed

The previous section discussed how to target precise estimates of predictive performance by recruiting an appropriate number of participants to the external validation study. As is often the case in model evaluation, the initial external validation of the Hudda fat-free mass model was conducted in a previously collected dataset (from the ALSPAC study [199, 200]), recruited for a different purpose. Where there are limited resources for prospective recruitment of the external validation cohort, researchers might seek an existing dataset from the population of interest and so they (along with other stakeholders such as funders and patient representatives) will need to assess whether the sample is large enough for a reliable external validation.

Where the size of an existing dataset is fixed, the number of participants available could simply be compared to the results of the above calculations. This would identify whether or not the available data meet each of the above criteria, but would not give any indication of how precise estimates of performance measures might be expected to be. To assess this expected precision, the calculations in for criteria (i) to (iv) can be re-expressed to calculate the expected $SE^2_{\hat{R}^2_{val}}$, $SE^2_{\hat{CITL}}$, $SE^2_{\hat{\lambda}_{cal}}$, and MMOE conditional on the known sample size. As before, assumed values of R^2_{val}, $var(Y_i)$, R^2_{CITL}, R^2_{cal}, and λ_{cal} should be specified, which can be based on the model's known performance (preferably optimism-adjusted) on internal validation.

Steps one to four, below, demonstrate the application of the re-expressed equations to assess the expected precision Hudda could expect in values of R^2_{val}, CITL, the calibration slope, and the residual errors of the calibration models, based on their available sample of 176 children for external validation.

4.5.1 STEP 1: Expected precision in R^2_{val}

To assess how accurately one could estimate the overall model fit to their existing external validation dataset, the expected value of $SE^2_{\hat{R}^2_{val}}$ can be obtained by rearranging equation 7 from criterion (i). To apply this adaptation of criterion (i), an expected value of $\hat{R}^2_{val}$ must be specified. As previously, a good starting point would be to assume model performance on external validation will be consistent with that seen in the model development data. Prior to Hudda's initial external validation, their best estimate of the anticipated $\hat{R}^2_{val}$ came from internal validation using bootstrapping, which gave an adjusted R^2 value of 0.948. Thus, assuming $\hat{R}^2_{val} = 0.948$,

$$n = \frac{4 * \hat{R}^2_{val} * (1 - \hat{R}^2_{val})^2}{SE^2_{\hat{R}^2_{val}}}$$

$$176 = \frac{4 * 0.948 * (1 - 0.948)^2}{SE^2_{\hat{R}^2_{val}}}$$

Rearranging gives,

$$SE^2_{\hat{R}^2_{val}} = \frac{4 * 0.948 * (1 - 0.948)^2}{176}$$

$$= 0.000058$$

Therefore, if $\hat{R}^2_{val} = 0.948$ in the validation data, we would expect $SE_{\hat{R}^2_{val}} = 0.0076$ and, correspondingly, a 95% confidence interval with width of 0.0299 around the estimate of $\hat{R}^2_{val}$, with an expected interval from 0.933 to 0.963. In this case, the anticipated estimate of $\hat{R}^2_{val}$ would be very precise, given a sample size of 176 for external validation.

If we were to assume worse overall model fit in the external validation data, as discussed previously, we might instead expect a $\hat{R}^2_{val} = 0.9$ or 0.8,

	$\hat{R}^2_{val} = 0.9$	$\hat{R}^2_{val} = 0.8$
$SE^2_{\hat{R}^2_{val}}$	$= \frac{4*0.9*(1-0.9)^2}{176}$	$= \frac{4*0.8*(1-0.8)^2}{176}$
	$= 0.0002$	$= 0.0007$
95% confidence interval width	0.0561	0.1057
Expected 95% confidence interval:	0.872 to 0.928	0.747 to 0.853

The observed $\hat{R}^2_{val}$ in Hudda's external validation was in fact 0.9 (as noted above), with a 95% confidence interval width of around 0.056, from 0.872 to 0.928. Clearly, where an accurate estimate of the expected $\hat{R}^2_{val}$ is used, exact confidence intervals for the overall model fit could be derived *a priori*.

Given the true value for $\hat{R}^2_{val}$ will not be known at the time of sample size assessment, we focus on the result that the expected 95% confidence interval width for $\hat{R}^2_{val}$ would be between 0.0299 and 0.1057, depending on the $\hat{R}^2_{val}$ value used. As discussed in the derivation of criterion (i), a reasonable target confidence interval width would be 0.1 (corresponding to a $SE_{\hat{R}^2_{val}}$ of 0.0255), thus the validation sample size of 176 would only be sufficient to provide a suitably precise estimate of $\hat{R}^2_{val}$ where $\hat{R}^2_{val} > 0.8$.

4.5.2 STEP 2: Expected precision in calibration-in-the-large

Similarly, anticipated precision in the estimate of CITL can be assessed from equation 10, after some rearrangement. First, assuming $\hat{R}^2_{CITL} = \hat{R}^2_{val} = 0.948$, from the estimate on internal validation, and with an expected $\hat{var}(Y_i) = 0.089$ as was seen in the development data, equation 10 can be rearranged to reveal,

$$n = \frac{var(Y_i)(1 - \hat{R}^2_{CITL})}{SE^2_{C\hat{I}TL}}$$

$$176 = \frac{0.089 * (1 - 0.948)}{SE^2_{C\hat{I}TL}}$$

Rearranging gives,

$$SE^2_{C\hat{I}TL} = \frac{0.089 * (1 - 0.948)}{176}$$

$$= 0.000026$$

Meaning an expected $SE_{C\hat{I}TL} = 0.005$ and a 95% confidence interval around $C\hat{I}TL_{val}$ with a width of approximately $0.020 \ln kg$. This corresponds to an anticipated 95% confidence interval width of 1.020kg, on the more clinically interpretable kilogram scale.

Assuming a lower value of $\hat{R}^2_{CITL} = \hat{R}^2_{val}$ the external validation data, as in step one, gives anticipated confidence interval widths as follows.

	$\hat{R}^2_{CITL} = 0.9$	$\hat{R}^2_{CITL} = 0.8$
$SE^2_{C\hat{I}TL}$	$= \frac{0.089*(1-0.9)}{176}$	$= \frac{0.089*(1-0.8)}{176}$
	$= 0.00005$	$= 0.00010$
95% confidence interval width, ln(kg)	0.028	0.039
95% confidence interval width, kg	1.028	1.040

On external validation in the ALSPAC study data [199, 200], Hudda reported a CITL value of -1.58kg (-2.29kg to -0.86kg), giving a 95% confidence interval width of 1.43kg. The observed precision in the estimate of CITL was in fact lower than that suggested by the above calculation, implying the above assumptions did not well match the external validation data. A key difference may have been been in the distribution of the observed fat-free mass levels in the validation population, perhaps due to population differences.

The children in the external validation population were older on average, weighed more, and followed a different distribution of ethnic origins to the model development data. The reported median (LQ to UQ) fat-free mass in the validation data did differ from that in the model development data, with a higher median and a narrower interquartile range at 33.8kg (29.8kg to 37.4kg) compared to 24.8kg (20.8kg to 30.6kg).

4.5.3 STEP 3: Expected precision in the calibration slope

Assuming the model would be well calibrated in the new dataset, such that $\lambda^2_{cal} = 1$, and that $\hat{R}^2_{cal} = \hat{R}^2_{CITL} = \hat{R}^2_{val} = 0.948$ allows the estimation of $SE_{\hat{\lambda}_{cal}}$ from equation 14 from criterion (iii). Rearranging equation criterion iii here would give a conservative estimate of the expected $SE_{\hat{\lambda}_{cal}}$ in the validation population, as mentioned above, given the assumption of good calibration ($\lambda^2_{cal} = 1$) would result in the a higher required sample size, corresponding to a higher expected $SE_{\hat{\lambda}_{cal}}$ for a fixed sample size, when compared to the more common situation of overfitting of the model to the development data (with $\lambda^2_{cal} < 1$).

$$n = \frac{\lambda^2_{cal}(1 - R^2_{cal})}{SE^2_{\hat{\lambda}_{cal}} R^2_{cal}} + 1$$

$$176 = \frac{1 * (1 - 0.948)}{0.9 SE^2_{\hat{\lambda}_{cal}}} + 1$$

Rearranging gives,

$$SE^2_{\hat{\lambda}_{cal}} = \frac{1 * (1 - 0.948)}{0.948 * 175}$$

$$= 0.0003$$

This would suggest an expected $SE_{\hat{\lambda}_{cal}}$ of 0.018, corresponding to a narrow 95% confidence interval around the calibration slope of width 0.069 (from 0.965 to 1.035, given the assumption that $\lambda^2_{cal} = 1$).

Expanding this to accommodate lower values for $\hat{R}^2_{cal}$ of 0.9 and 0.8, as above, suggests slightly wider confidence intervals, but still acceptable levels of precision when compared to the maximum

interval width of 0.2 suggested in Section 4.3.3.

	$\hat{R}^2_{cal} = 0.9$	$\hat{R}^2_{cal} = 0.8$
$SE^2_{\hat{\lambda}_{cal}}$	$= \frac{1*(1-0.9)}{0.9*175}$	$= \frac{1*(1-0.8)}{0.8*175}$
	$= 0.0006$	$= 0.0014$
95% confidence interval width	0.0988	0.1482
Expected 95% confidence interval:	0.951 to 1.049	0.926 to 1.074

The observed calibration slope in Hudda's external validation was 1.02, with a 95% confidence interval from 0.97 to 1.07, with a width of 0.10. This closely matches the expected precision in the estimate of $\hat{\lambda}_{cal}$ when assuming an $\hat{R}^2_{cal}$ of 0.9, matching the $\hat{R}^2_{val}$ value that was observed in the external validation data.

4.5.4 STEP 4: Expected precision in σ^2_{CITL} and σ^2_{cal}

The final criterion is needed to assess the expected precision around the residual variances of the calibration model and for the calibration-in-the-large, which in turn affect the expected precision around estimates of CITL and the calibration slope. The sample size available for external validation was lower than the 235 recommended for precise estimation of σ^2_{CITL} and σ^2_{cal}, and so the MMOE for these estimates is expected to be $> 10\%$. Referring back to equation 15 allows the exact calculation of the MMOE of the residual variances, to infer how much less precision we would expect:

$$MMOE = \sqrt{max\left(\frac{\chi^2_{1-\frac{\alpha}{2},n-1}}{n-1}, \frac{n-1}{\chi^2_{\frac{\alpha}{2},n-1}}\right)}$$

$$MMOE = \sqrt{max\left(\frac{\chi^2_{1-\frac{\alpha}{2},175}}{175}, \frac{175}{\chi^2_{\frac{\alpha}{2},175}}\right)}$$

where $\chi^2_{1-\frac{\alpha}{2},175}$ and $\chi^2_{\frac{\alpha}{2},175}$ are the critical values of a χ^2 distribution with 175 degrees of freedom. This equates to

$$MMOE = \sqrt{max\left(\frac{214}{175}, \frac{175}{140}\right)} = \sqrt{1.25} = 1.12$$

Thus the error is expected to be 12%, only just over the 10% recommendation.

4.5.5 STEP 5: Summary of expected precision

A summary of the expected precision for each criteria, based on assessment in a dataset of 176 children, is given in Table 4.2. Given the relevant assumptions, the existing dataset appears to have a reasonable sample size for precisely estimating model calibration measures on external validation, though the MMOE for σ^2_{CITL} and σ^2_{cal} was higher than recommended, at 12%.

Estimated confidence interval widths for the CITL are narrower than the previously stated acceptable level of ± 1kg, though it is worth noting that the observed confidence interval for CITL in the external validation data was wider than expected. This suggests that some of the assumptions made, most likely relating to the distribution of fat-free mass in the validation population, did not hold. While the observed confidence interval was still sufficiently narrow, this difference in expected and observed widths demonstrates the importance of gaining a realistic estimates for the anticipated variance of the outcome variable, Y_i, in the target population, especially where the validation population differs from the model development population.

Furthermore, while 176 participants would achieve the recommended precision in estimating R^2_{val} where the R^2_{val} is large (as in the model development data), confidence interval width exceeded the recommend 0.1 when R^2_{val} was assumed to be 0.8. While this width was only just wider than recommended, it is likely the 176 children would not have been sufficient to achieve the desired precision in populations where the proportion of the outcome variance explained by the model is lower.

Table 4.2: Summary of the expected precision in model performance estimates for the Hudda fat-free mass prediction model, for an external validation of using 176 participants from the ALSPAC study

Criterion	Assumptions	Expected precision	Expected confidence interval width
(i) Precise estimate of R^2_{val}	$R^2_{val} = 0.948$	$SE_{\hat{R}^2_{val}} = 0.0076$	0.0299
	$R^2_{val} = 0.9$	$SE_{\hat{R}^2_{val}} = 0.0143$	0.0561
	$R^2_{val} = 0.8$	$SE_{\hat{R}^2_{val}} = 0.0270$	0.1057
(ii) Precise estimate of CITL	$R^2_{CITL} = R^2_{val} = 0.948$, $var(Y_i) = 0.089$	$SE_{\hat{CITL}} = 0.0051$	0.0201ln(kg) = 1.020kg
	$R^2_{CITL} = R^2_{val} = 0.9$, $var(Y_i) = 0.089$	$SE_{\hat{CITL}} = 0.0071$	0.0279ln(kg) = 1.028kg
	$R^2_{CITL} = R^2_{val} = 0.8$, $var(Y_i) = 0.089$	$SE_{\hat{CITL}} = 0.0101$	0.0394ln(kg) = 1.040kg
(iii) Precise estimate of λ_{cal}	$R^2_{cal} = R^2_{val} = 0.948$	$SE_{\hat{\lambda}_{cal}} = 0.0177$	0.0694
	$R^2_{cal} = R^2_{val} = 0.9$	$SE_{\hat{\lambda}_{cal}} = 0.0252$	0.0988
	$R^2_{cal} = R^2_{val} = 0.8$	$SE_{\hat{\lambda}_{cal}} = 0.3780$	0.1482
(iv) Precise $\hat{\sigma}^2_{CITL}$ and $\hat{\sigma}^2_{cal}$	–	$1.0 \leq MMOE \leq 1.12$	-

4.6 Discussion

4.6.1 Summary of key findings

This chapter discussed the development and possible applications of closed-form sample size calculations, aiming to target precise estimates of predictive performance for studies externally validating a prediction model with a continuous outcome. These calculations aim to ensure the sample size is large enough to precisely estimate key measures of predictive performance (R^2, CITL, and calibration slope), as well as precise estimation of the residual variances in calibration models. These requirements led to four criteria, with the largest sample size required across all four criteria being the recommended minimum sample size needed in the external validation dataset. This work builds on minimum sample size calculations for development of a prediction model with a continuous outcome [75].

As with any sample size calculation, assumptions are required to implement the proposed approach, without prior knowledge of what observed values will be. In particular, researchers must specify the values of $\hat{R}^2_{val}$, $\hat{var}(Y_i)$, and $\hat{\lambda_c al}$ expected in the external validation dataset. A simple starting point for these values is to assume similar performance to what was seen in the original model development study, thus using the same value as those reported for internal validation (after optimism-adjustment). This is likely to give an accurate assessment of the sample size needed if the target population for the external validation is similar to that used in the model development study, with similar predictive performance, though. Where populations and expected model performance differ, the researcher then might also consider required sample sizes based on adjustments to these expected values. In particular, adjustments to these assumed values must be made to allow for worse performance in the validation dataset, as is often the case in practice, and

for key differences in expected outcome distributions.

Lower values of performance measures, implying worse calibration and overall model fit in the external validation population, are important to consider where the model development dataset was small (where the reported predictive performance statistics were estimated with large uncertainty); the model development process did not include adjustment for overfitting (for example, using penalisation and shrinkage techniques), resulting in reported performance statistics are likely to be optimistic; and in situations where the intention is to validate the model in a different population or setting from that used in the development study. Larger sample sizes may also be needed if missing data are expected, or if the model's predictive performance in key subgroups (such as age groups, or in those of different ethnic origins) is of interest during assessments of algorithmic fairness.

The proposed calculations can also be used to estimate the expected precision in performance measures when aiming to externally validate the given model in an existing dataset of a fixed sample size. This would allow the researcher (and other key stakeholders) to gauge the expected precision of estimates conditional on the sample size available. Similarly, this approach could be used to identify the precision expected in predictive performance estimates within clinically relevant subgroups, if demographic details of the participants in the existing dataset are known. Ideally the dataset should be large enough to ensure precise estimates, as then more robust conclusions about predictive performance in the population of interest will be possible.

4.6.2 Strengths and limitations

Although the proposed sample size calculation targets precise estimates of calibration statistics (calibration slope, CITL), these values are summaries across the whole population, thus large variability in calibration curves across all individuals may still arise. Larger sample sizes may be needed to ensure that there is not excessive variability in this curve. Ideally calibration curves should be precise across the whole range of predicted values, though at a minimum should be precisely estimated across the range of values that is important for clinical decision making. For example, in a model concerned with identifying babies who are most likely to be born at a low birthweight, precise estimates of the calibration curve would be most important in the lower range.

A notable limitation of the approach currently proposed in this chapter is the difficulty in specifying appropriate values for the standard errors on external validation, especially for the model calibration statistics. The values of the standard error in the calibration slope, $SE_{\lambda_{cal}}$, and CITL, SE_{CITL}, that are needed to ensure precision in the calibration curve itself is hard to determine. This precision would be better assessed through visual assessment of simulated calibration curves for a given sample size.

Though these methods have not been considered here, future extensions to this work should involve consideration of precision in the calibration curve, in addition to precision in calibration statistics [96].

4.6.3 Conclusions and next steps

The sample size proposal presented in this chapter has been published in *Statistics in Medicine* [92]. Important extensions of this work include assessment of the necessary sample size for external validations of prediction models with non-continuous outcomes, building on the work of others [76, 172]. Simulation-based extensions to binary [95] and time-to-event [93] outcome settings were beyond the scope of this thesis, but have been investigated as a part of a wider research team and published subsequent to the work described in this chapter. Closed-form sample size calculations, as proposed here, are transparent and quick to implement, when compared to simulation-based approaches, but are more difficult to derive for binary and time-to-event outcomes. Approximate closed-form solutions for a binary outcome have been proposed [94], though closed-form alternatives to the calculation for external validation of a time-to-event model are not yet available. Future work might also consider an extension of the proposed criteria to include precise estimation of calibration curves in the validation of models to predict any outcome type, which was not considered in this chapter, nor in subsequent publications for binary or time-to-event outcomes [79, 188].

Is it important to note that a large sample size alone does not overcome issues in quality and applicability, and so meeting the proposed criteria in this chapter does not mean that the external validation alone can be used to recommend model use [65, 202, 111]. To draw valid conclusions about model performance, it is also important that model evaluations are high quality and applicable to the target population and setting where the model might be implemented in practice. Further, the criteria proposed in this chapter focus only on certain statistical measures of predictive performance, and not on clinical utility or the expected impact of using a model to inform healthcare decisions (for example, the initiation of treatment).

Furthermore, just as large sample sizes are no guarantee of validity, assessment of model performance in datasets that do not meet the criteria proposed here are not futile. Even when available validation datasets are small, obtaining estimates of predictive performance can still be highly useful. In particular, estimates could be combined in a future meta-analysis of model performance estimates on external validation: a situation that will be the focus of the next chapter of this thesis [65, 203].

CHAPTER 5

External validation of prediction models for birthweight and Fetal Growth Restriction (FGR) with complications: Individual Participant Data (IPD) meta-analysis

5 Chapter 5: External validation of prediction models for birthweight and Fetal Growth Restriction (FGR) with complications: Individual Participant Data (IPD) meta-analysis

5.1 Introduction and objectives

Chapter 4 discussed the importance of the sample used to externally validate a clinical prediction model with a continuous outcome, demonstrating how the number of participants included in a validation directly influences the precision then seen in estimates of key performance statistics. Many external validations are conducted in a relatively small dataset from a single setting, which, while giving useful information on out-of-sample performance, may mean a high level of uncertainty in that observed performance. For example, the external validation of the pain intensity score model in Chapter 3 was conducted in a small sample from a single, previously recruited randomised trial [91], and Hudda's initial external validation of their fat-free mass model, discussed in Chapter 4, took place in a pre-existing dataset of a modest, fixed size [17]. In both of these cases, the available data gave confidence intervals for at least one of the performance statistics that was wider than the recommended level, that was too wide to make clear conclusions about the model's predictive performance.

However, as mentioned in Chapter 4, estimates of predictive performance can still be highly useful even when validation datasets are small, if information across multiple validations can then be combined. Indeed, this was the next step for Hudda et al, who followed up their model development and initial external validation with an assessment of their model's performance across a variety

of different populations, from 19 existing datasets worldwide [203]. The new external validation datasets range in size from just 42 individuals (less than a fifth of the recommended minimum), up to 1010 (over seven times higher than the minimum required to gain precise estimates of performance measures).

Figure 5.1 demonstrates how the majority (12/19, 63%) of the available datasets contained sample sizes smaller than the recommended minimum, though 95% (18/19) met at least one of the proposed criteria. Clearly, when data were combined across populations, the information available was well over the required 235 participants for external validation, though it is worth noting that the criteria in Chapter 4 target only precise estimates of within study performance and do not include allowance for variations in performance between studies. To account for this additional variance, the methods used to meta-analyse performance measures and their associated standard errors across studies must be considered when considering what constitutes a precise estimate [65, 63, 204].

External validations in data from a single pre-existing source often have limited ability to assess the transportability of the model in question. Indeed, for the examples in Chapters 3 and 4, the external validation data was generally of a very similar demographic composition to that of model development. There are many potential causes for heterogeneity in model performance across populations, thus promising results in an external validation in one population might not imply the model works well elsewhere. Just as Hudda was able to assess model performance on external validation across 19 different datasets, comprising children from very different populations, this chapter will employ appropriate meta-analysis methods to examine the performance of clinical prediction models across different settings.

Within this chapter, models for predicting birthweight or FGR risk (identified in Chapter 2) will be externally validated using data from multiple cohorts, to demonstrate how confidence in the calculated performance estimates varies with sample size, and to discuss the level of precision seen when combining performance estimates from external validations across differing populations.

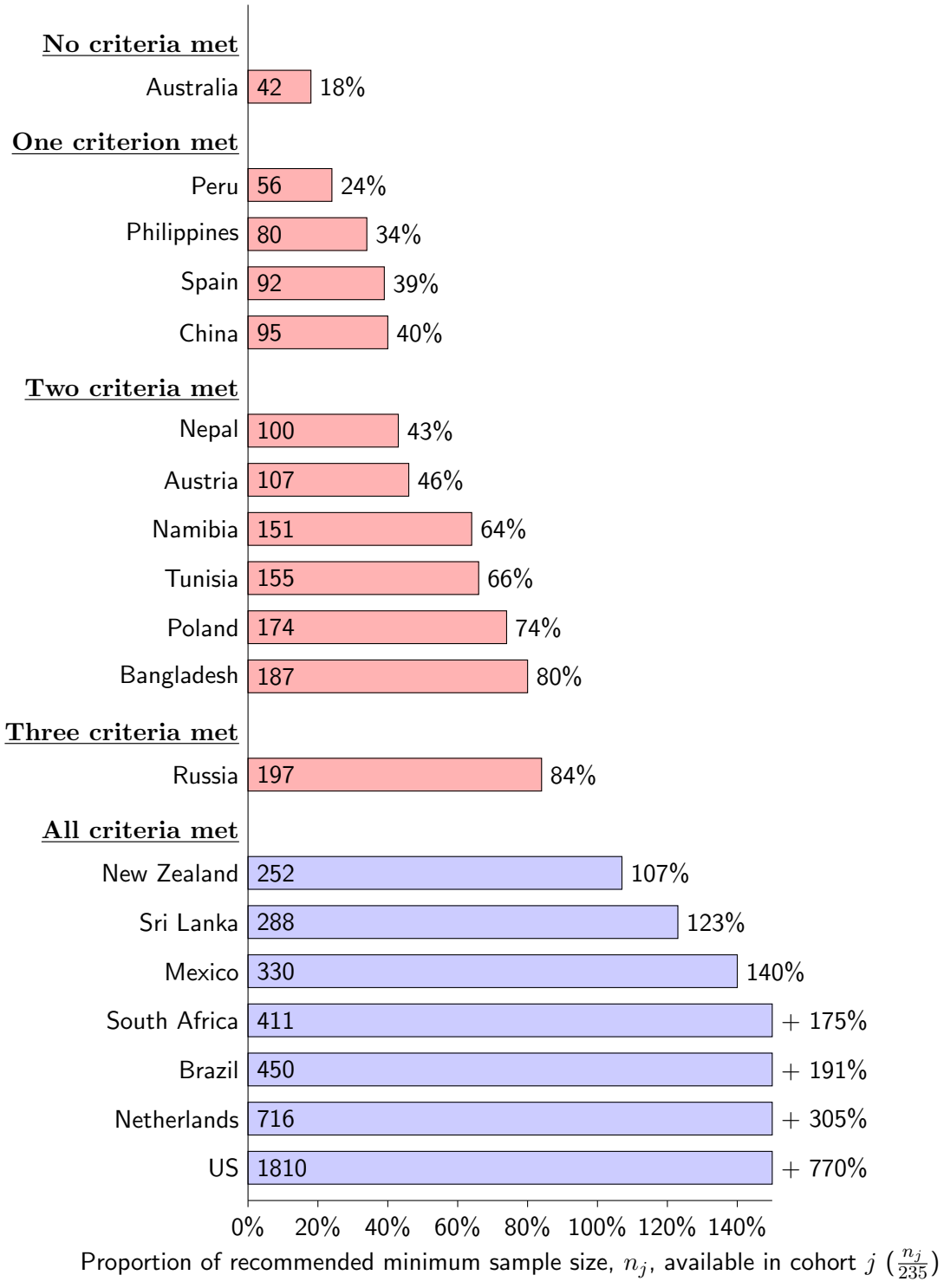


Figure 5.1: Bar chart showing the sample size in each of Hudda's datasets, relative to the minimum recommended for a single external validation. Colours indicate those with a sample size greater (blue) and lower (red) than the recommended minimum from Chapter 4.

5.1.1 Clinical scenario

The HTA funded IPPIC (International Prediction of Pregnancy Complications) collaborative network consists of Individual Participant Data (IPD) from 14 UK and 66 international datasets, including information on around 3 million pregnancies in total. These data contain information on many important predictors of birthweight, and of FGR with associated complications, along with measurements of those two outcomes, so were ideal for the external validation of models to predict FGR that were identified and discussed in Chapter 2 [109].

Models that aim to predict rare outcomes, such as serious pregnancy complications including FGR, require large sample sizes for external validation, which may not be possible in just one dataset. Robust external validation across multiple populations and settings is of great importance in settings such as these, meaning this is an ideal clinical setting in which to explore the potential benefits of increasing sample size for external validation through use of IPD from multiple existing studies.

The external validation discussed below formed part of an Health Technology Assessment (HTA) report "*External validation and development of prediction models for fetal growth restriction (FGR) and birthweight: an Individual Participant Data (IPD) meta-analysis and cost-effectiveness analysis*", along with subsequent research involving model development and internal-external cross-validation (summarised in the Appendices to this chapter). This report is due for publication in 2024.

5.1.2 Objectives

Broadly, this chapter discusses the external validation of all applicable published models for predicting birthweight, or FGR with severe complications (birthweight less than 10th percentile adjusted for gestational age at delivery, complicated by stillbirth or neonatal death or delivery before 32 weeks). Model performance was assessed in datasets included in the IPPIC collaboration data collection, where relevant predictors and outcomes were measured.

Thus, this chapter has a number of applied and methodological aims, as follows:

1. assess predictive performance of published models for predicting birthweight or FGR with severe complications, using measures of calibration, discrimination, and clinical utility.
2. demonstrate how external validation of a prediction model in multiple IPD datasets can be used to determine both the generalisability and the transportability of the model
3. show how analysing the information across multiple validations as an IPD meta-analysis can boost the available sample size for model validation.

The sample size of individual datasets and the combined data as a whole were compared to sample size recommendations for external validation of continuous outcome models, proposed in Chapter 4, and how this is reflected in the precision of the model performance estimates is discussed.

The method described in Chapter 3, for gaining predicted probabilities from a model for predicting a continuous outcome, are further utilised. This chapter includes a demonstration of how this method can be combined with traditional decision curve analysis to explore the net benefit of

using birthweight prediction models in the management of patients at high risk of FGR.

5.2 Methods

5.2.1 Identifying existing models to predict birthweight or FGR

Details of the literature search and eligibility criteria for the identification of models to predict birthweight and FGR with complications are given in Chapter 2. The models identified in the review from Chapter 2 were further restricted, so that those eligible for external validation were only those which predicted either birthweight on its continuous scale, or FGR outcomes adhering to a definition agreed by consensus among the clinical members of the research team. This definition was as follows: birthweight less than 10th percentile adjusted for gestational age at delivery, complicated by stillbirth, or neonatal death, or delivery before 32 weeks. No existing models met this definition of FGR, therefore this chapter reports only on the external validation of models for predicting continuous values of birthweight.

5.2.2 Identifying available datasets for external validation of existing models

All eligible models were externally validated, where possible, using IPD from studies included in the IPPIC collaboration data collection, that was provided for the purposes of this research project. In the formation of this data collection, primary studies identified during the review discussed in Chapter 2, among others, were invited to share their data. Invitations were also extended to investigators of primary studies and population-based cohorts not included in review, but identified through links within the wider collaborative group. The IPPIC collaboration data collection has previously been used in the external validation of models to predict other pregnancy complications, such as pre-eclampsia [205] and stillbirth [206]. For the analyses discussed in this chapter, no restrictions were placed on the type of study design that would be eligible.

Identified models could only be externally validated in datasets where the available information included all necessary model variables needed to calculate predictions, and information regarding the relevant outcome (birthweight). For the assessment of birthweight prediction as a proxy for FGR, validation datasets were also required to contain the necessary information to determine the presence of FGR with complications, namely gestational age at delivery, presence of stillbirth, or neonatal death.

5.2.3 Calibration performance measures

To assess the performance of each model using the IPD for each study cohort, each model equation was applied to each participant to calculate the predicted birthweight value for that individual, conditional on their predictor values, as described in previous chapters. The calibration statistics, discussed previously and recapped below, were calculated within each cohort separately and for each model separately. Performance for each model was then summarised across studies using the meta-analysis methods described in Section 5.2.6.

Calibration-in-the-large

This measure indicates the extent to which the predicted birthweight values calculated from the model are systematically too high or too low (on average, across all individuals). The estimate of CITL and its standard error were calculated by fitting the calibration model $\text{Birthweight}_i = \alpha + \beta(Y_{PREDi})$ where Birthweight_i is the observed birthweight for individual i, Y_{PREDi} is the predicted birthweight from the model, and α is the estimate of CITL when β is constrained to equal one (fitted using a regression constraint). The ideal value of CITL is zero, which would imply that the predicted birthweight values were on average no higher or lower than

the observed values.

Calibration slope

The calibration slope indicates whether there is agreement between observed and predicted birthweight outcome values across the range of predicted risks. The calibration model, $\text{Birthweight}_i = \alpha + \beta(Y_{PREDi})$ was fitted, this time with no constraint on the value of β. This β value gave the estimate of the calibration slope. Ideally, the calibration slope should be equal or very close to one for good calibration across the full range of predicted probabilities, implying that for every 1g increase in predicted birthweight we expect a corresponding 1g increase in observed birthweight. A slope < 1 suggests predicted birthweight values are too extreme (predicted birthweight values that are high are too high, while those that are low are too low) and may indicate overfitting of the original model to the development data. A slope value > 1 indicates the range of predicted birthweight values is too narrow, when compared to the observed values for the same individuals.

Calibration plots

These plots, as seen in previous chapters, show a scatter of the observed to expected birthweight values for each participant. A lowess smoother was calculated across all participants, and shown on the same axes as the scatter, to demonstrate the shape of the overall calibration curve across the full range of possible predicted birthweight values.

Calibration plots (with calibration curves) were generated in each cohort separately, for each imputation. Where prediction distributions and calibration plots were consistent across

imputations, it was concluded that predictions were similar enough for pooling across imputations to be appropriate, for visualisation purposes. Where plots were not similar across imputations, a selection of example plots were to be displayed, though this was not necessary in practice as imputations were sufficiently consistent on visual inspection to allow pooling in all cases.

5.2.4 Decision curve analysis

Decision curve analysis (DCA) was introduced in Chapter 1, and is an approach used to evaluate and compare prediction models in terms of clinical utility [70, 69, 71]. DCA identifies whether there is an overall benefit of using a prediction model to guide treatment decisions within clinical practice, based on various decision thresholds of predicted outcome risk. This potential benefit is considered net of any harm arising from misclassification of individuals as being at high risk where no outcome event was then observed. Prediction models are evaluated across a range of different probability thresholds, where those with predicted outcome risks above the given threshold would receive an altered treatment pathway.

The net benefit of using birthweight prediction model to identify pregnancies at high risk of being Small for Gestational Age (SGA, defined as being in the lowest 10% birthweights for their observed gestational age at delivery), a proxy for FGR, at given threshold probabilities was calculated. To obtain the decision curve, the prediction model is evaluated over a range of different probability thresholds, where the threshold is taken as a point above which a patient would be treated for being at high SGA risk, and below which a patient would receive usual care. Decision curves for the model were then compared to the net benefit expected to arise from using alternative "treat all" and "treat none" strategies to allocate treatment [207]. These alternative strategies

corresponded to hypothetical situations where every pregnancy was treated as though it were at high risk of being SGA ("treat all"), or where no one was treated as though they were at high SGA risk ("treat none"). Similarly, decision curves can also be plotted for multiple models on the same axes to facilitate comparison, and to help decide which model offers the most clinical benefit.

As noted, traditional DCA methods incorporate the use of threshold probabilities to assign treatment group, and thus require an outcome on a binary scale to allow for probability calculation. Predicted probabilities of being SGA were gained from the continuous predicted birthweight value (obtained from the prediction model being evaluated) following the methods described in Chapter 3. Given gestational age at delivery was a key factor in determining SGA based on birthweight, but in practice would be unknown at the time of model implementation, two different methods were considered for assessing model benefit. When determining SGA risk from continuous birthweight predictions, the following definitions were assessed:

1. Lower than the 10th percentile for 40 weeks gestation: Predictions were standardised with respect to gestational age at delivery, with all individuals having their predicted birthweight calculated at 40 weeks' gestation. The predicted individual-level probabilities were calculated and compared to whether the observed birthweight was lower than the 10th birthweight percentile value for their observed gestational age at delivery, recorded in completed weeks.

2. Lower than the 10th percentile for observed gestational age at delivery: Predicted birthweight was based on observed gestational age at delivery, a value that would not be available when the model would be used to make a prediction in practice. The probability of the predicted birthweight being lower than the 10th percentile for the true gestational age at delivery was calculated and compared to whether the observed birthweight was lower than the 10th

percentile value for their observed gestational age at delivery, recorded in completed weeks.

The comparison to relevant percentile values was conducted with reference to a normal range of birth weights for gestational age reported by Poon et al in 2016 [208], from 92,018 live births across two UK-based hospitals. The relevant percentile values used in determining FGR in the external validation populations are shown in Table 5.1.

Table 5.1: Normal range of birthweights (in grams) according to gestational age (GA) at delivery in 92,018 live births, reported by Poon et al 2016. Values are reported for 1st, 3rd, 5th and 10th percentiles, representing varying degrees of smallness

	Percentile			
GA at delivery (weeks)	1st	3rd	5th	10th
24	508	551	574	609
25	521	566	590	626
26	556	606	633	674
27	612	671	701	749
28	689	757	793	849
29	783	863	906	971
30	896	988	1038	1113
31	1024	1130	1186	1273
32	1167	1286	1349	1447
33	1322	1454	1524	1632
34	1488	1631	1707	1825
35	1660	1815	1896	2022
36	1837	2001	2087	2221
37	2014	2185	2276	2416
38	2187	2365	2459	2604
39	2351	2534	2631	2780
40	2501	2688	2787	2939
41	2633	2822	2922	3077
42	2740	2931	3033	3188
43	2818	3010	3112	3269

5.2.5 Missing data

The numbers of missing values for each required predictor variable in each cohort were summarised, and preliminary checks for associations between missingness and predictor values were conducted to check for obvious violations of the missing-at-random assumption. Where a variable required for calculating model predictions was not present within an individual cohort, or was present in fewer than 10% of participants, this variable was considered to be systematically missing. Although methods have been proposed to impute values for systematically missing variables based on the IPD from other studies [209, 210], for practical reasons, imputation was not performed for such systematically missing information in this case. Thus, only studies that recorded information (for at least 10% participants) on all required predictors were used in the validation of any given prediction model.

Multiple imputation by chained equations was used to estimate values for partially missing predictor and outcome variables within individual cohort separately, to maintain the clustering of participants within studies and to preserve heterogeneity between cohorts [211]. Missing values for continuous variables were imputed using linear regression, for binary variables using logistic regression, and for categorical variables using predictive mean matching [173]. Other variables were included in the imputation models as auxiliary variables, such that the imputation model included all candidate predictors, along with both birthweight and FGR outcome variables. Observations with imputed outcomes were then deleted prior to analysis [176, 212].

To ensure that the number of imputed datasets would be at least equal to the percentage of incomplete observations, 100 imputations were generated in all cohorts to exceed the largest percentage of incomplete observations in any individual cohort [173]. Imputations were assessed for

consistency by comparing density plots, histograms, and summary statistics across imputations and back to the complete values, within and across studies. Performance statistics were summarised across imputations using Rubin's rules [177], where appropriate, to obtain one estimate and standard error (SE) for each performance statistic in each cohort, prior to meta-analysis of performance estimates across cohorts [213].

Complete case analysis, assessing model performance only in individuals with complete information on all predictors and outcome, was conducted for completeness, as a sensitivity analysis. The main results from the complete-case analyses are included as an appendix to this chapter.

5.2.6 Data synthesis

Meta-analysis methods were used to summarise a model's performance across all IPD datasets used for the external validation. Random-effects meta-analysis was used to allow for differing true model performance (heterogeneity) across cohort populations, for example, due to differing case-mix and overall risk [65, 63]. Heterogeneity in model performance across studies was summarised using the estimates of τ^2, with approximate 95% prediction intervals calculated using Higgins' approach [214], giving an indication of likely model performance in a new cohort from a similar setting [215].

Predictive performance measures were summarised across the IPD datasets using a two-stage IPD meta-analysis approach: performance measures and their variances were first estimated for each cohort separately and then pooled using restricted maximum likelihood estimation of the aforementioned random-effects meta-analysis model, which weights cohort contributions by a combination of the within-study and between-study variances [216]. The calibration slope and

calibration-in-the large were pooled on their original scales [204], giving the average and 95% confidence interval for the average of each performance statistic. These confidence intervals were derived using a Hartung-Knapp-Sidik-Jonkman variance correction, to account for uncertainty in variance and heterogeneity estimates due to relatively few studies being present in the meta-analysis [217].

Model performance across cohorts is shown graphically using forest plots for each performance statistic and through scatter plots to show both measures of calibration in combination,to give a view of the overall calibration performance of the model.

5.2.7 Other considerations

Participants may have been included in a dataset multiple times if they had more than one pregnancy during the study period. For the purpose of external validation, models were validated for each pregnancy of each participant separately. Although two or more pregnancy outcomes from the same women are likely to be correlated in reality, the number of multiple pregnancies was expected to be very small relative to the total number of pregnancies. Considering each pregnancy as a distinct observation meant that no allowance was made for correlation between multiple pregnancies of the same woman in neither the model performance statistics, nor the contribution of such pregnancies to the validation sample size (given correlated pregnancies within the validation data would impact the observed precision).

5.3 Results

5.3.1 Identified models for external validation

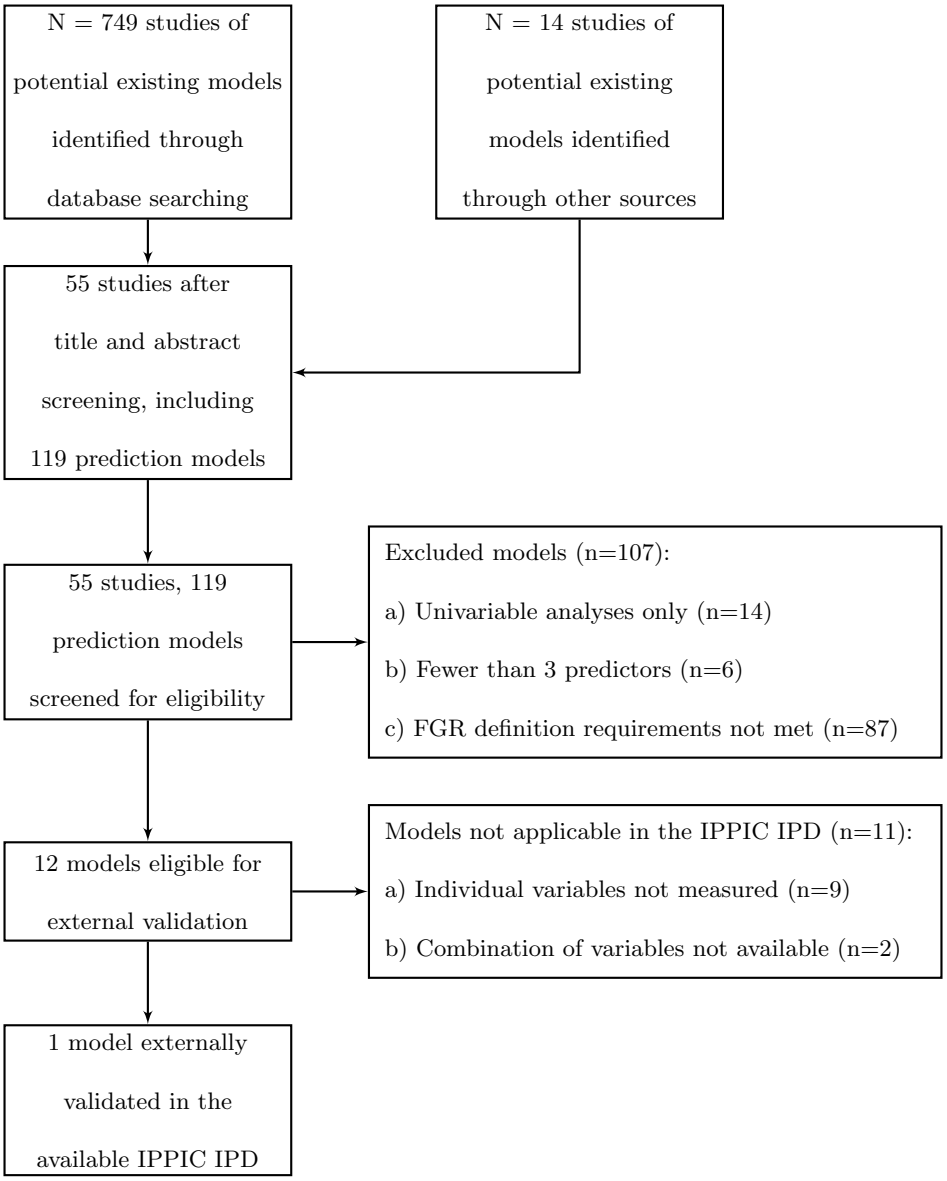


Figure 5.2: PRISMA flowchart, as shown in Chapter 2, extended to demonstrate birthweight/FGR models available for external validation

Of the prediction models identified in Chapter 2, none were found to adhere to the binary FGR definition included above, thus all 87 of the binary outcome models were excluded from the external validation analyses. This left 12 models of potential interest, all predicting birthweight on a continuous scale, based on a combination of maternal and pregnancy features.

5.3.2 Available datasets for external validation

Cohorts within the IPPIC collaboration data collection were compared to the list of necessary predictors for implementation of each of the identified 12 birthweight models, to discern which contained all the necessary information to be included in the external validation. Unfortunately, of these eligible models, nine could not be externally validated as they contained variables (such as biparietal diameter, or birthweight from previous pregnancies) that were not measured in any of the individual IPPIC FGR cohorts. A further two models could not be included for external validation as they included variables that, although all available within the IPPIC collaboration data collection, were not present in combination in any single cohort.

Thus, despite the huge effort that went into obtaining IPD from multiple studies, external validation in the IPPIC FGR data was possible for only one model, for predicting birthweight on its continuous scale. This model was published by Poon et al in 2011, and included predictor variables for gestational age at delivery, mother's weight, mother's height, mother's age, ethnic origin, key comorbidities (chronic hypertension, diabetes), and whether the pregnancy was a result of assisted conception [208].

5.3.3 External validation of the Poon 2011 birthweight model

The remainder of the results in this chapter describe the external validation of the Poon 2011 model, with predictions of birthweight in the external validation populations calculated using the below equation, where: predictor variable GA refers to gestational age at delivery (measured in weeks); height, weight, and age of the mother were measured on their continuous scale, in centimetres (cm), kilograms (kg), and years, respectively; and the remaining predictors were binary variables, with value of one indicating that the individual has the feature in question, and a value of zero indicating that the feature was absent.

$$\begin{aligned}\log_{10}\text{Birthweight} = &-0.935219 + 0.186853(\text{GA}) - 0.002078(\text{GA})^2 \\ &+ 0.003726(\text{weight}) - 0.000030(\text{weight})^2 + 8.820640e^{-08}(\text{weight})^3 \\ &+ 0.000965(\text{height}) + 0.001466(\text{age}) - 0.000026(\text{age})^2 \\ &+ 0.016986(\text{if parous}) - 0.024867(\text{if current smoker}) \\ &- 0.021769(\text{if African ethnicity}) - 0.017824(\text{if South Asian ethnicity}) \\ &- 0.005543(\text{if East Asian}) - 0.009063(\text{if mixed ethnicity}) \\ &- 0.020995(\text{if chronic hypertension present}) + 0.03143(\text{if diabetes present}) \\ &- 0.004015(\text{if assisted conception})\end{aligned}$$

A total of nine individual cohorts were identified that included all required predictor variables and outcome information for external validation of the Poon 2011 model, these were: STORKG [218], Allen [219], Odibo [220], Baschat [221], Rumbold [222], POP [223], Generation R [224], ALSPAC [200], and Chie [225]. Key cohort features, including recruitment location and time period, and available sample size are shown in Table 5.2 and Figure 5.3, respectively.

Table 5.2: Features of external validation cohorts for the Poon 2011 model

Cohort	Location	Study type	Time period	Population
STORKG [218]	Norway	Prospective cohort	2008-2010	Unselected
Allen [219]	UK	Prospective cohort	2010-2014	Unselected
Odibo [220]	US	Prospective cohort	2009-2011	Unselected
Baschat [221]	US	Prospective cohort	2007-2010	Unselected
Rumbold [222]	Australia	Randomised trial	2001-2005	Low risk only
POP [223]	UK	Prospective cohort	2008-2012	Unselected
Generation R [224]	Netherlands	Prospective cohort	2002-2006	Unselected
ALSPAC [200]	UK	Prospective cohort	1991-1992	Unselected
Chie [225]	Japan	Prospective registry	2013-2014	Unselected

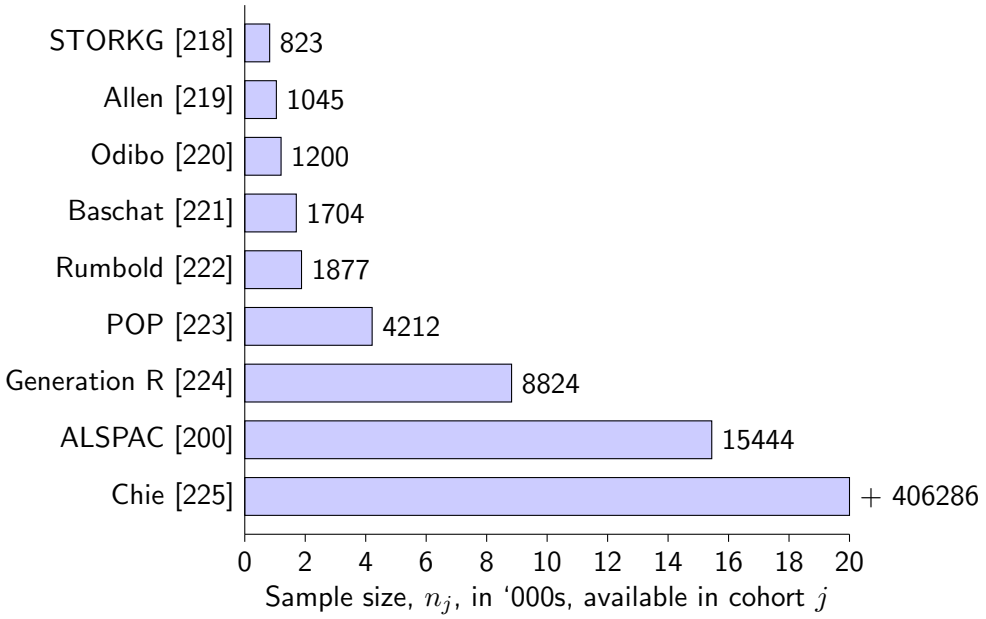


Figure 5.3: Bar chart showing the available sample size in each validation cohort.

5.3.4 Sample size requirements for external validation

Precisely estimate R^2_{val} (criterion (i))

The first stage of the calculation discussed in Chapter 4 requires an assumed value $\hat{R}^2_{val}$, and uses a $SE_{\hat{R}^2_{val}}$ of 0.0255 to target a confidence interval width of 0.1. From the model performance in the development data, $R^2_{val} = 0.625$ was assumed, along with the lower assumed model fit of $R^2_{val} = 0.5$

	$\hat{R}^2_{val} = 0.625$	$\hat{R}^2_{val} = 0.5$
n	$= \frac{4R^2_{val}(1-R^2_{val})^2}{SE^2_{\hat{R}^2_{val}}}$	$= \frac{4R^2_{val}(1-R^2_{val})^2}{SE^2_{\hat{R}^2_{val}}}$
	$= \frac{4*0.625*(1-0.625)^2}{0.0255^2}$	$= \frac{4*0.5*(1-0.5)^2}{0.0255^2}$
Sample size to meet criterion (i)	$= 541$	$= 769$

Precisely estimate calibration-in-the-large (criterion (ii))

The second step in the calculation requires specification of a value for $\hat{var}(Y_i)$, the anticipated variance of outcome values in the external validation population. Given the clinical importance of birthweight predictions on their original scale, the Poon 2011 model's external validation performance on this scale was of primary interest (with performance statistics calculated after transforming predictions back to the gram-scale). Thus, calculations of required sample size for precise estimation of CITL was based on values for birthweight in grams, rather than for log_{10}Birthweight, to ensure appropriate precision on this scale.

The authors of the development study did not report the standard deviation of the model outcome (log_{10}Birthweight) in their model development data, nor did they report the standard deviation of birthweight on its original scale, thus information from the model development population could not be used to inform the calculation in this case. Given the external validation population was

known prior to analysis, with the data already available, an estimate of the standard deviation of birthweight in the validation population could be calculated. The standard deviation of birthweight on the grams scale was 594g.

The precision required to estimate CITL again needs to be placed in context of the mean birthweight value in the population ($\bar{Y} = 2722$g). Considering this scale, an accuracy of around $\pm$50g around $\bar{Y}$ seems an acceptable level of precision, which corresponds to a target $SE_{\hat{CITL}}$ of about 25.5g.

In applying the equation, the expected value for R^2_{CITL} was first assumed to equal $R^2_{val} = 0.625$, as in the published model development, with the more conservative estimate of $R^2_{CITL} = 0.5$ also assessed:

	$\hat{R}^2_{CITL} = 0.625$	$\hat{R}^2_{CITL} = 0.5$
n	$= \frac{var(Y_i)(1-R^2_{CITL})}{SE^2_{\hat{CITL}}}$	$= \frac{var(Y_i)(1-R^2_{CITL})}{SE^2_{\hat{CITL}}}$
	$= \frac{594^2*(1-0.625)}{25.5^2}$	$= \frac{594^2*(1-0.5)}{25.5^2}$
Sample size to meet criterion (ii)	$= 204$	$= 272$

Precisely estimate calibration slope (criterion (iii))

Targetting a confidence interval width for the calibration slope of at most 0.2 ($SE_{\hat{\lambda}_{cal}} = 0.051$) and assuming good calibration ($\hat{\lambda}^2_{cal} = 1$), with $\hat{R}^2_{cal} = \hat{R}^2_{val}$:

	$\hat{R}^2_{cal} = 0.625$	$\hat{R}^2_{cal} = 0.5$
n	$= \frac{\lambda^2_{cal}(1-R^2_{cal})}{SE^2_{\hat{\lambda}_{cal}} R^2_{cal}} + 1$	$= \frac{\lambda^2_{cal}(1-R^2_{cal})}{SE^2_{\hat{\lambda}_{cal}} R^2_{cal}} + 1$
	$= \frac{1*(1-0.625)}{0.051^2*0.625}$	$= \frac{1*(1-0.5)}{0.051^2*0.5}$
Sample size to meet criterion (iii)	$= 232$	$= 386$

Precisely estimating residual variances (criterion (iv))

As shown in the previous chapter, to ensure sufficiently precise residual variance estimates from the calibration models (10% margin of error), at least 235 participants are required regardless of clinical context.

Minimum sample size to meet all criteria simultaneously

Thus, the minimum sample size to ensure precise estimation of R^2, CITL, calibration slope, and the residual variances of the calibration models was 541 pregnancies, assuming the overall model fit would be consistent between model development and external validation populations. When anticipating poorer model fit on external validation, with an $R^2_{val} = 0.5$, this minimum sample size requirement increases to 769 pregnancies.

The sample sizes required to meet each of the criteria are summarised in Table 5.3. Not only did the combined external validation IPD across eligible cohorts easily surpass these requirements, with a total of 441,415 pregnancies included, all individual cohorts had sample sizes greater than the recommended minimum even where assuming the lower value of $R^2_{val} = 0.5$.

Table 5.3: Summary of the sample size calculation for external validation of the Poon 2011 birthweight prediction model

Criterion	Target precision	Assumptions	Minimum sample size required
(i) Precise estimate of R^2_{val}	$SE_{\hat{R}^2_{val}} = 0.0255$	$R^2_{val} = 0.625$	541
		$R^2_{val} = 0.5$	769
(ii) Precise estimate of CITL	$SE_{\hat{CITL}} = 25.5$	$R^2_{CITL} = R^2_{val} = 0.625$, $var(Y_i) = 594^2$	204
		$R^2_{CITL} = R^2_{val} = 0.5$, $var(Y_i) = 594^2$	272
(iii) Precise estimate of λ_{cal}	$SE_{\hat{\lambda}_{cal}} = 0.051$	$R^2_{cal} = R^2_{val} = 0.625$	232
		$R^2_{cal} = R^2_{val} = 0.5$	386
(iv) Precise $\hat{\sigma}^2_{CITL}$ and $\hat{\sigma}^2_{cal}$	$1.0 \leq MMOE \leq 1.1$	–	235
Minimum sample size			**769**

5.3.5 External validation cohort characteristics

As would be expected due to differences in time period and countries for recruitment, population demographics varied across the nine external validation cohorts. Some notable differences include the proportion of mothers smoking, ranging from a low of only 2.7% (Chie) up to 19.4% (Rumbold). The proportion of nulliparous women across cohorts varied, in some cases by design, with Rumbold and POP containing only nulliparous women (0% parous) as a part of their recruitment criteria.

Ethnic origins of mothers also varied across populations, with mothers from most European (ALSPAC, Generation R, STORKG and POP), Australian (Rumbold), and North American (Odibo) cohorts being predominately white. Allen, a UK-based cohort recruiting from a hospital in East London, included 47% South Asian mothers, while 47% mothers in Baschat were of Black ethnic origin. As the recruitment for the Chie cohort was conducted in Japan, 100% of mothers were reported as being of Eastern Asian origin, the most highly represented ethnicity in the external validation data.

Both Hispanic and Mixed ethnicities were poorly represented in the available external validation data, with the maximum representation in an individual cohort being only 2% for each (Odibo). The level of detail on ethnic groups varied across the IPD from different cohorts, thus it is possible that some of those women included in the "other ethnicity" group in some cohorts might have been better suited to other category. This information was not available in sufficient detail in the IPD to be sure of alternative classifications, thus those coded as "other" in the IPD were retained in this group for analysis.

Table 5.4: Characteristics of the studies used in the external validation of the Poon 2011 prediction model. Values are number (percentage) unless otherwise stated.

	Allen [219]	ALSPAC [200]	Baschat [221]	Generation R [224]	Odibo [220]	Rumbold [222]	Chie [225]	STORKG [218]	POP [223]
n	1,045	15,444	1,704	8,824	1,200	1,877	406,286	823	4,212
GA, median (LQ-UQ)	40 (39.3 to 40.6)	40 (39 to 41)	39.1 (37.9 to 39.9)	40.1 (39 to 41)	39.1 (38 to 39.6)	40 (39 to 41)	38 (37 to 40)	40 (38.9 to 40.9)	40.3 (39.1 to 41.1)
Weight, median (LQ-UQ)	62 (55 to 69)	55 (50 to 60)	71.8 (61.3 to 87.9)	67 (60.5 to 76)	68.9 (59.9 to 83.9)	66 (58.5 to 76)	52 (47 to 57)	64.6 (56.9 to 72.9)	66 (59 to 75)
Height, mean (SD)	161.5 (7.4)	164.3 (6.8)	164 (7)	167.2 (7.4)	164.6 (6.8)	165.3 (6.7)	158.3 (5.5)	163.6 (6.7)	165.2 (6.4)
Age, mean (SD)	29.9 (5.1)	27.8 (4.9)	30.2 (6.5)	29.7 (5.3)	31.5 (5.6)	26.4 (5.7)	32.2 (5.4)	29.9 (4.9)	29.9 (5.1)
Parous	461 (44.11)	5828 (37.74)	736 (43.19)	4834 (54.78)	518 (43.17)	0 (0)	210896 (51.91)	381 (46.29)	0 (0)
Smoking	38 (3.64)	2645 (17.13)	162 (9.51)	1438 (16.3)	97 (8.08)	364 (19.39)	10952 (2.7)	50 (6.08)	211 (5.01)
Ethnicity									
White	398 (38.09)	12075 (78.19)	775 (45.48)	4933 (55.9)	735 (61.25)	1777 (94.67)	-	379 (46.05)	3900 (92.59)
Black	108 (10.33)	131 (0.85)	803 (47.12)	2146 (24.32)	325 (27.08)	3 (0.16)	-	62 (7.53)	25 (0.59)
South Asian	495 (47.37)	113 (0.73)	88 (5.16)	496 (5.62)	94 (7.83)	1 (0.05)	-	200 (24.3)	91 (2.16)
Eastern Asian	-	-	-	-	-	-	406286 (100)	-	-
Hispanic	-	-	27 (1.58)	-	23 (1.92)	1 (0.05)	-	12 (1.46)	-
Mixed	12 (1.15)	-	-	-	22 (1.83)	4 (0.21)	-	-	1 (0.02)
Other	30 (2.87)	82 (0.53)	11 (0.65)	767 (8.69)	1 (0.08)	87 (4.64)	-	170 (20.66)	195 (4.63)
Chronic hypertension	10 (0.96)	1822 (11.8)	162 (9.51)	147 (1.67)	109 (9.08)	9 (0.48)	3421 (0.84)	13 (1.58)	220 (5.22)
Diabetes	11 (1.05)	126 (0.82)	81 (4.75)	33 (0.37)	58 (4.83)	8 (0.43)	2926 (0.72)	-	16 (0.38)
Assisted conception	23 (2.2)	365 (2.36)	35 (2.05)	140 (1.59)	59 (4.92)	49 (2.62)	57082 (14.05)	13 (1.58)	184 (4.37)
Birthweight, mean (SD)	3298.3 (524.5)	3347.7 (608.7)	3147.5 (674.6)	3391.1 (578.4)	3227.9 (676)	3382 (608.9)	2840.4 (581.1)	3418.3 (570.1)	3401 (534.5)

GA - Gestational Age, LQ - Lower Quartile, UQ - Upper Quartile, SD - Standard Deviation. Scales: GA (weeks), weight (kg), height (cm), age (years), birthweight (g).

5.3.6 Missing data

All cohorts had missing data in measurements of at least one of the predictor values, though the level of missingness varied considerably across cohorts. Four of the cohorts were over 90% complete (Allen, Baschat, Odibo, POP), meaning complete information on predictor and outcome values were available in at least 90% of the women included.

Two cohorts had missing information for at least one predictor or the outcome in more than half of their participants. In STORKG, this was driven by missing measurements for mother's weight, which were missing for 51% of women. The greatest proportion of observations with missing information was seen in ALSPAC, which only had complete data for 11% of women. Within the ALSPAC cohort mother's height (63%) and mother's weight (66%) were the most commonly missing predictor variables.

Birthweight, the outcome of interest, was missing for at least some women in every cohort. The level of missingness ranged from only 0.05% missing (Chie) up to 67% (ALSPAC). With the exception of the ALSPAC cohort, birthweight was at least 95% complete in all cohorts.

No notable associations were seen between the missingness of a given predictor variable and the values of other predictors or the outcome (where reported), thus there were no obvious violations of the missing-at-random assumption. Multiple imputation was therefore considered a reasonable method to account for missing information in this case.

Table 5.5: A summary of missingness by variable, for cohorts used to validate the Poon 2011 model. Values are number (percentage) missing.

	Allen [219]	ALSPAC [200]	Baschat [221]	Generation R [224]	Odibo [220]	Rumbold [222]	Chie [225]	STORKG [218]	POP [223]
n	1,045	15,444	1,704	8,824	1,200	1,877	406,286	823	4,212
Complete	99%	11%	99%	78%	95%	89%	73%	46%	96%
Gestational age	1 (0.1)	1334 (8.64)	3 (0.18)	5 (0.06)	24 (2)	-	180 (0.04)	22 (2.67)	-
Weight	5 (0.48)	10156 (65.76)	-	41 (0.46)	23 (1.92)	103 (5.49)	54600 (13.44)	421 (51.15)	146 (3.47)
Height	-	9735 (63.03)	-	33 (0.37)	8 (0.67)	138 (7.35)	34978 (8.61)	-	6 (0.14)
Age	1 (0.1)	2107 (13.64)	-	2 (0.02)	1 (0.08)	-	1147 (0.28)	-	-
Parous	-	2505 (16.22)	-	108 (1.22)	1 (0.08)	-	1586 (0.39)	-	-
Smoking	-	2736 (17.72)	-	1123 (12.73)	9 (0.75)	39 (2.08)	84755 (20.86)	-	-
Ethnicity	2 (0.19)	3043 (19.7)	-	482 (5.46)	-	4 (0.21)	-	-	-
Chronic hypertension	-	3095 (20.04)	-	1265 (14.34)	-	-	-	-	-
Diabetes	-	2899 (18.77)	-	1410 (15.98)	-	-	-	-	-
Assisted conception	-	2949 (19.09)	-	595 (6.74)	5 (0.42)	39 (2.08)	-	-	-
Birthweight	4 (0.38)	10312 (66.77)	25 (1.47)	82 (0.93)	36 (3)	6 (0.32)	223 (0.05)	38 (4.62)	26 (0.62)

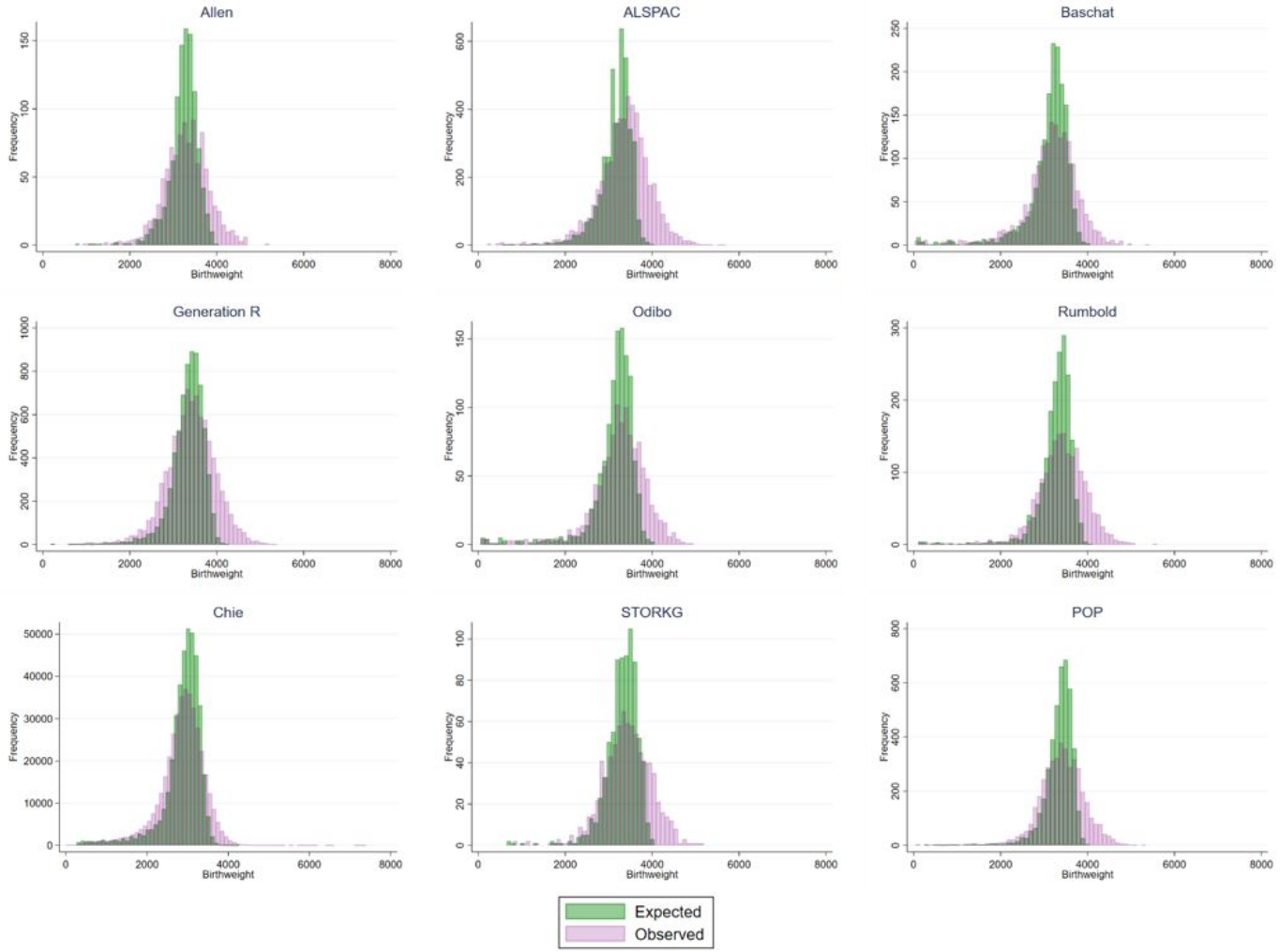
5.3.7 Predicted birthweight distribution

Predicted birthweight on the grams scale was slightly skewed on visual inspection of histograms (Figure 5.4), with a long left tail. This left tail was also seen in the distribution of observed birthweights for all cohorts, with notably few babies been born at weights below 2000g. The Chie cohort was the only setting where a larger number of both lower predicted and lower observed birthweights (below 2000g) were seen.

In all cohorts the distribution of predicted birtweights was narrower than that seen in the observed birthweights. Though the lower tail in observed birthweight distributions was well matched by the prediction distributions, higher birthweights were less well represented. While the left tail of the true distribution was well modelled by the Poon 2011 model, the more extreme right observations were poorly identified, with very few predicted birthweights exceeding 4000g (4kg) in any of the cohorts. Though very few observed birthweights exceeded 5000g, observed weights between 4000g and 5000g were common, and were not reflected in the distribution of predictions.

Distributions for both observed and expected birtweights were reasonably consistent across all cohorts. The only differences were noted in the Chie cohort, although these were minor, where more data was observed in the extremes. The largest observed babies were seen within the Chie cohort, where 27 babies were born larger than 5kg, potentially a reflection of the larger size of this dataset, allowing rarer more extreme observations to occur. Overall, the Poon 2011 predictions reasonably well mimicked the distribution of the observed outcome, where the majority of babies were born at a larger, healthier weight, with gradually fewer small babies, reflecting those born at unusually early gestational ages or with some level of growth restriction.

Figure 5.4: Distributions of Expected (based on the Poon model predictions) and Observed birthweights (g), by external validation cohort



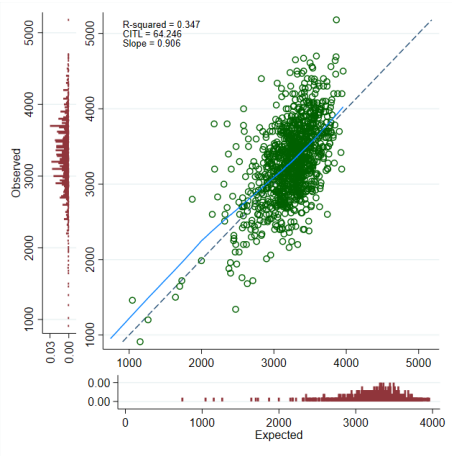
5.3.8 Model calibration performance

Given the Poon 2011 model was developed to predict values of $\log_{10} birthweight$, it was expected that calibration performance would be optimal for predictions on this scale, though given birthweight predictions on the grams scale were more easily clinically interpreted, and were more likely to be used in practice, reasonable calibration on the grams scale was important. Thus, calibration plots were produced for predictions on the more clinically interpretable grams scale, as follows. Assessment of calibration performance when using predictions on the $\log_{10} grams$ scale is given in Appendix IVa.

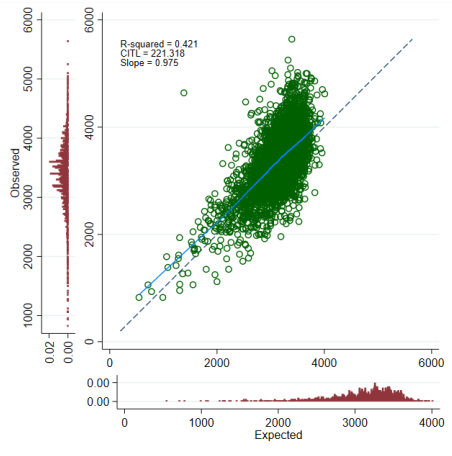
Calibration curves in Figure 5.5 suggest very good calibration on average for all cohorts. The individual-level scatters, however, show a wide spread of observed birthweight values for those with a given predicted birthweight, especially at higher predicted birthweight values. This spread is less pronounced on the $\log_{10} grams$ scale due to the shape of the log transformation. The most notable variation is seen in the largest cohort (Chie), where, for example, those with an expected birthweight of 3500g had observed values ranging from 500g up to over 7000g.

When focusing on the range of lower predicted birthweights, those at higher concern for FGR, calibration was good on average, and generally showed lower variability in observed values for a given predicted value. Given the clinical intention of identifying low birthweight babies at risk of FGR for early intervention, good calibration on average and small variation in predicted birthweights in the lower ranges is promising.

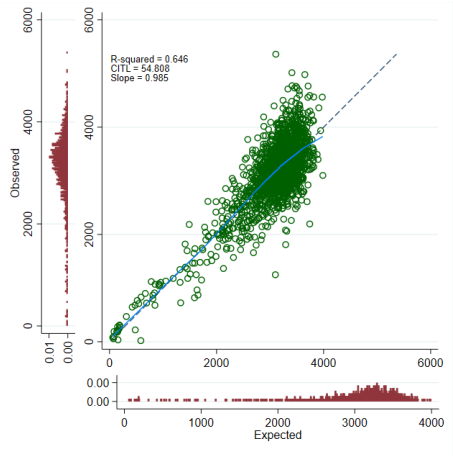
Figure 5.5: Calibration plots for the Poon 2011 model when assessed on the grams scale, by external validation cohorts.



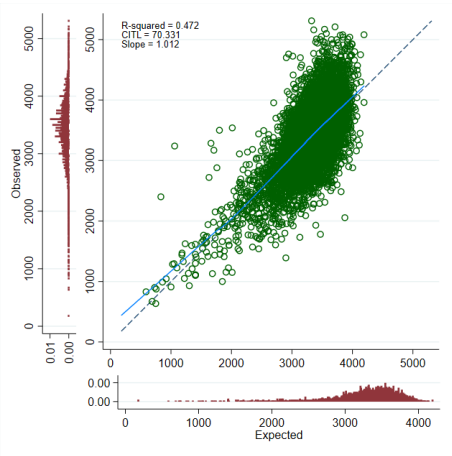
(a) Allen



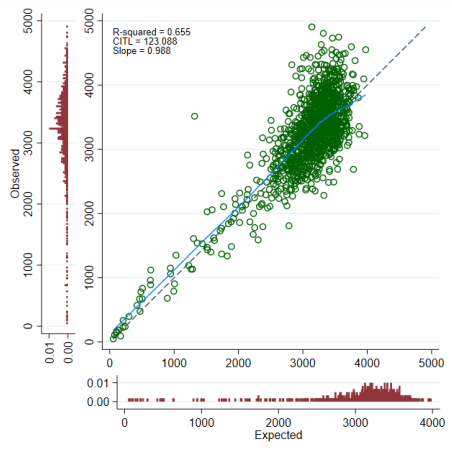
(b) ALSPAC



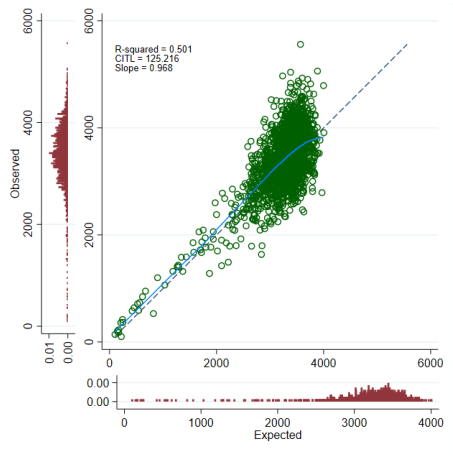
(c) Baschat



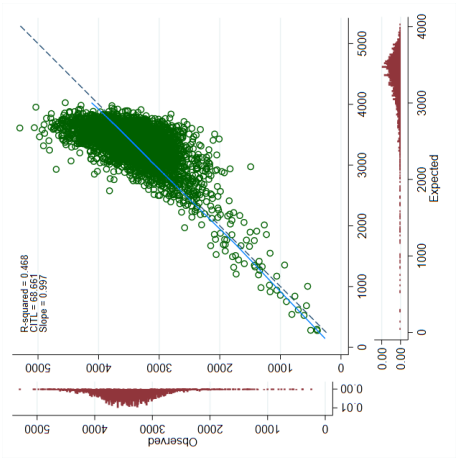
(d) Generation R



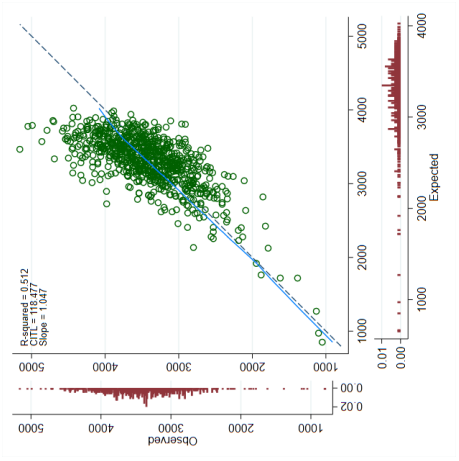
(e) Odibo



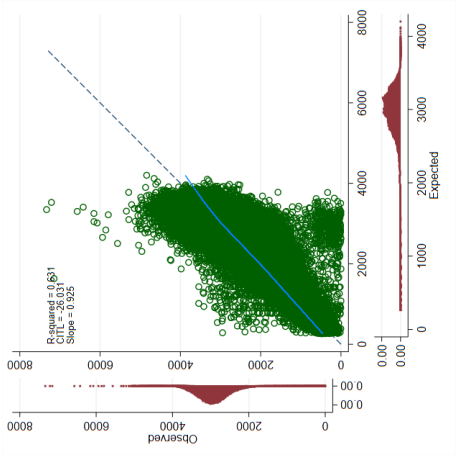
(f) Rumbold



(i) POP



(h) STORKG



(g) Chie

Calibration slope

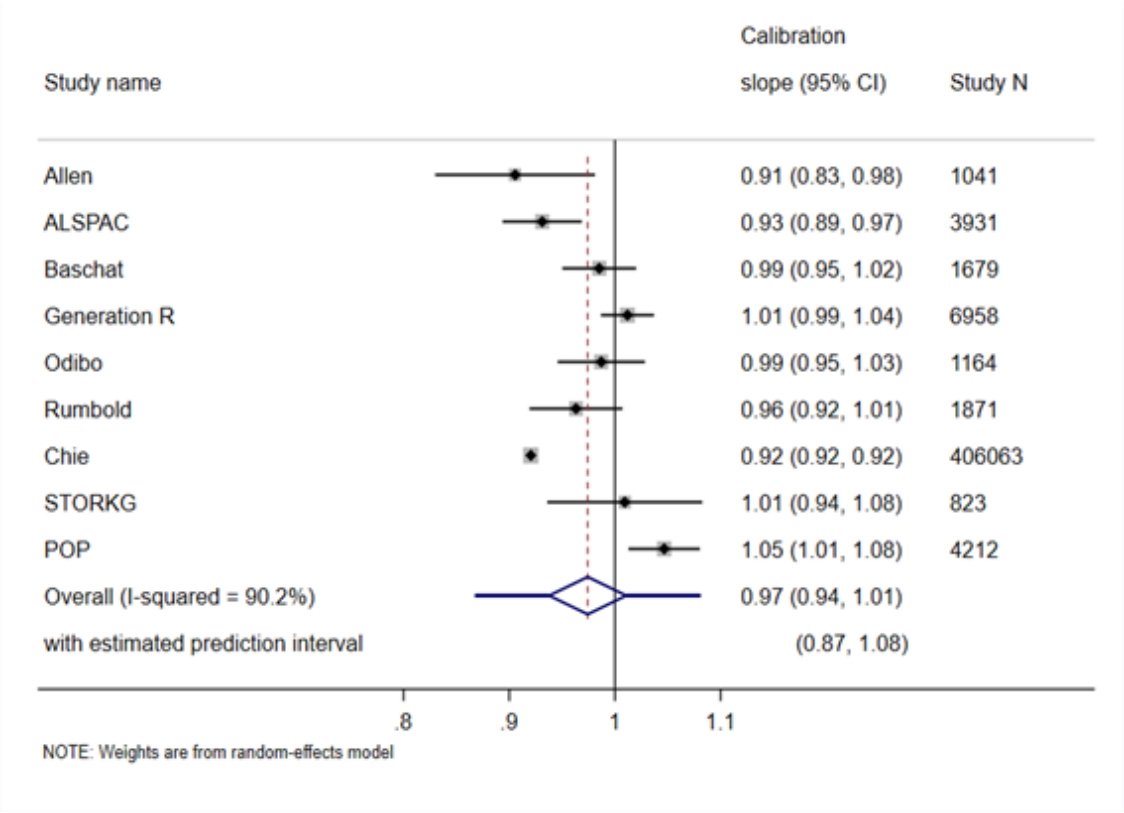
The pooled calibration slope across all cohorts was 0.97 (95% CI: 0.94 to 1.01, $\tau^2 = 0.0018$), implying near ideal calibration performance on average across settings, when using the model to predict birthweight in grams. Some heterogeneity was evident in calibration slope estimates across cohorts, as would be expected given case-mix differences in the samples included. Point estimates varied from as low as 0.91 (Allen) to 1.05 (POP), suggesting the range of predicted birthweights was too wide in the former and too narrow in the latter, relative to the range of observed birthweights. The 95% prediction interval for the calibration slope implied that the value expected in a new cohort from a similar setting has a 95% chance of falling between 0.87 and 1.08.

All cohorts comfortably exceeded the 386 participants required to achieve the recommended maximum confidence interval of width around the calibration slope (≤ 0.2), as shown in Section 5.3.4. Figure 5.6 shows how all cohorts give narrow confidence intervals around the calibration slope estimate, meeting the desired level of precision. This is also true of those cohorts with estimated calibration slope values above one (Generation R, STORKG, POP), though the sample size calculation conducted in Section 5.3.4 assumed a calibration slope on external validation of $\hat{\lambda}^2_{cal} \leq 1$. Chapter 4 demonstrated how the same size required to achieve precision in the calibration slope estimate would be higher where the value of $\hat{\lambda}^2_{cal}$ exceeded one: in this case, the high sample sizes available in all cohorts were sufficient to accommodate this difference from the assumed value.

The confidence interval around the pooled calibration slope was also narrow, with a width of only 0.05, despite some heterogeneity between estimates from the different cohorts. This implies the calibration slope of the Poon 2011 model was sufficiently consistent across populations to maintain

high precision in the pooled estimate.

Figure 5.6: Forest plot for the calibration slope of the Poon 2011 birthweight prediction model, across all external cohorts



Calibration-in-the-large

On average across cohorts, CITL was 90.39g (37.86g to 142.92g, $\tau^2 = 4578g^2$) on the grams scale, implying a systematic under-estimation of birthweight by around 90g on average, across all cohort populations (see Figure 5.7). The only cohort in which the Poon 2011 model over-predicted birthweights on average was the Chie cohort, where a CITL of -26.4g (-27.5g to -25.3g) suggests systematic over-prediction of just 26.4g: a clinically insignificant amount.

The largest absolute CITL value suggested an under-prediction of birthweight by 220.3g (206.5g

to 234.0g) on average in the ALSPAC cohort. While this amount is relatively small for a full-term, healthy weight baby, it was noted among the clinical team that 220.3g would be considered a clinically significant amount for small babies with suspected FGR, thus this level of error was concerning when considering application of the model in practice.

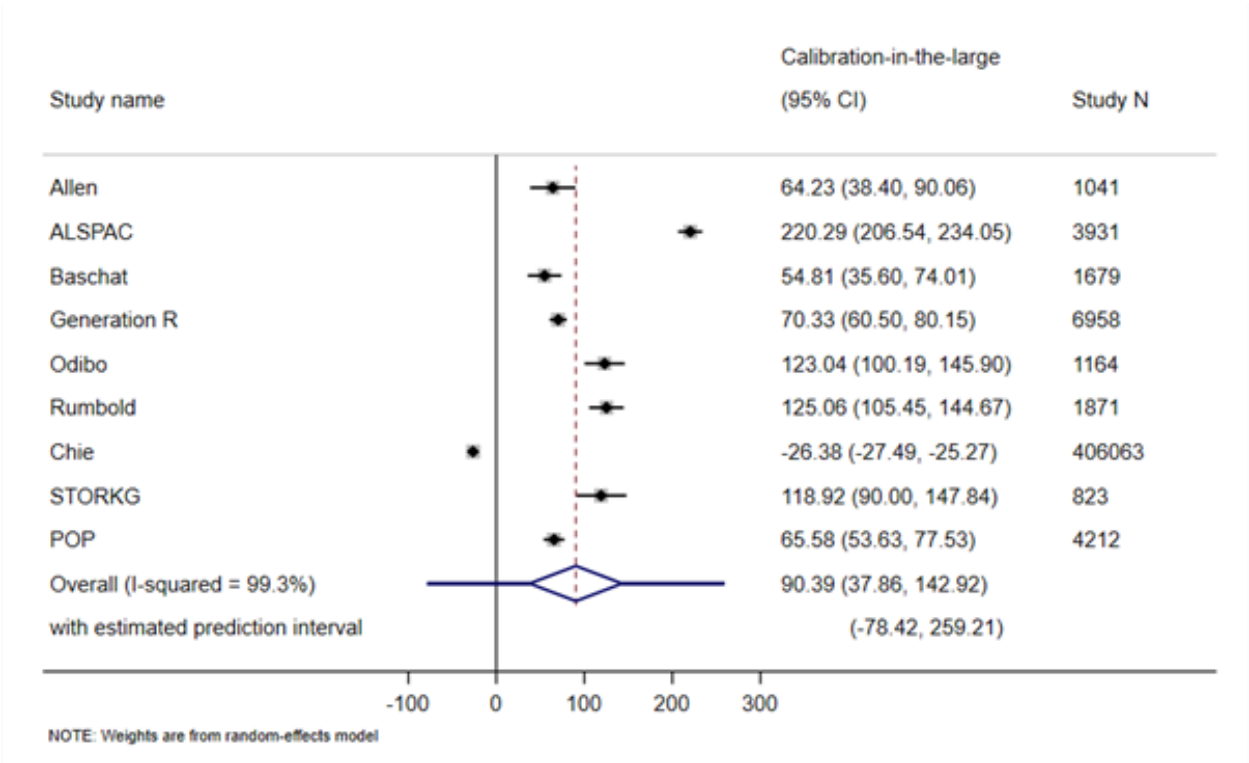
The sample size available in all cohorts was sufficient to meet the minimum requirement of 272 participants to ensure confidence interval widths of at most 100g. In fact, confidence intervals in most cohorts were considerably narrower than this, with widths ranging from just 2g (Chie) to 58g (STORKG). The confidence interval for the pooled CITL estimate, however, was wide (with a width of 105g) and so narrowly missed the desired precision level.

As can be seen from the large value for τ^2 and the spread of points on the forest plots in Figure 5.7, there was considerable heterogeneity in CITL estimates across the nine cohorts, supporting the notion that underlying birthweight distributions in the different populations (and so correspondingly, CITL values) were different to one another. Incorporation of this large τ^2 value in the variance of the pooled CITL estimate, to accommodate the variation in CITL across studies, resulted in a larger variance value overall.

Further to this, as mentioned in Section 5.2.6, a Hartung-Knapp-Sidik-Jonkman approach was employed in the calculation of the confidence interval around the pooled CITL. Given the Hartung-Knapp-Sidik-Jonkman confidence interval is derived using critical values from the student's t distribution, which are larger than the corresponding values from the normal distribution used in the derivation of intervals for individual cohorts, the resulting confidence interval is expected to be wider than if using other methods of estimation.

Thus, in this case, the heterogeneity in model performance across cohorts, along with analysis methods to properly allow for this, resulted in lower precision in the pooled CITL result than in any individual cohort. The concept of precision in estimates alone does not directly extend to an IPD meta-analysis situation, where different studies contain different populations resulting in different levels of model performance.

Figure 5.7: Forest plot for the calibration-in-the-large of the Poon 2011 birthweight prediction model, across all external cohorts



Further details on the consistency of model calibration between analysis in the multiply imputed data and analysis in complete cases only are given in Appendix IVb.

5.3.9 Decision curve analysis

In order to assess the clinical utility of using the Poon's birthweight prediction model in the identification of SGA babies early in pregnancy, decision curves were produced from the number of true positive (those correctly identified as SGA) and false positive (those incorrectly identified as SGA) cases. The observed number and proportion of SGA babies in each of the external validation cohorts, and over all external validation data combined, are summarised in Table 5.6. The outcome prevalence varied across cohorts, between 3.3% (STORKG) and 16.5% (Chie), while the overall prevalence across all cohorts was heavily influenced by the considerably higher sample size and SGA prevalence seen in the Chie cohort.

Table 5.6: Numbers of observed SGA events, defined as being below the birthweight 10th percentile cut-off for observed gestational age at delivery, by cohort

Cohort	Total participants	Events (10th percentile)	%
Allen	1041	90	8.6%
ALSPAC	5132	305	5.9%
Baschat	1679	121	7.2%
Generation R	8742	489	5.6%
Odibo	1164	63	5.4%
Rumbold	1871	92	4.9%
Chie	403,284	66,347	16.5%
STORKG	823	27	3.3%
POP	4212	168	4.0%
Total	427,948	67,702	15.8%

Decision curves in Figures 5.8 and 5.9 display the expected benefit of using the Poon 2011 model to identify those at a high risk of FGR, net of any harm caused by incorrectly identified high risk

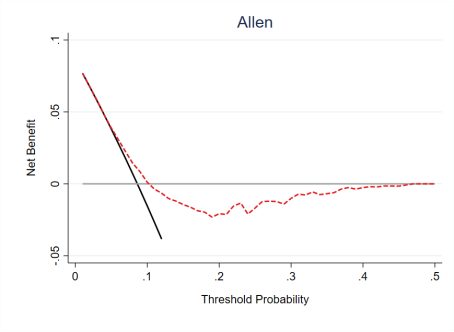
cases (where the baby was of healthy weight at birth), for each of the external validation cohorts. These figures have been generated using the methods described in Chapter 3, from the probability that the birthweight prediction from the Poon 2011 model would imply a predicted birthweight (a) in the lowest 10% birthweights at an assumed 40 weeks' gestation at delivery (Figure 5.8); and (b) in the lowest 10% birthweights for their true gestational age at delivery, which would be unknown if the model were to be used in practice (Figure 5.9). These probabilities were compared to the observed, dichotomised birthweight outcome: observed birthweight in the lowest 10% birthweights for their observed gestational age at delivery, based on the cut-offs given in Table 5.1.

The below curves show the net benefit of treatment based on model predictions compared to that of simple treat-all and treat-none strategies. Eight of the nine studies show the net benefit of the Poon 2011 model closely following the treat-all strategy up to the point where the treat-all curve crosses the x-axis. This suggests that, in these cohorts, using the model to guide SGA treatment based on treatment decision thresholds up to this point has no benefit over treating everyone as if they were at high risk. For decision cut-off values beyond the point where both the treat-all and Poon 2011 model curves cross the x-axis (where the threshold probability corresponds to the outcome prevalence for each cohort, in the treat-all case), the best strategy is to treat-none despite having a net benefit of zero, as both treat-all and treat-per-model options result in net harm overall.

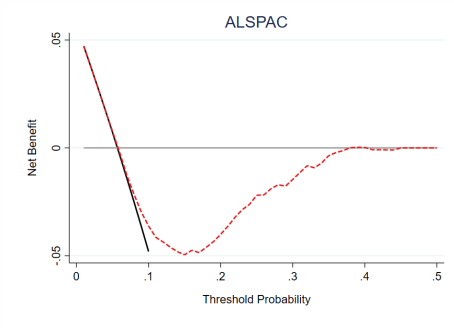
The only cohort for which a net benefit seems to be indicated is Chie: the cohort with both the largest sample size, and the largest proportion of events. In this case, probability thresholds between 15% and 25% appear to have some benefit over other strategies, though this range is clearly data dependent and was not defined *a priori* as being of clinical relevance.

These results seem to be consistent across both definitions for defining outcome probabilities, though interestingly using the observed gestational age to generate predicted values lead to decision curves that indicated net benefit slightly below that of the treat-all strategy for lower threshold probabilities.

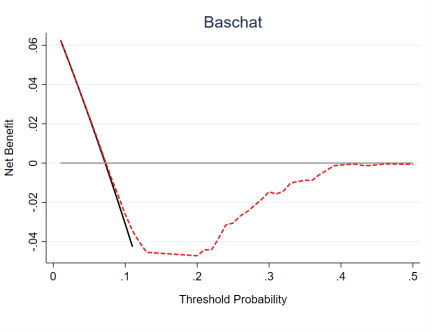
Figure 5.8: Decision curve analysis by validation data set, based on a cut-off for predicted birthweight at a standardised predicted gestational age at delivery of 40 weeks. Red lines indicate the net benefit of using the Poon 2011 model, black and greys lines indicate the treat-all and treat-none alternatives, respectively.



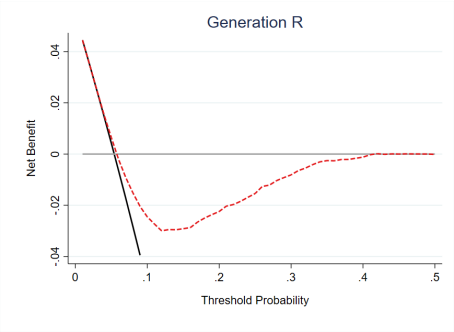
(a) Allen



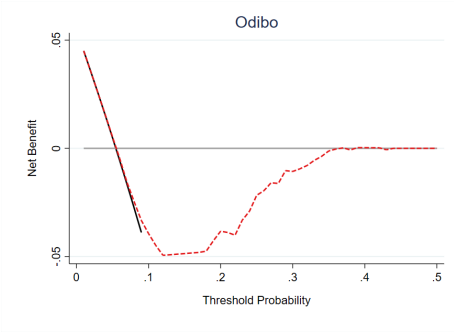
(b) ALSPAC



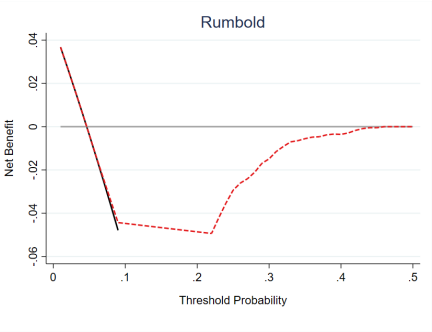
(c) Baschat



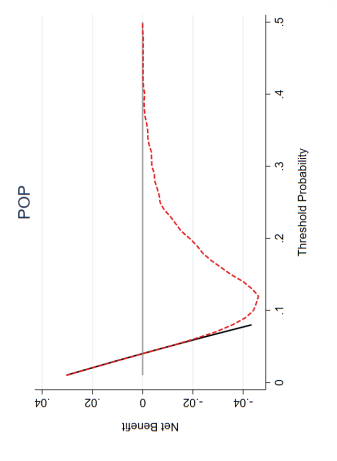
(d) Generation R



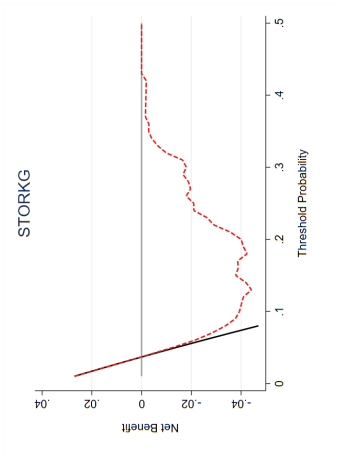
(e) Odibo



(f) Rumbold



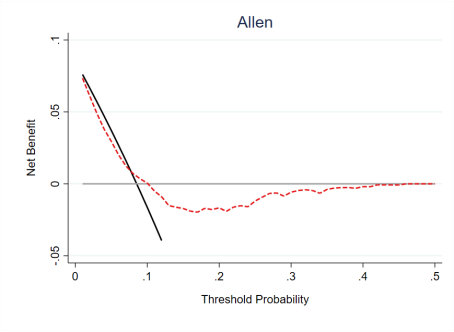
(g) Chie



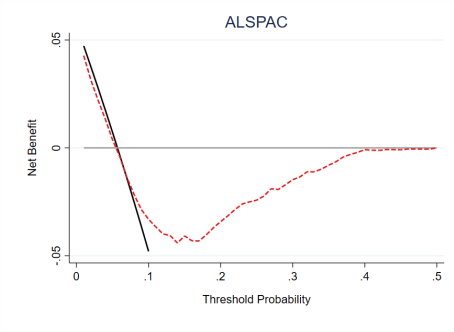
(h) STORKG

(i) POP

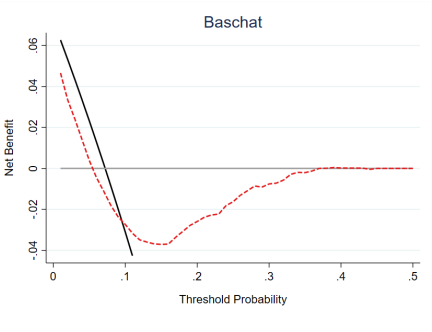
Figure 5.9: Decision curve analysis by validation data set, based on a cut-off to include those with predicted birthweight $\leq$ 10th percentile for their observed GA at delivery (weeks). Red lines indicate the net benefit of using the Poon 2011 model, black and greys lines indicate the treat-all and treat-none alternatives, respectively.



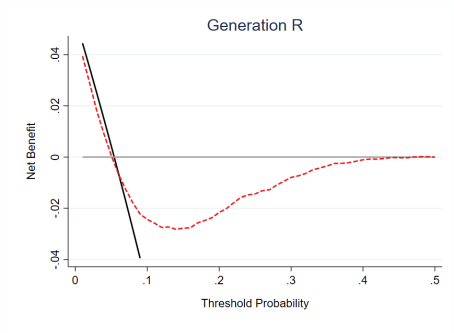
(a) Allen



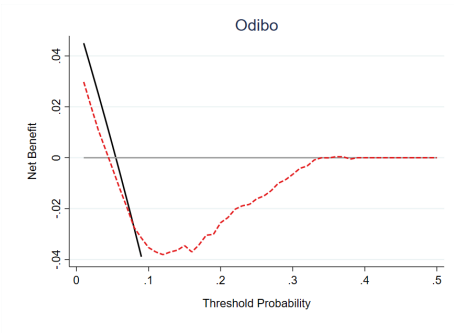
(b) ALSPAC



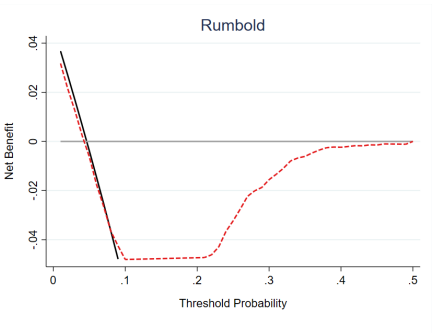
(c) Baschat



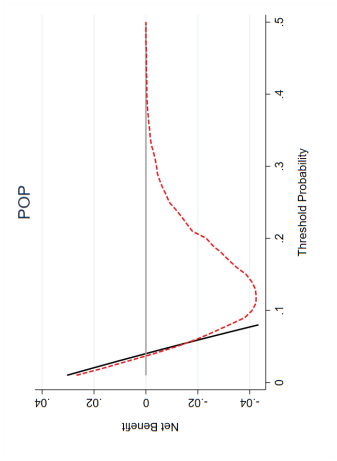
(d) Generation R



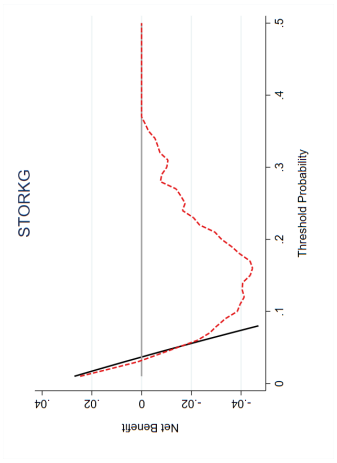
(e) Odibo



(f) Rumbold



(g) Chie



(h) STORKG

(i) POP

5.4 Discussion

5.4.1 Summary of key findings from this chapter

This chapter demonstrates how external validation of a published prediction model in multiple individual datasets can be used to assess both model generalisability and transportability. From the 119 models identified in Chapter 2, no model predicted FGR by a definition that was compatible with that specified by clinical collaborators (birthweight <10th centile adjusted for gestational age, with stillbirth or neonatal death or delivery before 32 weeks). Of the 11 models predicting birthweight on a continuous scale, eight included variables that were not recorded in the IPPIC collaboration data collection, and two contained variables that were not available in combination within a given cohort. Just one model, for birthweight on its continuous scale could be externally validated.

External validation was possible in nine cohorts from across six countries, with recruitment periods ranging from 1991 to 2014. Data contained information on 441,415 pregnancies, though the bulk of these came from a single, Japan-based cohort (Chie). Birthweight distributions were similar across cohorts, as were prediction distributions, though predictions poorly matched the higher range of the observed birthweight distribution. Overall, predictions well mimicked the distribution of observed birthweights in the lower range of weights, with the majority of babies born at a larger, healthier weight and with relatively few small babies.

All cohorts individually met the minimum sample size requirements for a single external validation of the Poon 2011 model, per criteria introduced in Chapter 4. Increased numbers of participants from combining information across cohorts, however, did not result in corresponding increases in

precision (defined by a narrower confidence interval width around pooled estimates). Heterogeneity stemming from between study variances lead to lower precision in the pooled CITL result than for the CITL in any individual cohort.

Estimates of the calibration slope were fairly consistent across cohorts, with the resulting pooled calibration slope value having a narrow confidence interval width of only 0.05 (within the precision targetted during the sample size calculation). Despite some heterogeneity between calibration slope estimates across cohorts, this implies the calibration slope of the Poon 2011 model was sufficiently consistent across populations to maintain high precision in the pooled estimate.

While the Poon 2011 model was well calibrated on average, large amounts of individual-level variation was not well accounted for by the model predictions. Individual-level scatters of expected to observed birthweight showed a wide spread of observed values for those with a given predicted birthweight, especially at higher predicted birthweight values. This variation was also present, though less pronounced, on the $\log_{10} grams$ scale on which predictions were generated, suggesting room for improvement in prediction of birthweight, despite promising calibration performance through pooled values.

Beyond assessments of calibration, this chapter further demonstrated an additional application of methods for gaining predicted probabilities from a linear regression model, described in Chapter 3, through an example of how this proposed method could be combined with traditional decision curve analysis to explore the net benefit of using a birthweight prediction model in the management of patients at high risk of FGR. In this case, decision curves suggested that using the model to guide treatment decision had no benefit over alternative strategies (treating all or treating none,

depending on threshold probability) for most cohorts. A clear net benefit over other strategies was only seen in one cohort, Chie, which was both the largest, and had the highest outcome proportion.

5.4.2 Strengths and limitations

The applied example in this chapter enabled further consideration of a potential drawback of the sample size approach suggested in Chapter 4. While on average CITL was considered small, larger absolute CITL values such as that seen in the ALSPAC cohort (220.3g, 95% confidence interval: 206.5g to 234.0g) were noted among the clinical team as being a clinically significant amount for small babies with suspected FGR. This level of error would be concerning when applying the model in truly small babies, those born at unusually early gestational ages or with some level of growth restriction, though would be considered a negligible amount in the assessment of full-term, healthy weight baby.

The acceptable level of precision in CITL included in the sample size calculation in Section 5.3.4 was $\pm$50g. Given the difference in acceptable error in predictions, it is feasible that the desired confidence interval width could vary considerably depending on the population under assessment. For those with an earlier gestational age at delivery, with relatively low birthweights, narrow confidence intervals would be desired, meaning the sample size calculated above may not be sufficient.

Thus, prior to sample size calculation the population composition must be considered carefully, to allow for varying precision requirements during an external validation. Targetting different

levels of precision for different subgroups, for example those defined by different gestational ages, might be preferable. This would mean conducting multiple sample size assessments, with appropriately and independently targetted precision in key subgroups, or establishing a precision in performance statistics that would be acceptable in all subgroups simultaneously, rather than just being acceptable for the majority.

Despite the large number of models identified in Chapter 2, the external validations conducted within this chapter were confined to a single published model. All published models identified to predict FGR failed to meet a suitable outcome definition, and of those predicting birthweight, suitable predictors were not available in combination in the available data.

A key challenge in the implementation of IPD-MA in practice, as clearly demonstrated here, is the difficulty in dealing with different definitions or measurements of necessary predictors and outcomes across different data sources [67]. In particular, completely (systematically) missing predictors in some cohorts limited the number of models that could be validated in this study, despite the wealth of data available and the huge effort that went into curating the IPPIC collaboration data collection.

While multiple imputation approaches were conducted to account for missing predictor data within a cohort, no attempt was made to account for predictor information that was entirely missing in some cohorts, though recorded in others. Methods to account for systematically missing information may have enabled the external validation of a further two models for predicting continuous birthweight, though were not considered in this example [209, 210].

IPD-MA for prediction model validation has great potential, allowing evaluation of predictive

performance and clinical utility across a range of patient subgroups, different populations, or even different care settings, if such datasets are included [67]. Setting-specific performances over a more variable case-mix allows more robust conclusions on the model's generalisability across different settings.

The external validation discussed in this chapter, along with subsequent research involving model development and internal-external cross-validation (summarised in this chapter's Appendices), have been detailed in an HTA report "*External validation and development of prediction models for fetal growth restriction (FGR) and birthweight: an Individual Participant Data (IPD) meta-analysis and cost-effectiveness analysis*", which is due for publication in 2024.

5.4.3 Conclusion and next steps

Through application in an example model for predicting birthweight, this chapter has demonstrated how combining performance estimates from external validations across differing populations may lead to wide confidence intervals, even where all cohorts meet precision recommendations for a single validation study. Heterogeneity in model performance across external validations may arise where populations differ, for example when coming from different geographical regions or having different case-mixes. When combining performance estimates across heterogeneous groups, targeting narrow confidence intervals for pooled estimates may not be suitable. Appropriate random-effects meta-analysis methods to allow for heterogeneity in model performance will rightly result in wider confidence intervals around pooled estimates. Thus, precision in pooled estimates depends on the consistency of performance across settings, not solely the sample size used.

The concept of precision in estimates alone does not directly extend to an IPD meta-analysis situation, where different population compositions can lead to different predictive performance. Assessment of sample size needed for an external validation involving meta-analysis across settings should allow for this, and precision in the pooled value should be interpreted differently to the equivalent in a single study. Wider confidence intervals may not be rectified when pooling across populations, thus narrow confidence intervals around a pooled estimate should not be the target of sample size calculations where heterogeneity is expected and random-effects meta-analysis methods are employed. Similarly, a sufficient sample size across studies combined should not be assumed to be sufficient where individual studies are smaller than the recommended minimum.

Chapter 6 will further expand on this theme of external validation across multiple populations and settings, with the IPD meta-analysis methods introduced in this chapter further utilised to demonstrate variability in model performance at a national (UK) level. The performance of a model to predict the risk of a serious fall (resulting in hospitalisation or death) is investigated in routinely collected data across many different GP practices, allowing assessment of the extent of heterogeneity in model performance across different settings, and identification of where precise overall estimates of model performance (given a large sample size for external validation) may have masked poor performance in smaller geographical subgroups with differing case-mix.

So far, this thesis has presented examples with an outcome of interest that has been measured on a continuous scale, with outcome dichotomisation being considered to model the data instead as a binary outcome. Chapter 6 now introduces an example of a prediction model for a genuinely binary outcome event (a serious fall), with the (continuous) time until this event occurred being of primary interest. In particular, the continuous aspect of this time-to-event outcome is investigated

through the use of pseudo-values for observed outcomes, to account for both the censoring of data and the competing risk of death, to measure the model calibration on a continuous scale. These concepts are explained further in the upcoming introduction section.

CHAPTER 6

Chapter 6: Methods for external validation of survival models in big data whilst accounting for competing risks: examining calibration on a continuous scale using pseudo-values

6 Chapter 6: Methods for external validation of survival models in big data whilst accounting for competing risks: examining calibration on a continuous scale using pseudo-values

6.1 Introduction and objectives

Thus far, this thesis has presented examples where an outcome dichotomisation was conducted with a view to modelling data simply as a binary outcome, without the time until this binary outcome occurred being of interest. Survival analysis combines a binary outcome (such as the high pain intensity and FGR examples from previous chapters) with the continuous aspect of the time until that event takes place. Many clinical prediction models are developed using a survival analysis approach, and these also require thorough external validation.

The suitability of survival models for application outside the model development dataset is assessed through measures of calibration, discrimination, and net benefit, just as with continuous and binary outcome model. This chapter continues with the theme of external validation, in this case of a survival model, through the use of pseudo-value estimates for observed outcomes, giving a representation of the individual's contribution to the survival function. These pseudo-values are derived from the cumulative incidence function for the outcome of interest, incorporating both right censoring (described in Chapter 1) and competing risks (introduced in Chapter 1 and discussed in further detail in Section 6.2.1, below) in the assessment of the observed outcome, and are measured on a continuous scale. Further details on the use of pseudo-values in the assessment of calibration for a survival model are given in Section 6.2.2 of this chapter.

The analyses shown in Chapter 5, demonstrate the potential of IPD meta-analysis techniques

in the investigation of heterogeneity in model performance across populatoins and setting. This chapter expands on these concepts, with IPD meta-analysis being employed on a much larger scale, in the assessment of model performance in big data from Electronic Health Records (EHR), comprising patients from numerous sub-populations within different GP practices across England. Summarising performance measures across these different GP practices, especially when considering those of smaller sizes, may reveal substantial variation in estimates across sub-populations. Such variability is a key consideration prior to the application of a prediction model in practice, as promising performance on average may obscure poor performance in subgroups, where, for example, model miscalibration could cause an overall harm to patients.

Thus, this chapter delves deeper into the importance of sample size for external validation, as discussed in Chapters 4 and 5, with overall sample size across the validation population not being the only concern. Given populations within national settings vary considerably, assessment of model performance in patient subgroups, such as those defined geographically through GP practice registration, could potentially reveal unacceptable levels of uncertainty in model performance within (or variation across) subgroups.

First, the motivating clinical example is introduced, including details of the clinical prediction model to be validated and the wider research project that this chapter contributes to. Next, a summary of further methods for survival analysis are introduced, building on the concepts described in Chapter 1, including discussion of methods to account for competing risks and the use of pseudo-values to estimate survival from incomplete outcome measurements (censoring). Part (i) then fully demonstrates methods to calculate the minimum sample size required for the external validation of this survival model, while part (ii) describes the results of the validation itself.

6.1.1 Clinical scenario

The motivating example behind this chapter stems from a programme of work surrounding the prediction of adverse events associated with antihypertensive medication in frail older adults. Medications to reduce blood pressure (antihypertensives) are widely prescribed in older adults [226] and are generally effective at reducing the risk of cardiovascular disease in the future [227]. These medications are also thought to be associated with many different adverse events, including acute kidney injury (AKI), electrolyte abnormalities, and syncope [13]. In particular, previous studies have suggested a possible association between antihypertensive treatment and falling in frail individuals [13, 228, 229].

Improved assessment of the risk of different adverse events could allow targeted treatment in those with an indication for antihypertensive medication, who are least likely to experience harm from associated adverse events. The STRAtifying Treatments In the multi-morbid Frail elderly (STRATIFY) project aimed to use routinely collected data from UK-based EHR to develop and externally validate clinical prediction models to estimate an individual's risk of experiencing hospitalisation or death due to different adverse events within 10 years of antihypertensive treatment being indicated. Thus, the clinical example discussed here formed a part of a larger body of work, developing and external validating models to predict the risk of serious falls (the focus of this chapter) [97], AKI [98], fractures, hypotension, syncope, hyperkalaemia, and hyponatraemia. The protocol for this study was assessed and approved by the Independent Scientific Advisory Committee (ISAC) (protocol number 19_042).

The clinical prediction model of interest in this chapter is for the prediction of hospitalisation or death associated with a fall, defined based on ICD-10 (International Classification of Diseases, 10th revision) codes in Hospital Episode Statistics (HES) and Office for National Statistics (ONS) mortality data. The population of interest were those aged 40 years or older, with health records available during the study period, from the time of their first systolic blood pressure reading above (or including) 130mmHg [230]. Patients with any systolic blood pressure reading greater than 180 mmHg were excluded, as antihypertensive treatment would be indicated for these patients regardless of the risk of adverse events. These eligibility criteria were applied both in the external validation data, reported in this chapter, and in the model development cohort.

Given the population of interest in this clinical case were older adults at a higher risk of cardiovascular disease, there was a notable chance that some participants may die from causes unrelated to the outcome of interest, before that outcome of interest could take place, with such death being considered a competing risk to the main outcome of serious falls risk.

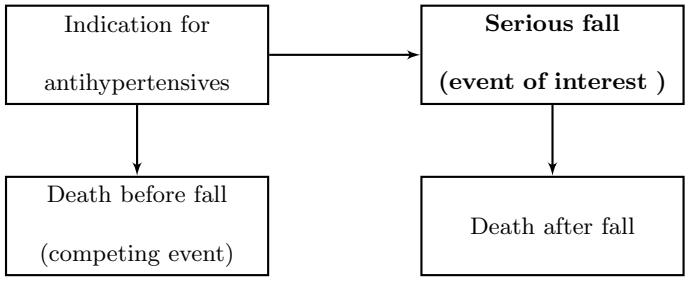


Figure 6.1: Demonstration of death as a competing risk, where the transition from indication for antihypertensives to serious fall is of primary interest

Existing model to predict the risk of a serious fall

The model to be externally validated was developed in a previous stage of the same project, after identifying a lack of suitable models for the desired population in the literature. The model to predict serious falls was developed using EHR data from the Clinical Practice Research Datalink (CPRD), from GP practices across the UK. Specifically, this model was developed using linked data from CPRD GOLD, which included data from GP practices using *Vision* electronic health record software (Cegedim Healthcare Solutions, London, England). Outcomes were determined through linkage to ONS mortality data and HES. All development analyses were conducted by members of the STRATIFY research team who were based at University of Oxford, independent to the external validation analyses.

Clinically relevant predictors of falls were identified from the literature [13] and through consultation with clinical experts. Prior to variable selection, 30 predictor variables were considered (44 predictor parameters, including fractional polynomial transformations of continuous predictors), covering demographics, clinical characteristics, comorbidities, and prescribed medications. A Fine-Gray sub-distribution hazard model was fitted to estimate falls risk over time, taking into account the competing risk of death by other causes [231]. The resulting apparent calibration plots in the model development data showed significant miscalibration, with under-prediction of falls risk for patients with low predicted risks and substantial over-prediction for those with high predicted risks. This original model was therefore recalibrated to the observed pseudo-values for the CIF of a serious fall, which improved apparent calibration (in the model development data) considerably. Both the Fine-Gray model and the pseudo-value recalibrated version ("STRATIFY-Falls model") were assessed as a part of the external validation analyses.

The primary outcome for prediction was falls risk ten years following a patient's first systolic blood pressure reading above (or including) 130mmHg. Pre-defined secondary outcomes of one- and five-year risks were also of interest, though are not discussed further in this chapter.

Available datasets for external validation

Identification of participants for the external validation followed a retrospective cohort design. The validation cohort came from CPRD Aurum, which contained IPD from GP practices in England that used recording software from Egton Medical Information Systems (*EMIS*, Leeds, England). Patients were eligible if they were registered at a linked practice (with outcome information available from HES or ONS data) between 01/01/1998 and 31/12/2018. The data contained within CPRD Aurum had previously been shown to be representative of the England patient population, in terms of age, ethnicity, and deprivation status, and so was considered suitable for assessing the model's generalisability in the UK population [232].

6.1.2 Objectives

This chapter describes the evaluation of a clinical prediction model for the risk of hospitalisation or death from a serious fall in those with an indication for antihypertensive treatment. The external validation cohort comprised information on patients from 738 different GP practices, routinely collected and stored in EHR.

In particular, the objectives of this chapter were to:

1. Demonstrate methods to calculate the appropriate sample size needed to accurately assess

the predictive performance of a survival model for serious falls in a single validation study.

2. Illustrate the use of pseudo-values in the assessment of calibration performance in the external validation of prognostic models with a time-to-event outcome.

3. Demonstrate IPD meta-analysis methods to summarise model performance across populations, while assessing heterogeneity in model performance across GP practices.

The sample size calculation demonstrated in the first part of this chapter, for the external validation of a clinical prediction model with a survival outcome, was developed as a part of a wider research team, and was published in *Statistics in Medicine* [93], following the publication of the calculation methods proposed in Chapter 4 [92], and other associated work [95, 94]. The clinical application discussed here was published in a journal article in the *BMJ* [97], and led to related publications into the prediction of falls risk in a wider population [99], and into further exploration of adverse events associated with antihypertensive medication [98, 13], as a part of a wider research group.

6.2 Further methods for survival analysis

6.2.1 Competing risks in prediction modelling

Given the length of follow up in many survival analysis prediction models, follow up for an individual can often end in one of many possible ways. Whilst in many cases, those without the event of interest during their follow up can be treated as right censored observations, in other cases an alternative (competing) event may have occurred that prevents the event of interest from taking place [233]. To treat such individuals as censored observations involves the unrealistic assumption that they are still at risk of the event of interest for the remainder of the time frame of interest, when this is not actually possible [234].

Standard survival analysis methods, without accounting for competing risks, are likely adequate when the competing risk is rare, but in older or frail populations where a competing event such as death is common, alternative approaches must be considered [235]. Where an event such as death precludes the event of interest, treating those who have died as though they were censored (for example, by using the Kaplan-Meier estimate of the survival function [28]) can result in overestimating the risk of that outcome in the remaining study population [236, 237]. Predictions from these methods in the presence of competing risks are said to refer to the risk of the event of interest in a hypothetical world where the competing event is not possible [234]. To inform clinical decision making in a real-world setting, where the competing event can actually happen, these hypothetical risks are not always relevant [238].

What constitutes appropriate analysis of those who have experienced the competing event will depend on the research question under investigation [236]. Two common approaches for accounting

for competing risks in the analysis of survival data are described below.

Cause-specific hazards models

The cause-specific hazard function $h_c(t)$, so named because it refers to the hazard function specifically for follow up ending due to cause c, gives the probability that event c occurs in the time interval immediately following time t given the subject has not experienced event c up until that time.

$$h_c(t) = \lim_{\delta \to 0} \frac{P(t \leq T_c < t + \delta | t \leq T_c)}{\delta}$$

where T_c denotes the time to the specific event type c.

When multiple competing events are present, survival analysis methods described in Chapter 1 can be used to separately estimate the probability of each type of event, c, while treating all other possible events as if they were censored [239]. These probabilities over time are known as the cause-specific hazard functions, for example based on a Cox proportional hazards model [29], giving functions of the form:

$$h_c(t) = h_{0c}(t) exp(\beta_{\mathbf{c}} \mathbf{X})$$

where h_{0c} gives the baseline hazard for event c, and β_c is the vector of cause-specific coefficients, allowing for the effects of covariates to differ between different event types.

This process leads to multiple cause-specific models, one for each of the competing events and one for the main event of interest. To subsequently make predictions for the main event accounting for the competing events involves a non-trivial combination of the hazards from all of these cause-specific models. [234, 237].

One potential drawback of implementing this approach is that the Cox model assumes independent censoring not only for those who were truly censored, but also for those who experience a competing event [29, 30]. This is the equivalent of assuming independence between the risk of different event types (an assumption which cannot be tested in practice, and is likely unrealistic in many cases), though evidence suggests that a cause-specific approach is valid, whether or not independent censoring is the case [239, 237]. Cause-specific approaches to modelling survival data in the presence of competing risks are generally preferred when the research question surrounds the effect of specific prognostic factors on different outcomes [236].

Sub-distribution hazards model

An alternative involves modelling the Cumulative Incidence Function (CIF) using a sub-distribution hazards approach, which is derived from the cause-specific hazards of each event, but does not assume their independence [240]. In the absence of any other censoring, for example through loss to follow-up, the approach is equivalent to including individuals who experienced the competing event in the risk set until the end of the study follow-up. If right censoring, then the contribution of these individuals is down-weighted over time, accounting for the probability of having been censored.

The CIF gives the marginal probability of each event type occurring at a specific time t: the probability of the event occurring regardless of whether the individual was censored or experienced a competing event. For a given event type, c, the CIF_c at time t can be expressed in terms of the

cause-specific hazard ($h_c(t)$) and overall survival ($S(t)$) functions [239], as

$$CIF_c(t) = P(T_c \leq t) = \int_0^t S(u)h_c(u)du$$

where T_c denotes the observed time that event c occurred. When no competing risks are present, the function $CIF_c(t)$ is equivalent to $1 - KM(t)$, where KM is the Kaplan-Meier estimator of the survival function [233, 239].

In 1999, Fine and Gray proposed a model akin to the Cox proportional hazards model, to predict the value of the CIF in the presence of other covariates [231]. This model treats the CIF as a sub-distribution, and so models the sub-distribution hazard function for the event of interest based on this CIF. This sub-distribution hazard function is defined as:

$$h_{c,CIF}(t) = \lim_{\delta \to 0} \frac{P(t \leq T_c < t + \delta | t \leq T_c \cup T_{c'} \leq t, c' \neq c)}{\delta}$$

The resulting model is of the form

$$h_{c,CIF}(t) = h_{0c,CIF}(t)exp(\beta \mathbf{X_i})$$

which estimates the sub-distribution hazard at a time t based on the value of the baseline CIF for event c at that time point, $h_{0c,CIF}(t)$, the vector β holding predictor effect estimates in the presence of the competing risk, and the vector of predictor values $\mathbf{X_i}$. These predictor effects are interpreted in much the same way as the coefficients from a Cox model, as the (adjusted) effects of the covariate on the cumulative incidence of the event of interest [234, 237].

There is still debate over what is the optimal approach to handling competing risks when developing and validating prediction models for survival outcomes, with evidence that, when using a Fine-Gray approach, the sum of the cause-specific estimates of risk (i.e., the combined CIFs

from all possible events obtained from sub-distribution hazard models) can exceed 100% for some patients [241]. In the development of the model being validated in this chapter, however, only the risk of a single event was of interest, rather than the probabilities or effect estimates for all event types, thus the adopted approach was that proposed by Fine and Gray [231], with the CIF of the event of interest (as defined above) being used [236].

6.2.2 Pseudo-values for observed survival estimates

A key feature of survival analysis is the incompleteness of outcome data due to censoring. Without censoring, the survival time T would be observed for all individuals, and standard regression methods could be used to model survival time directly, or to model a binary event indicator ($I(T \leq t)$) formed when dichotomising follow-up at some time t. Similarly, regression models could be used to assess any given function of the follow-up time, $f(T)$, if information on T were complete. One method proposed to allow such analyses in incomplete event history data is the use of *pseudo-values*: jack-knife estimators representing an individual's contribution to any function, $f(T)$, of interest [242].

For an unbiased estimator $\hat{\theta}$ of the expected value $E(f(T))$, the pseudo-value of estimate for individual i is defined as

$$\hat{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}^{-i}$$

This is a logical construction in the absence of censoring, where individual random variables can be reconstructed using the leave-i-out estimator $\hat{\theta}^{-i}$ of θ. For example, consider estimating the expected value, $E(f(T))$, of a function of event times across n individuals. In the absence of

censored observations, this is a simple average defined as $\hat{\theta} = \frac{1}{n}\sum_i f(T_i)$, thus

$$n\hat{\theta} = n\frac{1}{n}\sum_i f(T_i) = f(T_1) + f(T_2) + ... + f(T_{i-1}) + f(T_i) + f(T_{i+1}) + ... + f(T_n)$$

$$(n-1)\hat{\theta}^{-i} = (n-1)\frac{1}{n-1}\sum_{\rho \neq i} f(T_\rho) = f(T_1) + f(T_2) + ... + f(T_{i-1}) \qquad + f(T_{i+1}) + ... + f(T_n)$$

Evidently, here $n\hat{\theta} - (n-1)\hat{\theta}^{-i} = f(T_i), \forall i = 1, ..., n$ [243, 244].

When independent censoring is present, the function of interest is the survival function over time, thus $\hat{\theta} = \hat{S}(t) = E(1 - P(T \leq t))$, which can be gained using the Kaplan-Meier estimate [242, 28]. Extending the above concept, we construct individual-level estimates of $\hat{\theta}_i = \hat{S}_i(t)$ for a given value of t as $n\hat{S}(t) - (n-1)\hat{S}^{-i}(t)$ [243]. These $\hat{\theta}_i$ define the time-dependent pseudo-values for the incompletely observed outcomes in the same way for both censored and uncensored participants [242]. Similarly, in a competing risks setting, $\hat{\theta} = C\hat{I}F_c(t) = E(P(T_c \leq t))$, thus pseudo-value estimates can be derived as $nC\hat{I}F_c(t) - (n-1)C\hat{I}F_c{}^{-i}(t)$ for the observed risk of event c in individual i, in the presence of both right censoring and the competing risk [245, 246].

The derivation of pseudo-values leads to new, complete outcome observations for everyone in the dataset (no longer missing event indicators due to censoring), which are correlated across individuals due to their method of derivation. These new non-missing outcomes can then be used in 'standard' regression models, to relate the outcome with covariates of interest at a given time point. Generalised estimating equations can be used, with appropriate link functions to allow for logistic (logit-link), Cox (cloglog-link), or Fine-Gray (cloglog-link) type models to be fit on the continuous pseudo-value outcome, with robust standard errors to account for correlation between observations

[243, 245, 246]. Importantly, the resulting equations are approximately unbiased only where censoring is independent of covariates [242]. In the case of dependent censoring, modifications of the method are possible to achieve unbiased estimation [247], though these modifications are beyond the scope of this thesis. Instead independent censoring is assumed in the presented scenario.

A key benefit is that these pseudo-values provide a set of continuous outcome values for which various further analysis techniques are applicable, thus their use aligns with the themes of the thesis and with previous chapters, where continuous outcomes have been of particular interest. In prediction modelling research, calibration performance should presented visually, for example using a loess smoother to give a calibration curve across individuals. Relationships between observed and predicted outcomes in survival data are hard to visualise in this way. With the use of pseudo-values to define observed survival, comparisons through scatters and smoothers become possible [244]. For a pseudo-value regression model with a single covariate (predicted risk for the event of interest), a scatter-plot with a smoother overlaid is possible, giving a plot akin to a standard calibration plot.

Computation of pseudo-values can be time consuming, because the base estimator (for example, $CIF(t)$) needs to be re-calculated for each individual in the dataset [244]. Approximations to the pseudo-value calculation have been proposed, for example using an 'infinitesimal jackknife' process, to speed up this process in large datasets [248], giving values are asymptotically identical to full pseudo-values. Despite being conducted in large EHR data, the approach used in this chapter was that of full pseudo-value calculation, without approximation.

6.3 Part (i): Sample size calculation for external validation of a prediction model with a survival outcome

6.3.1 Sample size requirements for external validation

Sample size recommendations for the external validation of a prediction model have been developing in recent years. While criteria for the minimum sample size required to validate a prediction model with a continuous outcome have been discussed in Chapters 4 and 5, requirements for the external validation of a model for a survival outcome have not yet been discussed in this thesis.

Following the publication of the methods proposed in Chapter 4 [92], further recommendations have been published for validating both binary and survival outcome prediction models [95, 94, 93]. Prior to the publication of these new methods, sample size for external validation of such models was based on rule-of-thumb recommendations, such as those from Collins et al [172], which suggest at least 100 (thought ideally 200 or more) events in the external validation data. The following section of this chapter demonstrates the application of these recently published sample size assessment methods for the external validation of the survival model to predict falls risk, and compare the results to the available external validation data.

6.3.2 Process for determining appropriate sample size

Newly proposed simulation-based calculations for the sample size needed in the external validation of a survival model [93] allow tailoring of sample size calculations to a specific clinical question, potentially resulting in a higher recommended minimum than the 100 or 200 events of Collins et al [172]. An expanded version of the approach suggested by Riley et al follows in Figure 6.2 below.

Following the iterative steps described in Figure 6.2, the empirical distribution of calibration curves should also be assessed by considering the set of simulated calibration curves on the same set of axes. This could be presented, for example, as a 95% bootstrap confidence region around the anticipated calibration curve when estimated in a sample of the indicated minimum size to achieve the pre-specified target precision in the performance metrics. This reveals the variability in the calibration curve that might be observed in practice. If variability in the calibration curve is too high, then the calculation should be repeated with more stringent precision requirements for calibration measures.

This calculation method was applied in the assessment of the sample size available for the external validation of the STRATIFY-Falls prediction model. Not only was the full sample size available of interest (which, being from EHR across England, was likely ample for external validation of the model), but also the sample size available across the individual GP practices that formed clusters within the external validation data.

Figure 6.2: Summary of the steps involved in the sample size calculation for the external validation of a clinical prediction model with a survival outcome, as adapted from Riley et al 2021

STEP 1: Set up process
Specify the following:
(i) time point of interest for checking model performance, t
(ii) model's anticipated linear predictor (LP_i) distribution in the validation population
(iii) overall outcome risk, $F(t)$ (or $1 - S(t)$), in the validation population at time t
(iv) assumed distribution of survival times, T_i, in the validation population, conditional on LP_i
(v) assumed distribution of censoring times, C_i, in the validation population
(vi) maximum follow-up time in the validation population, C_{max}
(vii) target values for the standard error of key performance measures, e.g., for the calibration slope, $SE_{\hat{\lambda}_{cal}} \leq 0.051$

STEP 2: Choose starting sample size
Specify a starting sample size, n, and generate a dataset containing this number of individuals as a starting point

STEP 3: Simulate values of LP_i
For each individual i in the dataset, simulate a value of LP_i from the assumed linear predictor distribution, specified in step 1

STEP 4: Generate values of $\hat{F}_i(t)$ for each individual
For the time point of interest (t, specified in step 1), calculate individual-level predictions of outcome risk, $\hat{F}_i(t)$, based on the format of the existing prediction model equation. For example, this existing model will typically be of the form $\hat{F}_i(t) = 1 - \hat{S}_0(t)^{exp(LP_i)}$ where $\hat{S}_0(t)$ is the baseline survival probability at time t, and LP_i is the value of the linear predictor for individual i, from step 3.

STEP 5: Generate values of T_i for each individual
Randomly generate observed survival times for each individual (T_i) according to the assumed distribution from step 1, conditional on their LP_i value. For each individual, set their outcome status to be 1 ($D_i = 1$, an event) and their follow-up time to be their survival time, $\tilde{T}_i = T_i$.

STEP 6: Generate values of C_i for each individual
Randomly generate a censoring time (C_i) for each individual, under the censoring distribution assumed in step 1. For those with a survival time (from step 5) later than their censoring time ($T_i > C_i$) or the maximum follow-up time ($T_i > C_{max}$), allocate their event status as 0 ($D_i = 0$, no event) and update their follow-up time to be the earliest of their censoring time or the maximum follow-up time, $\tilde{T} = \min(C_i, C_{max})$.

STEP 7: Estimate model performance
For the chosen time point of interest for prediction, generate pseudo-values $\tilde{F}_i(t)$ for observed event probabilities $F_i(t)$, based on $\tilde{T}_i$ and D_i (generated in steps 5 and 6). Estimate key model performance estimates and their standard errors, including the calibration slope and any other measures of interest.

STEP 8: Repeat and store
Repeat steps 2 to 7 many times (Riley et al suggest 1000), each time storing the obtained estimate and standard error of the performance estimates.

STEP 9: Assess sample size suitability
Summarise the mean standard error of the desired performance measure across repetitions. If this mean is equal to the targeted value specified in step 1, then the sample size proposed in step 2 is the minimum sample size required to obtain a precise estimate of the given metric. Otherwise, repeat steps 2 to 9 with an alternative sample size, e.g., if targetting a precise estimate of the calibration slope, where $SE_{\hat{\lambda}_{cal}}$ is greater than the target value, increase n; where $SE_{\hat{\lambda}_{cal}}$ is smaller than the target value, decrease n.

6.3.3 Simulation set up in the context of the STRATIFY-Falls model

(i) Time point of interest for checking model performance, t

The primary time point of interest for checking model performance was ten years after antihypertensive treatment was indicated, thus $t = 10$. Follow up times of one and five years were also of interest as secondary outcomes, and, though the linear predictor distribution is consistent across all analysis times, the overall outcome risk was very different (with fewer severe falls events observed within the shorter time frames). This concern will be consistent across all prediction models for a time-to-event outcome where multiple time points are of interest for validation.

Targeting precise estimation of the STRATIFY-Falls model at the primary outcome time point seemed the most logical approach, though it is worth noting that an external validation set exactly meeting the required sample size for ten-year falls risk assessment would likely give wide confidence intervals for performance statistics measured at one year.

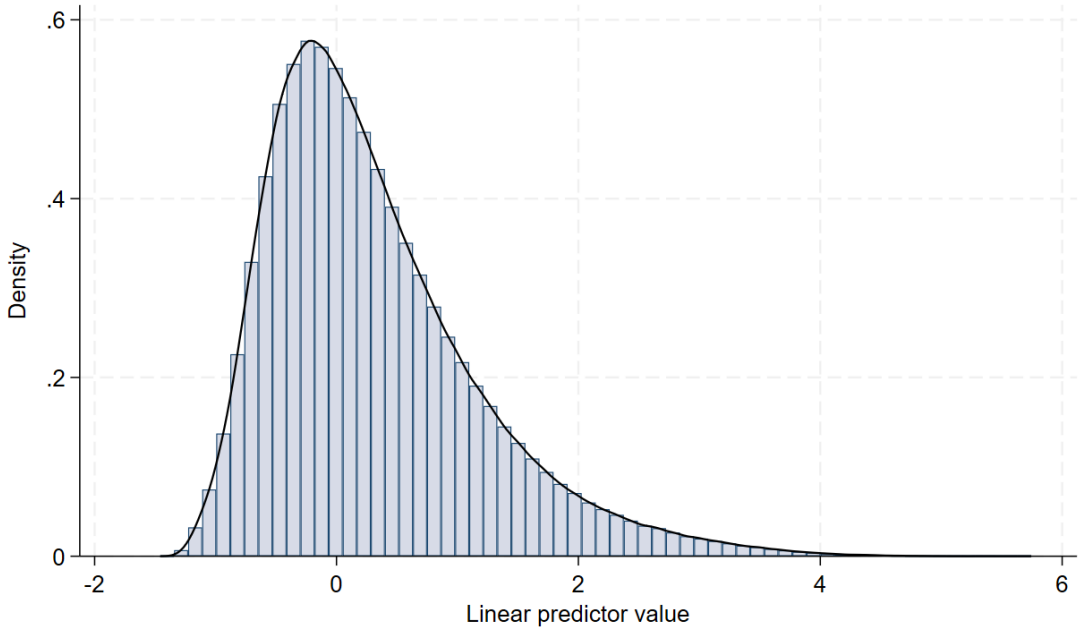
(ii) The model's anticipated linear predictor ($\mathbf{LP_i}$) distribution in the validation population

The prediction model equation was applied in the external validation data and a summary of the linear predictor distribution is given in Table 6.1. The linear predictor was found to follow an approximate skew-normal distribution, which was best simulated using the nearest available options in the *sknor* package in Stata software: a mean of 0.41, variance of 0.8, skewness parameter of 1, and kurtosis parameter of 4.

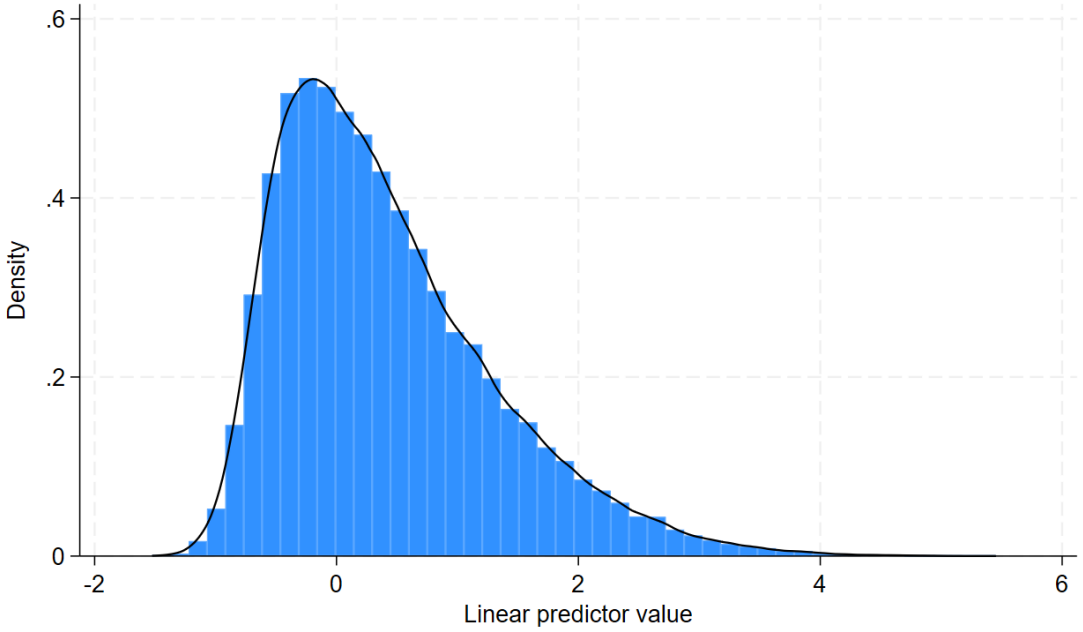
Table 6.1: Summary of the STRATIFY-Falls model linear predictor distribution, with parameters observed on application of the model in the external validation data, the nearest available options in the *sknor* package in Stata software, and observed in an example simulated dataset of 50000 LP_i values using these parameters to define the LP_i distribution

Parameter	Observed	Simulated	Example simulation
Mean	0.408	0.41	0.395
Variance	0.800	0.8	0.798
Skewness	1.124	1	1.015
Kurtosis	4.300	4	4.193

Figure 6.4 shows the distribution of the linear predictor, as observed in the external validation data and in a single simulated dataset of 50000 LP_i values using the above parameter values. When assessed visually, this histogram of simulated values closely matched the linear predictor distribution seen in the external validation data. Summarising the LP_i in this example simulation gave a mean of 0.395, variance of 0.798, skewness parameter of 1.015, and kurtosis parameter of 4.193 - a close match to the distribution summary values in the external validation data, again suggesting this was a reasonable approximation.



(a) Observed distribution



(b) Simulated distribution

Figure 6.4: Observed and simulated distributions of the linear predictor for the STRATIFY-Falls model, with parameters measured in the external validation data and the nearest available options in the *sknor* package in Stata software used in the simulation.

(iii) Overall outcome risk, $\mathbf{F(t)}$ (or $\mathbf{1 - S(t)}$), in the validation population at time t

The observed cumulative incidence of severe falls by ten years varied considerably across individual GP practices, ranging from $F(10) = 1.3\%$ up to $F(10) = 17.6\%$. This variation can be seen in Figure 6.5, and was more notable in smaller GP practices, with fewer eligible individuals, where uncertainty in the estimation of the outcome proportion was higher.

When pooled across all GP practices in the external validation data, there was an observed overall 10-year outcome risk of 8.7% (95%CI: 8.6% to 8.9%). Thus, there was an expected 91% survival (those with no severe fall event) at ten years, across the whole population. This value was taken forward to the simulation as, even though there was likely some variation in the true outcome proportion in individual practices (due to population and case-mix differences), no practices lay outside the 95% prediction region so this value was deemed appropriate for overall sample size calculation.

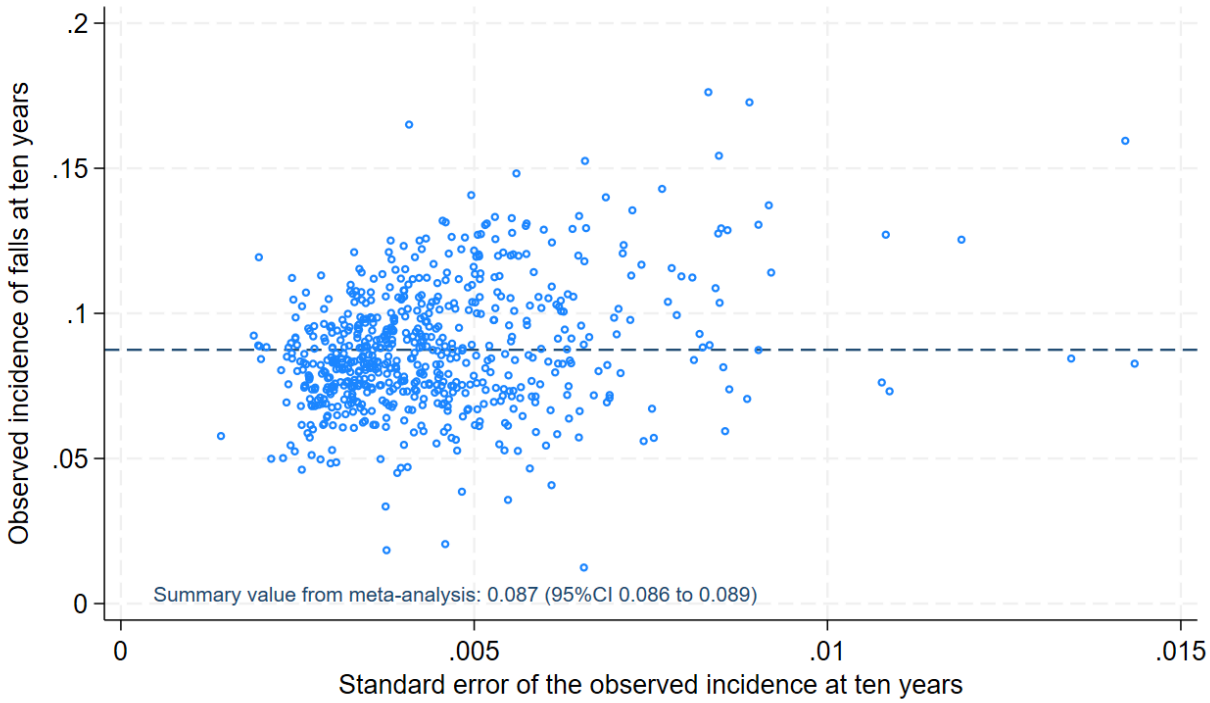


Figure 6.5: Observed outcome risk ($F(10)$) against standard error for each GP practice in the CRPD Aurum external validation data. Blue dashed line shows pooled incidence across all practices.

(iv) Assumed distribution of survival times, T_i, in the validation population, conditional on LP_i

An assumed parametric distribution of survival times was required to preserve the 91% survival at ten years that was observed in the external validation data. Conditional on the values of the linear predictor, LP_i, an exponential distribution with a baseline rate parameter 0.0042 was found to give an appropriate outcome risk at ten years. The methods proposed by Riley et al do not allow for the inclusion of the competing risk of death in the sample size calculation for external validation, thus this was not accounted for in the calculation shown here. Future research might consider adapting the simulation process to also account for this risk.

(v) Assumed distribution of censoring times, C_i, in the validation population

A constant censoring rate was assumed, with censoring times following an exponential distribution. Examining the censoring rate in the external validation dataset indicated that censoring was high, with 62% participants censored before 10 years. Through trial and error, rate parameter of 0.096 was found to give a probability of censoring by 10 years of about 62%.

(vi) The maximum follow-up time in the validation population, C_{max}

Follow-up times in the external validation data were capped at 10 years, thus the maximum possible follow-up for any individual $C_{max} = 10$.

(vii) Target values for the standard error of key performance measures

Riley et al suggest that the sample size calculation for the external validation of a time-to-event prediction model is most likely to be driven by the requirement for a precise estimation of model calibration, as defined by the estimate of the calibration slope ($\hat{\lambda}_{cal}$) at the time point of interest. Therefore, simulations were conducted to identify what sample size, n, would be sufficient to achieve a precision of $SE_{\hat{\lambda}_{cal}} = 0.051$, to target a 95% confidence interval width of no more than 0.2 around a calibration slope estimated using Cox regression for the calibration model (as was also suggested in the equivalent criterion for the precise estimation of the calibration slope of a continuous outcome prediction model, shown in Chapters 4 and 5).

6.3.4 Calculation in the context of the STRATIFY-Falls model

The simulation process shown in Figure 6.2 was set up using the information specified in Section 6.3.3, tailoring the calculation to the validation of the STRATIFY-Falls model in the CPRD Aurum data. This process needed to be repeated iteratively until the sample size defined in step 2 resulted in the target $SE_{\hat{\lambda}_{cal}}$ on average across the 1000 simulated datasets. Riley et al [93] do not recommend an approach to inform decisions on which sample size option should be tested at subsequent iterations of the process, thus two methods were trialled to assess their efficiency: using linear interpolation and using Newton's divided difference interpolation. For both approaches, the initial two values of n were chosen with one smaller ($n = 500$) and one bigger ($n = 20,000$) than the anticipated sample size to achieve $SE_{\hat{\lambda}_{cal}} = 0.051$, though the same concepts would apply (though possibly less efficiently) with any two starting values.

Linear interpolation

Starting sample size estimates of $n = 500$ and $n = 20,000$ were input into the simulation process, giving average $SE_{\hat{\lambda}_{cal}}$ across simulations of 0.208 and 0.013 respectively. Thus, the sample size needed to gain $SE_{\hat{\lambda}_{cal}} = 0.051$ an average would lie somewhere between these two values for n. A simple linear function $f(n) = a_0 + a_1 n$ was fitted between these two points, with a slope of $a_1 = \frac{f(n_2)-f(n_1)}{n_2-n_1}$ and intercept defined by $a_0 = f(n) - a_1 n$ for one of the known $(n_i, f(n_i))$ pairs. The point $f(n) = 0.051$ on this linear function was used to inform the next "best guess" of the required sample size.

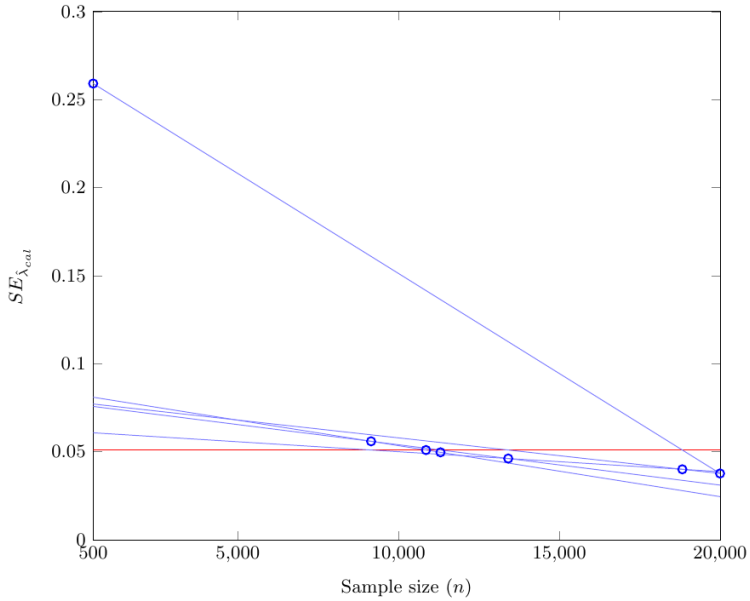
Accordingly, at each stage $f^{-1}(0.051)$ was calculated, giving the value of n to be used in the subsequent iteration:

$$n_{i+1} = n_{i-1} + \frac{(0.051 - f(n_{i-1}))(n_i - n_{i-1})}{f(n_i) - f(n_{i-1})}$$

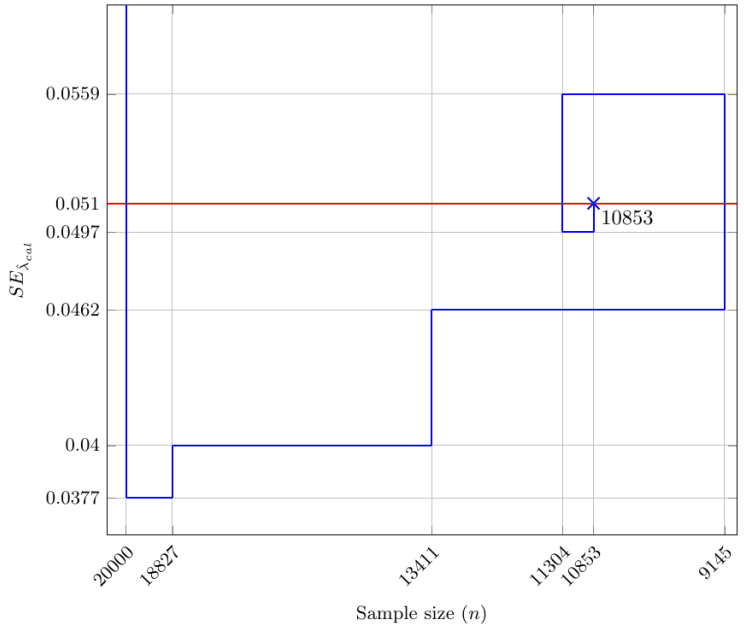
The sequence of n chosen by this method, with their corresponding mean $SE_{\hat{\lambda}_{cal}}$ across simulations, are presented in Table 6.2. Linear functions at each iteration, and proximity of sequentially tested sample sizes to the target $SE_{\hat{\lambda}_{cal}} = 0.051$ are further shown in Figure 6.7. This approach resulted in an estimated sample size requirement of 10,853 patients to ensure precise estimation of the calibration slope, taking only seven iterations to reach a conclusion on the required sample size.

Table 6.2: Sample sizes (n) considered, with their corresponding mean $SE_{\hat{\lambda}_{cal}}$ across simulations, for the sequence of n determined through linear interpolation.

i	n_i	$SE_{\hat{\lambda}_{cal}} = f(n_i)$	$\lvert f(n_i) - 0.051 \rvert$
1	500	0.2591525	0.2081525
2	20000	0.0376685	0.0133315
3	18827	0.0400418	0.0109582
4	13411	0.0461711	0.0048289
5	9145	0.0559455	0.0049455
6	11304	0.0496936	0.0013064
7	10853	0.0509753	0.0000247



(a) Linear function derived at each sample size iteration



(b) Sequence of tested sample sizes

Figure 6.7: Sample sizes (n) considered, with their corresponding expected standard error in the calibration slope ($SE_{\hat{\lambda}_{cal}}$), for the external validation of the STRATIFY-Falls model, for the sequence of n determined through linear interpolation. Red lines show the target precision of $SE_{\hat{\lambda}_{cal}} = 0.051$

Anticipated uncertainty in the calibration curve

To assess the anticipated uncertainty in the calibration curve resulting from an external validation using 10,853 individuals, example calibration curves with bootstrapped 95% confidence intervals were generated for a sample of simulated datasets of this size. These calibration curves and, importantly, their confidence intervals were inspected, to ascertain whether the precision of the curve was acceptable. An example calibration curve for one simulated dataset is shown in Figure 6.8. Confidence intervals suggested high precision in the calibration curve for lower predicted outcome probabilities. While the width of confidence intervals increased for predictions in the higher range, calibration curves are still anticipated to be estimated with adequate precision across the full range of predictions when using a sample of this size for external validation.

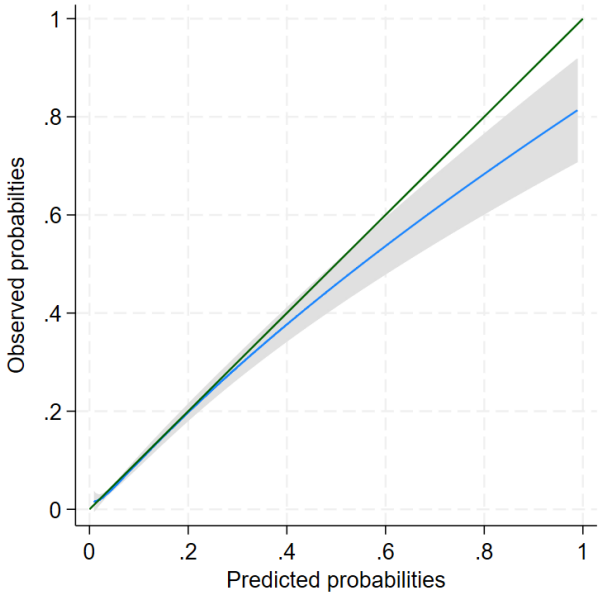


Figure 6.8: Anticipated uncertainty in the calibration curve with a simulated sample of 10,853 individuals. Grey shaded area shows the bootstrapped 95% confidence interval around the estimated calibration curve (blue line).

Newton's divided difference interpolation

As an alternative to linear interpolation, Newton's polynomial interpolation was trialled to inform the next sample size n in the sequence to test. This was done to assess whether an alternative approach might result in a more efficient sequence of iterations than the simple linear method. Newton's (divided difference) polynomials were used fit a polynomial function, $f(n)$, through all known $(n_i, f(n_i))$ pairs from previous iterations. These functions are traditionally defined as follows, for a set of ρ different $(x,\ f(x))$ pairs:

$$f_\rho(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + \cdots + a_\rho(x - x_0)(x - x_1)\ldots(x - x_\rho)$$

$$\therefore f_\rho(x) = \sum_{i=0}^{\rho} a_i\rho_i(x), \text{ where } \rho_i(x) = \prod_{j=1}^{i-1}(x - x_j)$$

In this context, $x = n$ (the sample size for external validation) and $f(x) = f(n) = SE_{\hat{\lambda}_{cal}}$ (the resulting standard error in the estimate of the calibration slope, averaged across simulations), with values for a_i determined sequentially, by substituting in values for known $(n_i, f(n_i))$ pairs and rearranging Newton's formulae:

1. $f_0(n_0) = a_0 \implies a_0 = f(n_0)$

2. $f_1(n_1) = a_0 + a_1(n_1 - n_0) \implies a_1 = \dfrac{f(n_1) - f(n_0)}{n_1 - n_0}$

3. $f_2(n_2) = a_0 + a_1(n_2 - n_0) + a_2(n_2 - n_0)(n_2 - n_1)$

$$\implies a_2 = \frac{\frac{f(n_2)-f(n_1)}{n_2-n_1} - \frac{f(n_1)-f(n_0)}{n_1-n_0}}{n_2 - n_0} = \frac{f_1(n_2) - f_1(n_1)}{n_2 - n_0}$$

4. $f_3(n_3) = a_0 + a_1(n_3 - n_0) + a_2(n_3 - n_0)(n_3 - n_1) + a_3(n_3 - n_0)(n_3 - n_1)(n_3 - n_2)$

$$\implies a_3 = \frac{\frac{\frac{f(n_3)-f(n_2)}{n_3-n_2} - \frac{f(n_2)-f(n_1)}{n_2-n_1}}{n_3-n_1} - \frac{\frac{f(n_2)-f(n_1)}{n_2-n_1} - \frac{f(n_1)-f(n_0)}{n_1-n_0}}{n_2-n_0}}{n_3 - n_0} = \frac{f_2(n_3) - f_2(n_2)}{n_3 - n_0}$$

$\vdots$

$\rho + 1.$ $f_\rho(n_\rho) = \sum_{i=0}^{\rho} a_i \rho_i(n_\rho) \implies a_\rho = \frac{f_{\rho-1}(n_\rho) - f_{\rho-1}(n_{\rho-1})}{n_\rho - n_0}$

As with linear interpolation, at each iteration, the point $f(n) = 0.051$ on this polynomial function was used to inform the next "best guess" of the required sample size. This was calculated as the real (non-complex) root of equation $f(n) = 0.051 - f_\rho(n)$, within the range of n_i from the current known $(n_i, f(n_i))$ pairs. This root was determined using the *uniroot* function in R version 4.2.2.

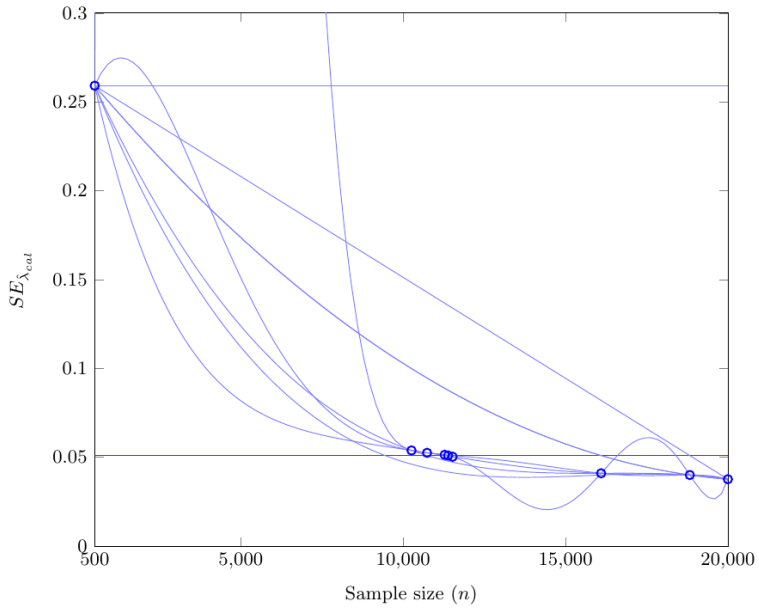
Table 6.3 shows the sequence of n chosen by this method, with the associated mean $SE_{\hat{\lambda}_{cal}}$ across simulations. Plots of the polynomial functions defined at each iteration are shown in Figure 6.10, along with the proximity of each tested sample size to the target $SE_{\hat{\lambda}_{cal}} = 0.051$. After 9 iterations, this approach gave an estimated sample size requirement of 11,385 patients. This value is slightly higher than that found with linear interpolation, likely due to chance differences arising during the simulation process.

As before, precision in calibration curves was high for lower predicted probabilities and decreased (with wider confidence intervals) for higher predictions. Overall, calibration curves were anticipated to be sufficiently precise across the full range of predicted values when using a sample of 11,385 for external validation of the STRATIFY-Falls models at ten years.

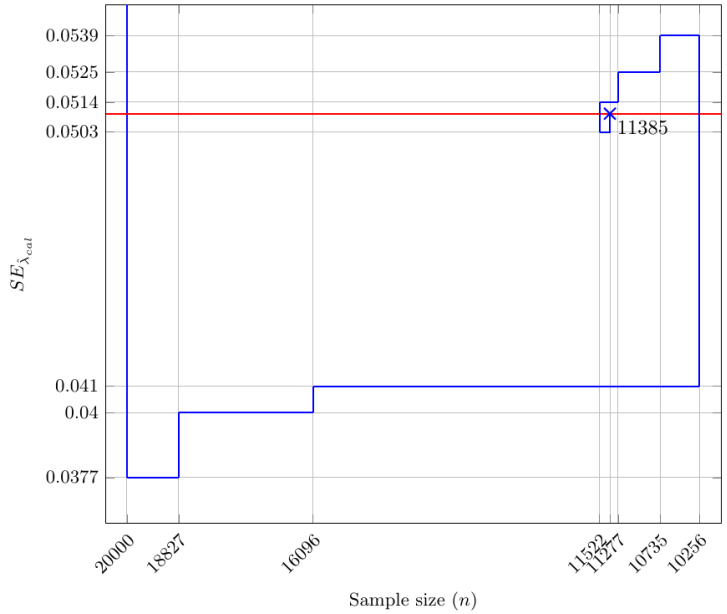
Notably, by the final iteration, the polynomial function (of degree eight) fit the data points very badly. Though the curve passed though all $(n_i, f(n_i))$ pairs from previous iterations by design, the shape of the function was far removed from the logical monotonic-decreasing function that would be expected of the relationship between standard error and sample size. Thus, the complexity introduced by such high order of polynomials not only does not facilitate estimation, but actively hinders it.

Table 6.3: Sample sizes (n) considered, with their corresponding mean $SE_{\hat{\lambda}_{cal}}$ across simulations, for the sequence of n determined through Newton's divided difference interpolation.

i	n_i	$SE_{\hat{\lambda}_{cal}} = f(n_i)$	$\lvert f(n_i) - 0.051 \rvert$
1	500	0.2591525	0.2081525
2	20000	0.0376685	0.0133315
3	18827	0.0400418	0.0109582
4	16096	0.0410136	0.0099864
5	10256	0.0538596	0.0028596
6	10735	0.0525219	0.0015219
7	11277	0.051419	0.000419
8	11522	0.0503237	0.0006763
9	11385	0.051	0



(a) Derived polynomials with each sample size iteration



(b) Sequence of tested sample sizes

Figure 6.10: Sample sizes (n) considered, with their corresponding expected standard error in the calibration slope ($SE_{\hat{\lambda}_{cal}}$), for the external validation of the STRATIFY-Falls model, for the sequence of n determined through Newton's divided difference interpolation. Red lines show the target precision of $SE_{\hat{\lambda}_{cal}} = 0.051$

6.3.5 Assessment of sample size available for external validation

Using the larger of these simulation-based calculations to be conservative, targetting a 95% confidence interval of width 0.2 around the estimate of the calibration slope resulted in a minimum sample size requirement of 11,385 patients, including approximately 991 serious fall events. Using CPRD Aurum, a large EHR database, for external validation allowed inclusion of many more than 991 events across the whole population, meaning these recommendations were easily met. When using such databases, however, it is also essential to consider the natural clustering of the data by GP practice. In this example, the variability in model performance across GP practices was also of interest. Ideally, there would also have been sufficient data in each practice to allow accurate assessment of the model's performance in these different populations.

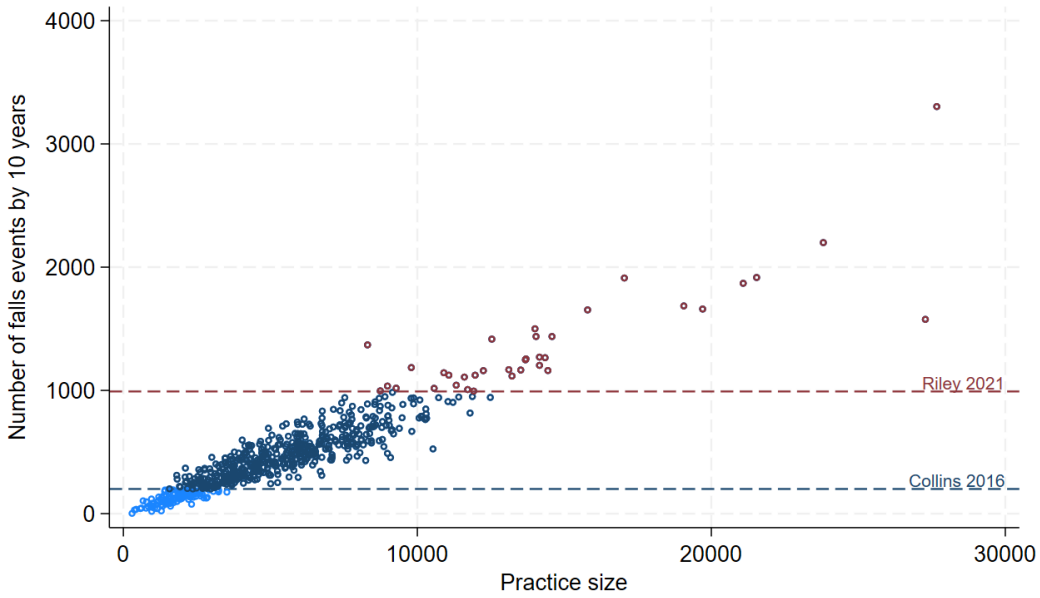


Figure 6.11: Numbers of falls events by GP practice size in the CRPD Aurum external validation data. The red line shows the recommended minimal sample size per Riley 2021, 991 falls events, while the blue line shows the recommended minimal per Collins 2016, 200 events.

When comparing to the sample size recommendations from tailored simulations, per Riley 2021 [93], 103 (14.5%) practices had a sufficient effective sample size to accurately estimate the calibration slope, recording more than 991 falls events at 10 years. An additional 488 (68.6%) practices recorded more than 200 falls events, though less than 991, meeting the Collins 2016 [172] criteria while failing to meet the Riley 2021 recommendations.

When comparing to the lower of Collins' suggestions (a minimum of 100 events, not tailored to the model of interest), 33 (4.6%) of GP practices in CPRD Aurum fell short of the sample size requirement, if the aim had been to evaluate model performance in that population specifically. Although sufficient data were included in the database overall to assess the model performance on average across the UK, it is important to carefully consider any conclusions around model performance in those smaller GP practices, especially where smaller practices might represent sub-populations that are under-represented in the full CPRD Aurum data.

6.4 Part (ii): External validation of the STRATIFY-Falls model

6.4.1 Methods for the assessment of model performance

Missing data

Where a patient's records had no entry describing the diagnosis of comorbidities or prescribed medications, it was assumed that no such diagnosis or prescription was present. Aside from comorbidities and medications, predictor variables included in the STRATIFY-Falls model but missing in some observations of the validation data were cholesterol level, ethnicity, deprivation score, smoking status, and alcohol consumption.

Multiple imputation by chained equations was used to impute these missing values across the whole of the external validation cohort, with no allowance for the clustering of patients within GP practices, potentially masking some of the heterogeneity between practices [211]. The imputation models included all predictor variables, along with binary event indicators for falls and for the competing event of death by 10 years, and a Nelson-Aalen estimator for the cumulative cause-specific hazards of each of these possible event types [249, 250]. A total of ten imputations were generated, a value lower than the percentage of incomplete observations (as was referenced in previous chapters) [173], which was chosen for practical reasons due to the size of the data and the computational intensity of generating and analysing a higher number of imputed datasets.

Imputations were assessed for consistency by comparing density plots, histograms, and summary statistics across imputations and back to the complete values. Predictive performance measures were then estimated in each imputed dataset separately, before combining estimates across imputations using Rubin's Rules [177], where appropriate.

Measures of performance

Model performance was assessed through measures of calibration, discrimination, overall model fit, and clinical utility, as were described in previous chapters.

Calibration performance was assessed through the use of pseudo-values, based on the CIF for falls, accounting for competing risk of death, calculated by the Aalen–Johansen method [251]. These pseudo-values gave jackknife estimates that represented an individual's contribution to the overall estimate of the cumulative incidence function for serious falls, giving a continuous outcome value for each individual that represented their observed 10-year falls risk while accounting for both the competing event of death and for right censoring. Observed pseudo-values for observed 10-year falls risks were compared to the predicted risks from the STRATIFY-Falls model in the calculation of the Observed/Expected ratio and allowed the estimation of smooth calibration curves across individuals, which were presented in calibration plots.

Discrimination performance was measured using an inverse probability of censoring weighted estimate of the time dependent area under the ROC curve (C-statistic) at 10 years [252] and the D-statistic, as proposed by Royston and Sauerbrei [253]. Overall model fit was assessed using Royston and Sauerbrei's R^2_D [253]. Clinical utility was assessed using net benefit, plotted against potential thresholds for clinical action on decision curves. Net benefit at each threshold probability was calculated using the CIF of falling (accounting for the competing risk of death) to define the number of true positive and true negative classifications [69].

Heterogeneity across GP practices

Heterogeneity in model performance across different GP surgeries was assessed using a random effects meta-analysis, by restricted maximum likelihood estimation (REML), given that the case-mix and incidence of falls were known to vary considerably between practices, so too were model performance estimates expected to vary [65, 63]. Confidence intervals for pooled estimates were derived using the Hartung-Knapp-Sidik-Jonkman variance correction, to account for any uncertainty in the estimate of the between-practice variance [217]. Performance estimates were first calculated for each GP practice, within each imputed dataset, with estimates combined across imputations (within GP practice) by applying Rubin's Rules where appropriate [177]. Estimates were then combined across practices using the random effects meta-analysis to gain estimates of average model performance across all populations [213].

6.4.2 External validation cohort characteristics

After exclusions for failure to meet the inclusion criteria, also being included in the model development data, and a lack of data linkage (Figure 6.12), a total of 3,805,366 patients were included in the cohort for external validation of the STRATIFY-Falls model, from across 711 GP practices, with 206,956 (5.4%) experiencing fall events during the ten-year follow-up. A further 334,552 (8.8%) patients died during follow-up from causes unrelated to a fall, prior to any fall occurring.

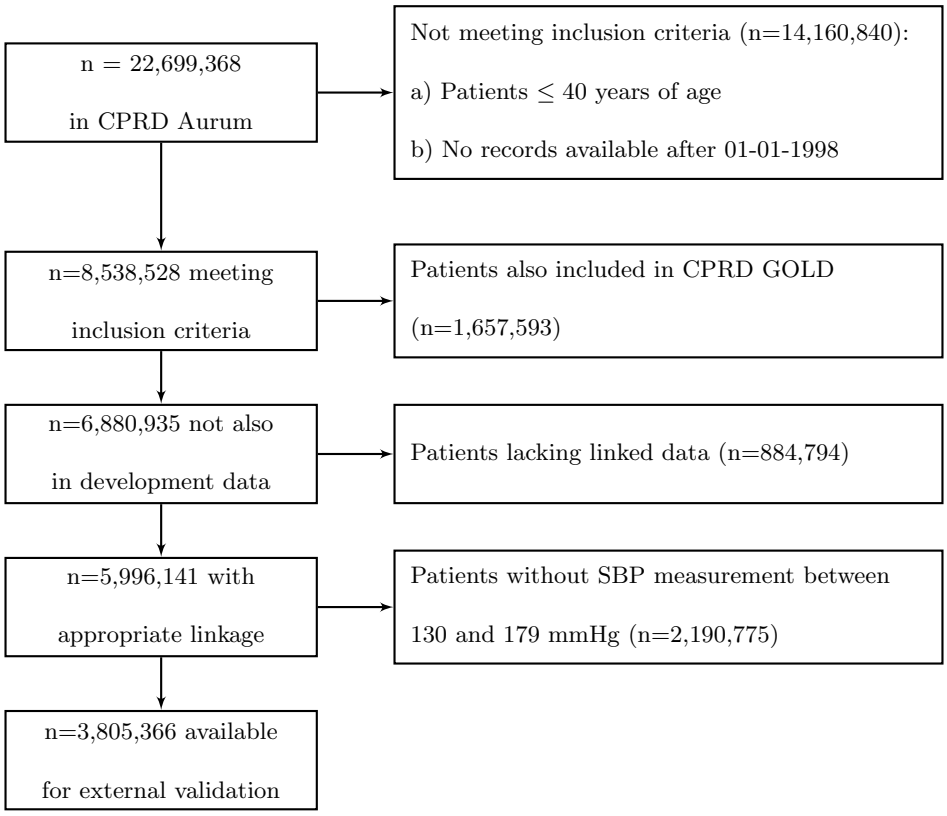


Figure 6.12: PRISMA flowchart, showing the number of eligible participants for external validation from the total CPRD Aurum population

Median follow up time in the validation cohort was 6.7 years (LQ to UQ: 2.7 to 10 years). The median time until a serious fall was 4.3 years (1.9 to 7.1), with those dying prior to any fall being followed up for a median of 3.8 years (1.6 to 6.5). Characteristics of the external validation cohort as a whole, and those who experienced the outcome and competing events, are given in Table 6.4.

The predictor most often missing in the validation data was total cholesterol, with 48.3% individuals missing this measurement overall. Data were also missing for some individuals for ethnicity (19%), deprivation score (9.7%), smoking status (7%), and alcohol consumption (18%). All other predictor values were assumed complete, with an absence of data entries describing comorbidity diagnoses or prescribed medications being taken to mean no such diagnoses or prescriptions were present for that individual. Summaries of imputed values showed no obvious differences from distributions of complete values.

Given the external validation data came from a very large UK database, containing information from 738 different GP practices, model performance was expected to vary across clusters, as was seen in the external validation across different populations in Chapter 5. Not only was the observed incidence of serious falls known to vary across practices (see Figure 6.5), case-mix in terms of age distribution, deprivation status, and ethnicity was also expected to be very different in different areas of the UK. This difference in demographics and outcome incidence would likely affect the performance of the model, with those clusters most similar in composition to the population average being most likely to show good performance. Heterogeneity in the model performance by different measures is therefore summarised in the following sections.

Table 6.4: Characteristics of the sample of CPRD Aurum used in the external validation of the STRATIFY-Falls prediction model. Values are number (percentage) unless otherwise stated.

	Total	Falls	Competing event
n	3,805,366	206,956	334,552
Age, mean (SD)	58.6 (13.3)	72.8 (12.7)	73.1 (12.3)
Sex (Female)	1959489 (52%)	134945 (65.2%)	165689 (49.5%)
Systolic blood pressure, mean (SD)	143.8 (12.3)	147.2 (13.2)	147.7 (13.3)
Diastolic blood pressure, mean (SD)	83.9 (9.8)	81.9 (10.2)	82.0 (10.3)
Cholesterol, mean (SD)	5.5 (1.2)	5.4 (1.3)	5.4 (1.3)
Missing	1839116 (48.3%)	109708 (53.0%)	195390 (58.4%)
Ethnicity			
White	2041505 (54%)	194311 (93.9%)	206384 (61.7%)
Black	115279 (3%)	2239 (1.1%)	4019 (1.2%)
South Asian	94485 (3%)	2449 (1.2%)	3673 (1.1%)
Other	832614 (22%)	3442 (1.7%)	21458 (6.4%)
Missing	721483 (19%)	4515 (2.2%)	99018 (29.6%)
Deprivation Score			
IMD 1	790311 (20.8%)	41786 (20.2%)	66606 (19.9%)
IMD 2	732246 (19.2%)	41820 (20.2%)	68147 (20.4%)
IMD 3	684288 (18%)	40665 (19.7%)	67130 (20.1%)
IMD 4	630482 (16.6%)	40383 (19.5%)	65342 (19.5%)
IMD 5	597180 (15.7%)	42141 (20.4%)	67024 (20.0%)
Missing	370859 (9.7%)	161 (0.1%)	303 (0.1%)
Smoking status			
Non smoker	1475708 (39%)	77990 (37.7%)	109249 (32.7%)
Ex-smoker	1236061 (33%)	39087 (18.9%)	75081 (22.4%)
Smoker	838404 (22%)	66836 (32.3%)	105363 (31.5%)
Missing	255193 (7%)	23043 (11.1%)	44859 (13.4%)
Frailty index, median (LQ to UQ)	0.06 (0.03 to 0.08)	0.08 (0.06 to 0.17)	0.08 (0.06 to 0.17)

	Total	Falls	Competing event
n	3,805,366	206,956	334,552
Alcohol consumption			
Non drinker	864865 (23%)	59364 (28.7%)	89537 (26.8%)
Trivial drinker	998948 (26%)	47088 (22.8%)	71739 (21.4%)
Light drinker	696369 (18%)	26635 (12.9%)	44924 (13.4%)
Moderate drinker	246468 (7%)	9378 (4.5%)	17491 (5.2%)
Heavy drinker	74005 (2%)	5124 (2.5%)	6845 (2.1%)
Unknown amount	237464 (6%)	9631 (4.7%)	12117 (3.6%)
Missing	687247 (18%)	49736 (24%)	91899 (27.5%)
Comorbidities			
Previous falls	140886 (3.7%)	21697 (10.5%)	25124 (7.5%)
Memory problems	99264 (2.6%)	15996 (7.7%)	28636 (8.6%)
Mobility problems	85675 (2.3%)	13999 (6.8%)	22928 (6.9%)
Stroke	111462 (2.9%)	15704 (7.6%)	26703 (8%)
Multiple sclerosis	11328 (0.3%)	975 (0.5%)	1373 (0.4%)
Antihypertensive medications			
ACE inhibitors	478778 (13%)	38867 (18.8%)	67787 (20.3%)
Angiotensin II receptor blockers	136926 (4%)	11018 (5.3%)	14308 (4.3%)
Alpha blockers	68131 (2%)	6335 (3.1%)	11388 (3.4%)
Beta blockers	461329 (12%)	36317 (17.6%)	59019 (17.6%)
Calcium channel blockers	426151 (11%)	37590 (18.2%)	63764 (19.1%)
Diuretics	397980 (11%)	36418 (17.6%)	55934 (16.7%)
Other antihypertensives	19235 (1%)	1437 (0.7%)	2471 (0.7%)
Other medications			
Opioids	1213876 (32%)	84108 (40.6%)	121303 (36.3%)
Hypnotics, anxiolytics	750584 (20%)	52854 (25.5%)	78627 (23.5%)
Antidepressants	793690 (21%)	52820 (25.5%)	71452 (21.4%)
Anticholinergic medications	388513 (10%)	31542 (15.2%)	46255 (13.8%)
Follow up, median (LQ to UQ)	6.7 (2.7 to 10)	4.3 (1.9 to 7.1)	3.8 (1.6 to 6.5)

SD - Standard Deviation, LQ - Lower Quartile, UQ - Upper Quartile.

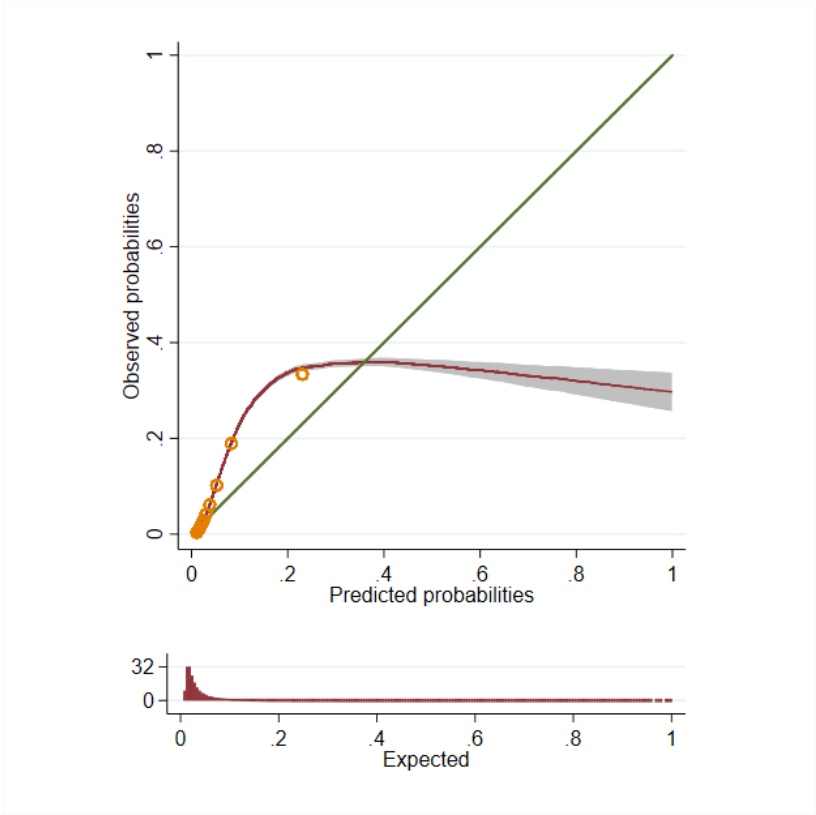
Scales: age (years), blood pressure (mmHg), cholesterol (mmol/L), follow up (years).

6.4.3 Model calibration performance

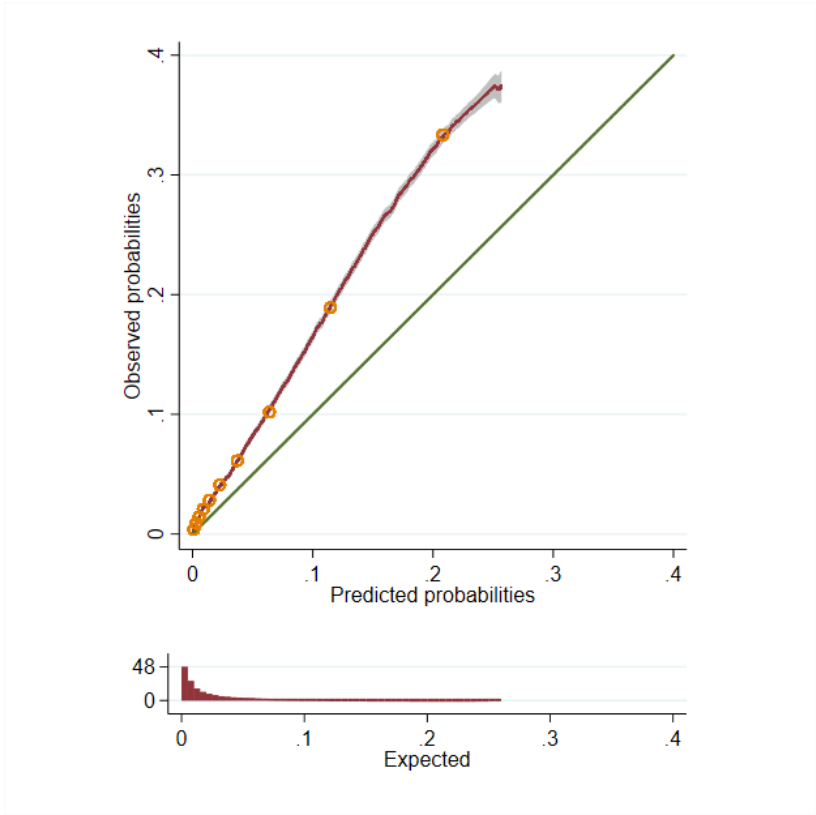
Calibration of the STRATIFY-Falls model was assessed through comparison of model predictions with the 10-year pseudo-values, calculated from the observed CIF in the external validation data. Overall summaries of calibration were visualised through average calibration curves, based on the pseudo-values calculated across all patients, with no allowance of clustering by GP practice. For practice-level calibration assessment, as subsequent pooling of calibration statistics, pseudo-values were generated in each practice individually to preserve heterogeneity in observed outcomes across practices.

When assessing calibration performance across the full external validation dataset, without accounting for the clustering by GP practice, pseudo-value-based calibration curves were estimated with high precision for both the Fine-Gray and the STRATIFY-Falls models. Plots in Figure 6.14 show these calibration curves and their 95% confidence intervals when generated in a random 10% of the data from a representative imputation. Even in this sub-sample of the full data, the uncertainty around both calibration curves was negligible due to the very large total sample size.

As was seen in the model development data, the Fine-Gray model was extremely miscalibrated at the 10-year time point. Predictions of risk below 30% were generally too low (with a calibration curve above the diagonal) and predictions higher than 30% falls risk were generally too high. The STRATIFY-Falls model had a more restricted range of predicted risks, with predictions going no higher than 26%. While recalibration of the model had corrected the miscalibration at 10 years in the model development cohort, under-prediction of risk on average was still evidence in the validation data.



(a) Fine-Gray model for 10-year falls risk



(b) STRATIFY-Falls model for 10-year falls risk

Figure 6.14: Average calibration plots (calculated without accounting for clustering by GP practice) for the Fine-Gray and STRATIFY-Falls models to predict falls in an example imputation of the CRPD Aurum external validation data.

Calibration performance varied considerably across practices, as can be seen through consideration of calibration curves of the STRATIFY-Falls model across practices (Figure 6.15). This variation was not evident in the average calibration curves shown in Figure 6.14. The vast majority of practice-level calibration curves can be seen to lie above the line of ideal calibration, suggesting the STRATIFY-Falls model consistently under-predicts the risk of a serious fall across most GP practices. The spread of curves across practices shows far more variation in calibration performance than might be implied from the uncertainty estimate in the overall calibration curves shown in Figure 6.14, likely due to the lack of allowance for practice-level clustering in the generation of this summary plot.

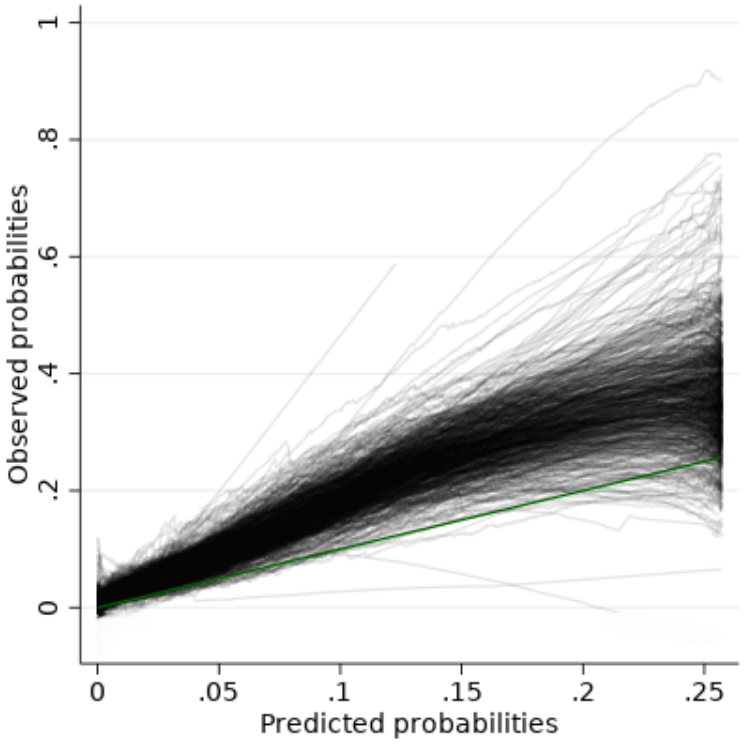


Figure 6.15: Calibration curves of the STRATIFY-Falls model to predict falls across GP practices in the CRPD Aurum external validation data. Green line indicates ideal calibration. "Predicted probabilities" axis cropped at maximum value of this model in the external validation data.

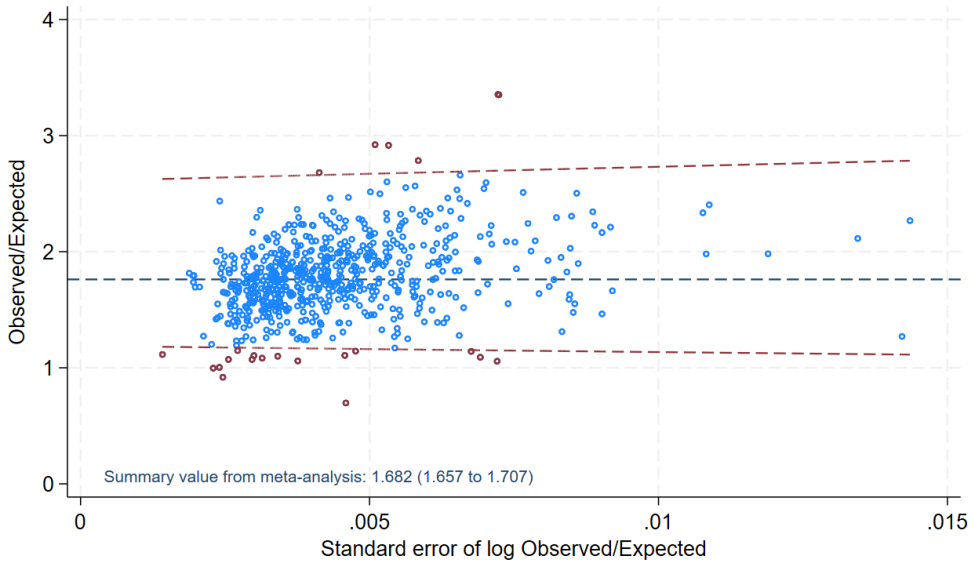
Heterogeneity in calibration performance is also evidenced in the assessment of the O/E ratio, through summary values in Table 6.6 and through scatter plots in Figure 6.17. Relatively high values for τ^2 and wide prediction intervals can be seen for both the Fine-Gray and the STRATIFY-Falls models, suggesting heterogeneous model calibration.

Though the entire of both prediction intervals indicate under-prediction of risk is expected in a new population, intervals are still very wide. The prediction interval from the STRATIFY-Falls model suggests that observed falls risks could be anywhere from 28% to 164% higher than predicted, if

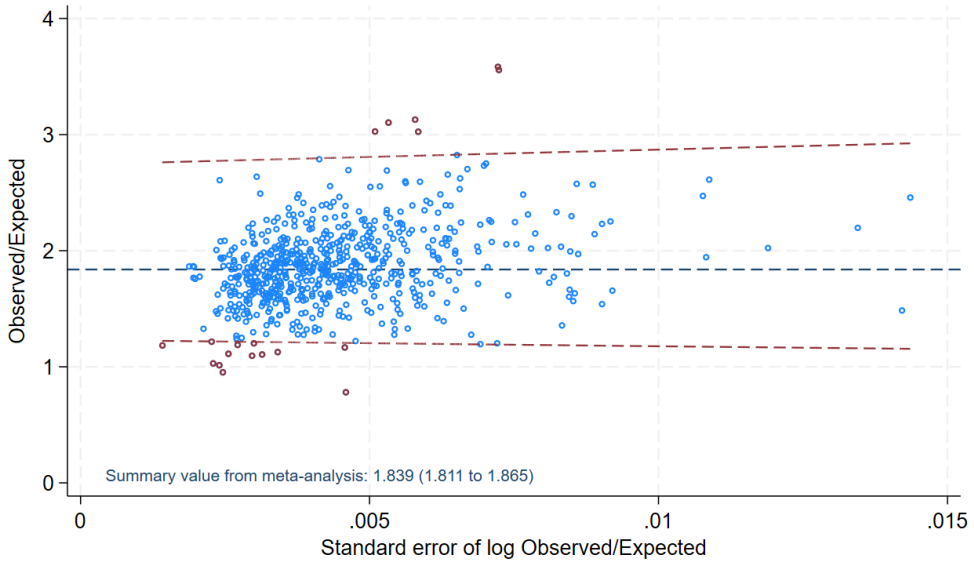
the model were applied in patients in a new GP practice from a similar setting to that used for this external validation.

Table 6.6: Observed/Expected ratio for the Fine-Gray and STRATIFY-Falls models on external validation, pooled across GP practices

	Fine-Gray model	STRATIFY-Falls model
Observed/Expected		
Pooled effect size (95% CI)	1.682 (1.657 to 1.707)	1.839 (1.811 to 1.865)
Prediction interval, 95%	1.139 to 2.484	1.284 to 2.638
τ^2 (95% CI)	0.038 (0.035 to 0.043)	0.0342 (0.031 to 0.038)



(a) Fine-Gray model



(b) STRATIFY-Falls model

Figure 6.17: Observed/Expected ratio of the Fine-Gray and STRATIFY-Falls models, by their standard errors, across GP practices in the CRPD Aurum external validation data. Blue dashed line shows summary value, red dashed lines show the 95% prediction interval.

6.4.4 Model discrimination performance

Discrimination performance of the Fine-Gray and STRATIFY-Falls model on external validation is summarised in Table 6.7, with the spread of C- and D-statistic values from individual practices displayed in scatter plots in Figure 6.19. The ordering of participants' predicted probabilities altered only slightly on recalibration; thus the discriminative abilities of the Fine-Gray model and the STRATIFY-Falls model were consistent with one another.

Table 6.7: Discrimination performance statistics for the Fine-Gray and STRATIFY-Falls models on external validation, pooled across GP practices

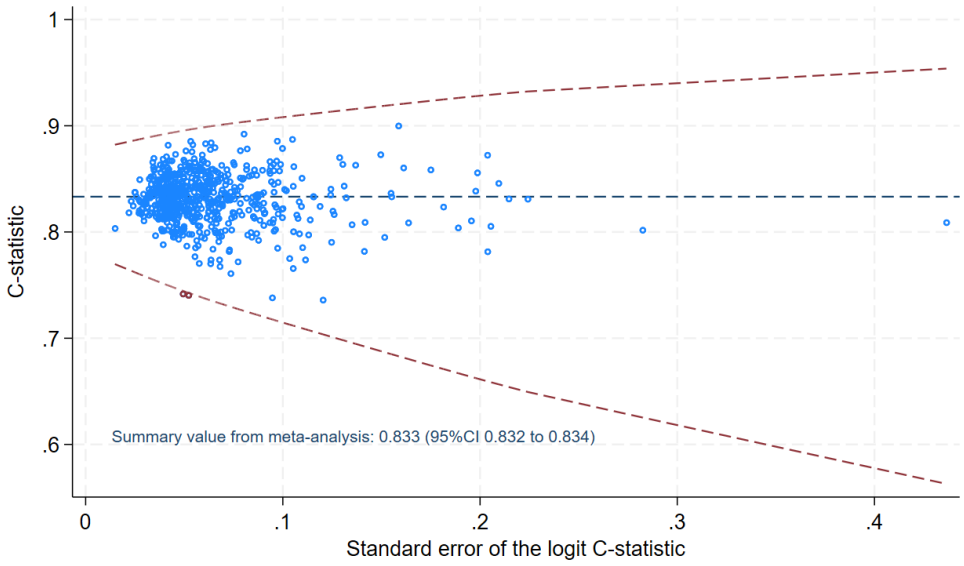
	Fine-Gray model	STRATIFY-Falls model
C-statistic		
Pooled effect size (95% CI)	0.833 (0.832 to 0.835)	0.833 (0.831 to 0.835)
Prediction interval, 95%	0.789 to 0.870	0.789 to 0.870
τ^2 (95% CI)	0.022 (0.019 to 0.025)	0.022 (0.019 to 0.025)
D-statistic		
Pooled effect size (95% CI)	1.643 (1.515 to 1.771)	1.597 (1.472 to 1.721)
Prediction interval, 95%	1.51 to 1.77	1.47 to 1.72
τ^2 (95% CI)	<0.0001 (<0.0001 to 0.0168)	<0.0001 (<0.0001 to 0.016)

When considering discrimination alone, small τ^2 values suggest that practices were reasonably homogeneous in terms of the C- and D-statistics. Similarly, 95% prediction intervals were relatively narrow, given the variation in case-mix that was expected, based on the different sizes and locations of the included practices. The analyses suggest that, were the Fine-Gray model to be used in a new GP practice (outside this external validation data), the C-statistic would be expected to

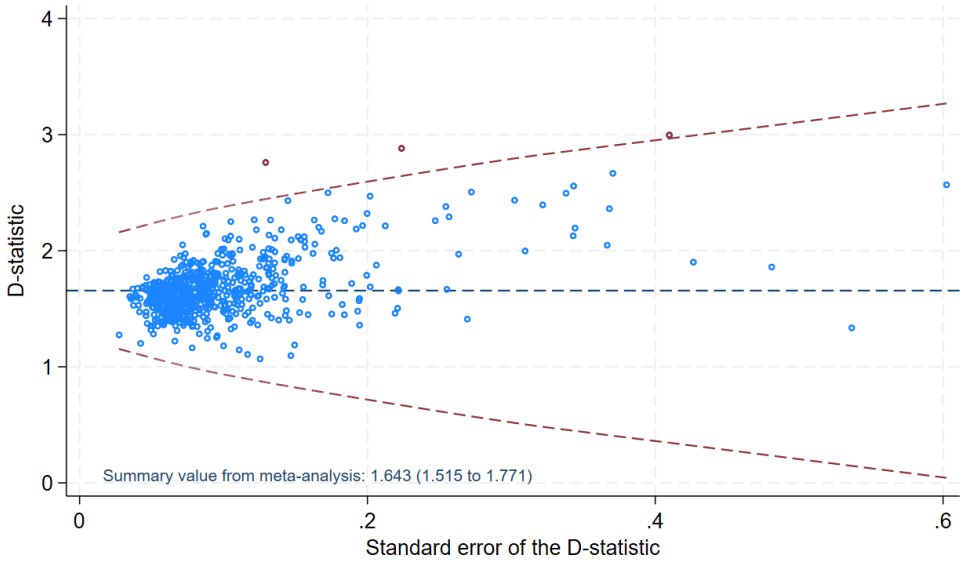
be somewhere between 0.789 and 0.870. The prediction interval for the D-statistic was similarly narrow, at 1.51 to 1.77.

Scatter plots in Figure 6.19 show that the majority of practices contained sufficient patients to give precise (standard error of logit C-statistic < 0.1, standard error of D-statistic < 0.2) estimates of discrimination, despite most falling short of the sample size requirement discussed in Section 6.3. Given this minimum sample size was calculated to target precise calibration, it seems precise estimates of discrimination are achievable with less stringent requirements.

Confidence intervals for both discrimination metrics were narrow for the Fine-Gray and STRATIFY-Falls models, with interval widths of only 0.003 for the C-statistic (0.831 to 0.835) and 0.256 for the D-statistic (1.515 to 1.771) of the Fine-Gray model. Therefore, despite some variation in discrimination performance across practices, pooled C- and D-statistics were both precisely estimated. This is in contrast to the example in Chapter 5, where heterogeneity in model performance across cohorts resulted in lower precision in the pooled performance estimates than in any single cohort.



(a) C-statistic



(b) D-statistic

Figure 6.19: Discrimination performance of the Fine-Gray model to predict falls, by their standard errors, across GP practices in the CRPD Aurum external validation data. Blue dashed line shows summary value, red dashed lines show the 95% prediction interval.

6.4.5 Overall model fit

Overall model fit was assessed using Royston and Sauerbrei's R^2_D for both the Fine-Gray model and the STRATIFY-Falls model. Table 6.8 summarises the model fit across GP practices.

Table 6.8: Summary of Royston and Sauerbrei's R^2_D across GP practices, for the Fine-Gray and STRATIFY-Falls models on external validation

	Fine-Gray model	STRATIFY-Falls model
R^2_D		
Range	21.3 to 91.4	21.6 to 91.4
Median (LQ-UQ)	39.9 (36.4 to 43.8)	38.6 (35.4 to 42.4)
Mean (SD)	40.8 (0.07)	39.4 (0.07)

Overall model fit by R^2_D for the Fine-Gray model varied considerably across GP practices, with values ranging from 21.3 to 91.4 on the 0-100 scale, suggesting exceptionally good performance in some practices, and relatively poor performance in others. This range was most notable among the smaller GP practices, where R^2_D would have been estimated with larger uncertainty, with estimates becoming more stable across larger the practices. Overall, half of practices gave an R^2_D value in the range 36.4 to 43.8.

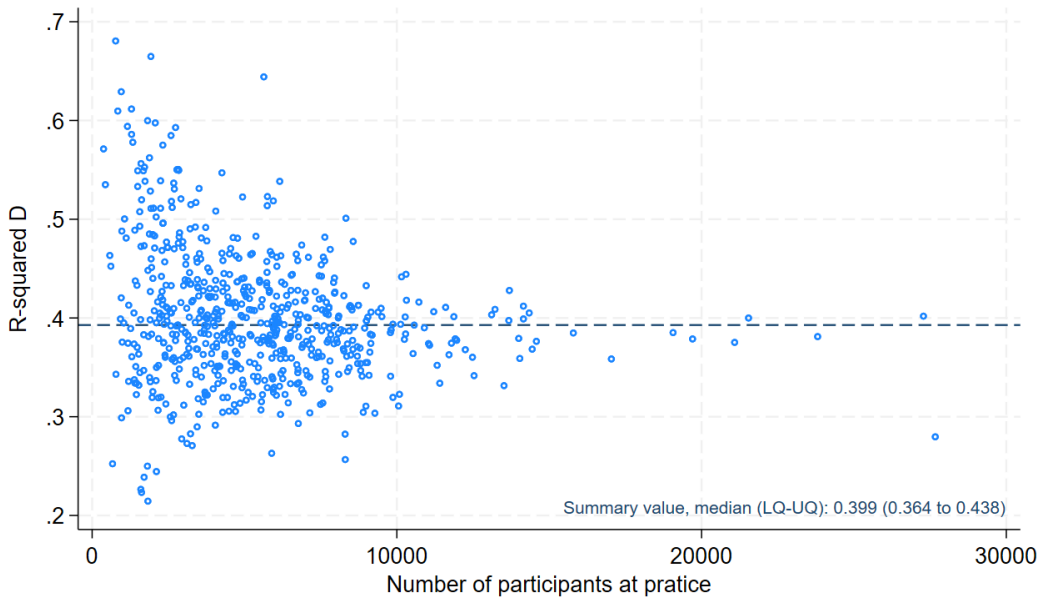


Figure 6.20: $R_D{}^2$ of the Fine-Gary model to predict falls, by GP practice size in the CRPD Aurum external validation data.

6.4.6 Net benefit

Decision curve analysis of the Fine-Gray and STRATIFY-Falls models indicated potential net benefit of using either model to predict 10-year falls risk, based on the pre-specified treatment decision threshold of 10% (Figure 6.21). Basing clinical decisions on either model with a 10% threshold probability yielded a benefit over the two alternative, model-blind strategies: altering care for all patients, for example by introducing falls prevention measures, which may include deprescribing of antihypertensives; and usual care for all patients (not introducing falls prevention measures or changing treatment.

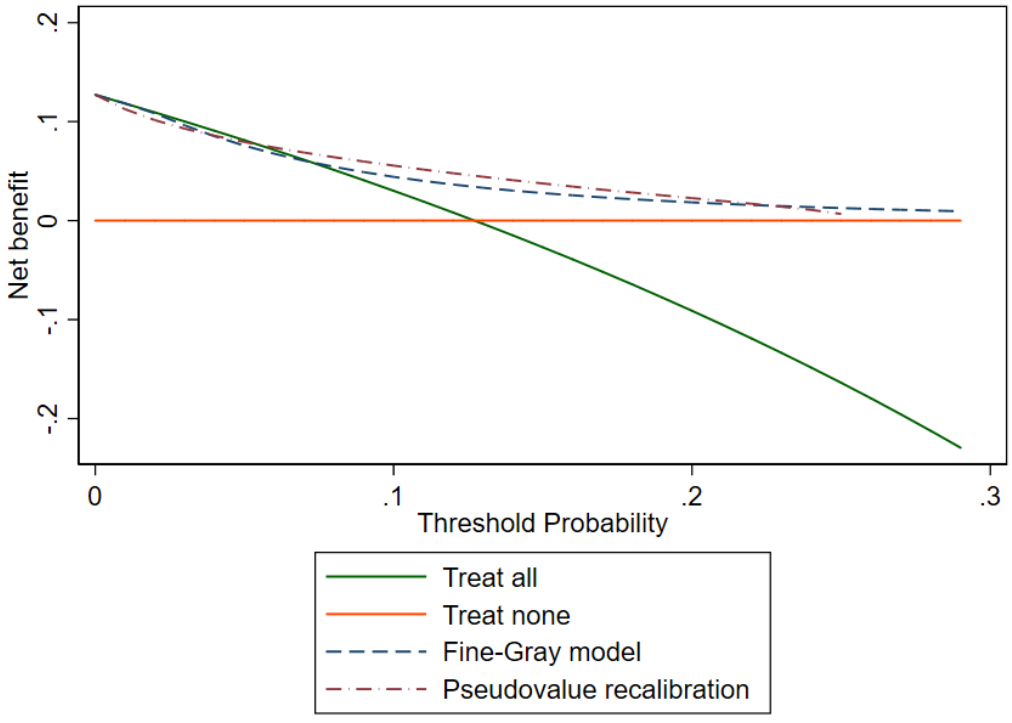


Figure 6.21: Decision curve analysis showing net benefit of using STRATIFY-Falls models across different threshold probabilities for assigning treatment

Benefit over other strategies was evident when using a treatment decision threshold of 7% or higher from the Fine-Gray model, or a treatment decision threshold of 6% or higher from the STRATIFY-Falls model. Curves were reasonably consistent with one another over the full range of threshold probabilities, though suggested that, on average over all GP practices (without accounting for clustering), the recalibration in the development data resulted in slightly improved net benefit for the STRATIFY-Falls model over the 0.1 to 0.2 range when assessed on external validation.

6.5 Discussion

6.5.1 Summary of key findings from this chapter

This chapter demonstrates the methods used in the external validation of a clinical prediction model for the risk of a serious fall within ten years of being indicated for antihypertensive treatment. Meta-analysis methods (as discussed in Chapter 5) were used to summarise model performance across populations, while assessing heterogeneity in model performance across GP practices, given the external validation cohort comprised routinely collected information from patients at 738 different GP practices.

Methods to calculate the appropriate sample size needed to accurately assess the predictive performance of this survival model were also demonstrated. Though the derivation of these methods does not form a part of this thesis, they form an extension to the calculations proposed in Chapter 4, and were developed and published as a part of a wider research team subsequent to that work [93]. When identifying the sample size value to test in subsequent iterations, two methods were trialled: a simple linear interpolation and a Newton's divided difference interpolation. Given the increased complexity in the Newton method, without a corresponding increase in efficiency, the linear interpolation approach was preferred for guiding this iterative approach to sample size calculation.

In this applied example, the sample size calculation suggested a minimum of 11,385 patients (approximately 991 serious fall events) were required to target precise estimates of the model's calibration performance. External validation data comprised information on patients from multiple different GP practices across the UK, many of which individually failed to meet the required

sample size for this external validation. Though the size of the external validation data overall surpassed this requirement by a considerable margin, only 14.5% of individual GP practices contained the required minimum number of falls events at 10 years.

Prediction intervals from meta-analyses across practices gave an indication of how well the models would be expected to perform in a new practice, outside the external validation population, helping to inform decisions on whether the model was suitable for implementation in practice. In this clinical example, the prediction intervals were relatively narrow across discrimination performance statistics, suggesting that the model's ability to discriminate between those with and without a serious fall would be consistent in a new practice from a similar population. Thus, the narrow 95% prediction intervals around pooled estimates for discrimination measures (C-statistic: 0.789 to 0.870, D-statistic: 1.51 to 1.77) gave reassurance that that the high level of discrimination would likely also be seen if the model were applied in a new setting.

In contrast, calibration performance was highly variable across practices, with wide prediction intervals for O/E in both the Fine-Gray (1.139 to 2.484) and STRATIFY-Falls (1.284 to 2.638) models. These intervals suggest that both forms of the model would likely under-predict the risk of serious falls if applied in a new but similar setting, though the extent of under-prediction is unclear. High heterogeneity was also seen in calibration curves across GP practices, with large amounts of variation in the level of under-prediction seen across sub-populations.

6.5.2 Strengths and limitations

The sample size calculation methods used in this chapter are not meant for use in a case with competing risks, and as such incorporates only the survival distribution for falls events (along with the censoring distribution) without allowing for the competing event of death. The data generating mechanism in the simulation process in Figure 6.2 would become more complicated if the competing risk were included, requiring additional information on both the distribution of times until the competing event, and on correlations between the risks of each event. While methods have been proposed for the simulation of data with competing risks [254], this alteration to the methods proposed by Riley et al 2021 [93] is beyond the scope of this thesis, though is an important consideration for future research.

In this model, drug treatments were defined as binary variables based on the presence of a prescription within the year before model application, without accounting for any changes to drugs during follow-up. This approach is known as "landmarking", with the treatment variables being considered only at a fixed point in time (the landmark), and the value of the predictor at the landmarking time being treated as a time-fixed covariate. Given the long time-frame for prediction, not allowing for the time-varying nature of treatment could have a substantial affect on the predicted risks obtained from the model. One alternative approach to account for important variations in prescriptions over time includes combining information across multiple landmarking times, by stacking landmark datasets and stratifying analyses by the times when treatment variables were reassessed (with corrections to the calculations of standard errors for repeated use of the same patients' data) [255]. However, this method would be highly computationally intensive, given the size of the model development and external validation datasets at just a single landmark (1,772,600 and 3,805,366 patients respectively).

Another option would have been to use multi-state modelling to account for time-varying treatments. In this approach, outcome probabilities are calculated based on transition rates between different "states", which describe the covariate position a patient occupies at any given time (for example, being prescribed an antihypertensive or not). Transitions between states may be two-way, where patients can move back and forth over time (such as starting and stopping different treatments); one-way, where there is no return to the prior state once the transition has occurred (diagnosis of a chronic comorbidity); or absorbing, where it is not possible to leave the state after transitioning to it (death, either related to a fall or as a competing event). A key drawback to the use of multi-state models in this example is that they would need to be formed under the Markov assumption: that transition rates depend only on the current state (current prescribed treatments at the point of prediction) and not the patient's history of state transitions. This assumption is unlikely to hold in practice, as the sequence of prescriptions in a patient's history is likely to be just as informative as their current prescription status, if not more so.

In the chapter, assessments of predictive performance were made across a range of GP practices, giving an opportunity for insight into the expected spread of performance across sub-populations with different case mix and outcome prevalence. The values of all performance measures varied considerably among smaller practices, with more consistency as practice sizes increased. This higher variation in small practices demonstrates both true variations in model performance and higher levels of uncertainty in the estimates derived in practices with small sample sizes. One likely source of the heterogeneity in model performance is differences in 10-year incidence of falls across the tested GP practices, which may have stemmed from differences in demographic composition, or different standards of care or availability of falls prevention programmes across different areas

of the UK.

Heterogeneity in baseline risk can lead to a lack of transportability of a model across different populations, resulting in systematic miscalibration. In such situations, model performance can often be improved by simple recalibration procedures [63]. Recalibration is the process of updating the model to tailor it to the population of interest and, in its simplest form, involves re-estimation of the model's intercept, or baseline risk term. The variation in the calibration performance of the STRATIFY-Falls models could have been substantially reduced if the CIF (in the Fine-Gray model), and the intercept term (STRATIFY-Falls model) were tailored to individual practices, to better represent the average falls risk in the given sub-population. It is worth noting, though, that these procedures would result in a new model for each practice, that had not been externally validated, thus the appropriateness of application in new individuals without further evaluation should be carefully considered.

Thus, there was potential for local recalibration of the STRATIFY-Falls model to specific sub-populations, to maximise calibration performance and the opportunity for patients to benefit from more accurate, tailored risk estimates. Though this approach to model updating was beyond the scope of the STRATIFY work programme, these methods were considered further and implemented in subsequent research into the risk of serious falls in a wider, frail population [99].

6.5.3 Conclusions and next steps

Big data from EHR offer great opportunities to more closely examine the performance of a clinical prediction model across populations. However, IPD meta-analysis techniques are key for

assessment of heterogeneity and to adequately account for differing predictive performance in the presented summaries. The use of a national GP database for external validation, provided the opportunity to assess heterogeneity across the different GP practices, where different model performance was expected. This analysis shows the power and importance of the appropriate consideration of heterogeneity in model performance, over-and-above what was shown in the previous chapter.

While the sample size in this applied example far exceeded the minimum requirement for the external validation of the STRATIFY-Falls model, the data comprised a combination of sub-populations, the majority of which did not meet the requirement in isolation. Heterogeneity in model performance across GP practices was high, in particular regarding the extent of miscalibration. Model updating would likely be required before these models could be considered for application in new settings, at a minimum involving recalibration of the baseline CIF and intercept terms to better account for baseline falls risk in specific sub-populations.

Thus, given populations within national settings vary considerably, consideration of model suitability is important not only on average, but within a variety of patient subgroups, such as those defined geographically through GP practice registration. Where such subgroups are represented in the validation data only in small numbers, with higher uncertainty in performance estimates, caution should be taken in the assumption that a model that performs well overall is suitable for application in the general population.

CHAPTER 7

Chapter 7: Discussion

7 Chapter 7: Discussion

7.1 Overview of thesis

Each year, thousands of prediction models are published in the medical literature, cementing prediction modelling as a core component of research in healthcare. Following a recent surge in interest in the topic during the COVID-19 pandemic [1], it is clear that the general quality of prediction research remains sub-standard, despite the publication of the PROGRESS partnership recommendations over 10 years ago [3, 4, 5, 6].

When appropriately accurate and precise, prediction models hold great potential for improving clinical decision making [186, 256]. Clearly, prediction modelling research has the potential to offer considerable clinical benefit, through the opportunity to enhance shared decision making, facilitate clinician-patient communication, and inform personalised care. Equally, poorly performing models have the capacity to cause considerable harm to individuals, for example if use of the model leads to an invasive treatment being used unnecessarily, or to advantageous treatments being wrongly withheld. Thus, the use of inappropriate or sub-optimal statistical methods is a real concern, leading to a gap between the potential and realised impact of research on patient outcomes.

Building on the PROGRESS recommendation to improve choice and implementation of statistical methods within prognosis research, this thesis has centred around the demonstration and development of leading methods for research involving clinical prediction models [5]. In particular, the research included here has revealed current practice and possible impacts of dichotomisation of continuous outcomes; demonstrated the importance of sample size consideration in the external validation of prediction models with continuous outcomes; and shown the impact of heterogeneity

in model performance on the precision of predictive performance estimates in clustered data. Furthermore, the preceding chapters have shown the implementation of robust statistical methods in prediction-based projects in clinical areas including musculoskeletal pain [91], pregnancy complications [109], and hypertension [97].

This final chapter gives a short summary of the chapters included in this thesis, followed by outlining the publications resulting from this body of work. Areas of contribution to the prediction modelling literature are discussed, for both methodological approaches and clinical applications, followed by suggestions of further projects that has been sparked by the analyses reported here. It concludes with a summary of strengths and limitations of this thesis as a whole, along with recommendations for future research involving prediction modelling in healthcare.

7.1.1 Summary of thesis chapters

Chapter 1 introduced key statistical concepts for the development and validation of clinical prediction models, including the different modelling methods often employed in the development of prediction models with outcomes of different types, and important considerations for evaluating the performance of a prediction model.

Chapter 2 described a review of recently published models for predicting Fetal Growth Restriction (FGR), demonstrating how the continuous outcome of birthweight is commonly dichotomised in practice, and how there is often a lack of any justification offered by authors for their choice of outcome treatment.

Chapter 3 discussed the development and validation (both internal and external) of prediction models for a patient's anticipated pain intensity outcomes following a primary care consultation for NLBP. Of particular interest was the comparison models' predictive performance when modelling pain intensity as a continuous versus a binary outcome variable, in a clinical scenario where the binary outcome had no notable clinical benefit above a continuous outcome at implementation.

This chapter further demonstrated a method proposal for how a model for a continuous outcome (such as pain intensity) could be used to make subsequent predictions of the probability of a dichotomised outcome, comparing these probabilities to those gained from a logistic model, where the binary outcome was modelled directly.

Chapter 4 proposed closed-form solutions to calculate the minimum sample size required when externally validating a clinical prediction model with a continuous outcome, to ensure sufficient precision around key performance statistics. This proposed method was demonstrated, walking through a calculation for the sample size required to ensure precise estimates in the external validation of a model to predict fat-free mass in children and adolescents.

Chapter 5 described the external validation of published prognostic models identified in Chapter 2, in particular evaluating one model to predict continuous birthweight as a proxy for the risk of delivering a growth restricted baby. External validation was conducted using IPD meta-analysis methods to combine model performance estimates across multiple cohorts, demonstrating heterogeneity in the accuracy of this birthweight prediction model across different populations.

Chapter 6 discussed the use of routinely collected data to externally validate a model to predict the risk of a serious fall (resulting in hospitalisation or death), in those eligible for antihypertensive treatment in a primary care population. IPD meta-analysis methods were further employed to show the variability in model performance across GP practices, demonstrating how precise overall estimates of model performance masked highly variable performance in settings with differing case-mixes.

7.1.2 Publications arising from this thesis

This thesis has lead to a number of publications, as a direct output of the research discussed here as well as from related, follow-on projects. These publications are summarised in Table 7.1.

In particular, first-author journal articles relating to the clinical applications discussed in Chapters 3 [91] and 6 [97], and from the sample size methodology discussed in Chapter 4 [92], have been published in *Physical Therapy*, *BMJ*, and *Statistics in Medicine*, respectively. Copies of these publications have been included as appendices to the relevant chapters.

Notable publications from extensions of the methodology work discussed in this thesis include co-authorship on sample size guidance for the external validation of both binary [95, 94] and time-to-event outcome models [93], an article discussing instability in prediction models developed on different sample sizes [182] and a three-part *BMJ* series on *Evaluation of Clinical Prediction Models*, including a paper on sample size recommendations [257, 258, 96].

Table 7.1: Summary of publications arising from, or related to, chapters in this thesis

Output	Title	Journal	Year	
Chapter 3				
Direct	Development and external validation of individualized prediction models for pain intensity outcomes in patients with neck pain, low back pain, or both in primary care settings	*Physical Therapy*	2023	[91]
Related	Musculoskeletal Health and Work: Development and Internal-External Cross-Validation of a Model to Predict Risk of Work Absence and Presenteeism in People Seeking Primary Healthcare	*J Occup Rehabil*	2024	[259]
Chapter 4				
Direct	Minimum sample size for external validation of a clinical prediction model with a continuous outcome	*Stat Med*	2020	[92]
Related	Minimum sample size for external validation of a clinical prediction model with a binary outcome	*Stat Med*	2021	[94]
	External validation of clinical prediction models: simulation-based sample size calculations were more reliable than rules-of-thumb	*JCE*	2021	[95]
	Minimum sample size calculations for external validation of a clinical prediction model with a time-to-event outcome	*Stat Med*	2022	[93]
	Clinical prediction models and the multiverse of madness	*BMC Medicine*	2023	[182]
	Evaluation of clinical prediction models (part 3): calculating the sample size required for an external validation study	*BMJ*	2024	[96]

Output	Title	Journal	Year	
Chapter 5				
Direct	Prediction of fetal growth restriction and complications: Individual Participant Data (IPD) meta-analysis with Decision Curve Analysis and Economic Evaluation International Prediction of Complications in Pregnancy: Fetal Growth restriction (IPPIC-FGR)	*HTA* [accepted]	-	
Related	Predicting birthweight: development and validation of prognostic model using individual participant data meta-analysis	*BMJ Med* [accepted]	-	
Chapter 6				
Direct	Development and external validation of a risk prediction model for falls in patients with an indication for antihypertensive treatment: retrospective cohort study	*BMJ*	2022	[97]
Related	Association between antihypertensive treatment and adverse events: systematic review and meta-analysis	*BMJ*	2021	[13]
	Predicting the risk of acute kidney injury: Derivation and validation of STRATIFY-AKI	*BJGP*	2022	[98]
	Development and external validation of the eFalls tool: a multivariable prediction model for the risk of ED attendance or hospitalisation with a fall or fracture in older adults	*Age and Aging*	2024	[99]

7.2 Areas of contribution to prediction modelling research

This thesis has contributed to both applied and methodological research in the field of prediction modelling. The findings and recommendations for each research contribution are presented in the discussion section of individual chapters, and the key methodological and clinical areas of contribution are now considered once more below.

7.2.1 Methodological approaches

Development of prediction models with a continuous outcome

The literature review described in Chapter 2 demonstrated current modelling preferences for just one clinical scenario where the prediction of a continuous outcome is important to inform patient care. In reality, continuous outcomes are relevant across the whole spectrum of clinical areas, with this thesis demonstrating examples in maternal, musculoskeletal, and metabolic health (birthweight, pain intensity score, and fat-free mass, respectively). Prediction models that fail to keep continuous outcomes on their continuous scale may suffer from loss of information and reduced power to detect predictor effects, thus avoidance of unnecessary dichotomisation is crucial [106, 102, 103].

The method proposed in Chapter 3, for generating predicted probabilities from the output from a linear regression model, could allow researchers to make maximum use of their development data at the modelling stage, while still allowing the dichotomisation of a continuous outcome to facilitate clinical decision making, if desired. This proposal could be beneficial to research in any clinical setting where outcomes of interest are on (or are formed from) a continuous scale. The proposed calculation can be applied post-modelling, meaning that there is no need to dichotomise a continuous outcome prior to prediction. Reserving the choice of cut-point until after the model

development means that continuous outcome models could then be applicable in a much wider range of populations. For example, where cultural differences contribute to different conventions around the interpretation of pain intensity, a different cut-point in pain score might be more relevant in identifying those in a high amount of pain.

External validation of prediction models with a continuous outcome

Conclusions on whether a prediction model is suitable to help inform clinical practice are often drawn on the basis of the model's performance in data from just a sample of the population where the model might be used. Clearly, accurate estimation of the model's performance here is vital, as imprecise estimation from assessment in only small samples of patients could lead to inappropriate decisions on whether or not to use the model in new patients. However, sample sizes used for external validation are often too small to provide reliable conclusions about a prediction model's performance [23, 20, 74].

Chapter 4 introduced new criteria to inform the minimum sample size needed for external validation of a clinical prediction model with a continuous outcome. The aim of these proposed criteria was to ensure their external validation gains sufficiently precise estimates of key performance measures, to ensure researchers can have confidence in the model's predictive ability. The proposed sample size calculations are closed-form, so are quick and easy to implement, and could be used by researchers to help improve the quality of external validation studies across all clinical areas. The paper associated with this work [92] has been cited 69 times as of 4th June 2024, and the work has been embedded in the *pmvalsampsize* package in R and Stata, allowing users to easily apply the method [260]. Calculation methods are also easily adapted to estimate the expected precision

in estimates based on a fixed sample size for external validation, allowing researchers to gauge expected precision of estimates prior to analysis, for example to inform grant applications.

External validation of prediction models with a survival outcome

Chapter 6 demonstrated the potential of pseudo-values as a complete, continuous representation of the risk of a survival outcome, where the observed data is incomplete due to right censoring or competing risks [251, 245]. The value of pseudo-values in the assessment of survival data is known, and their benefit in terms of improved plotting over traditional survival analysis approaches is acknowledged [242, 244, 248]. The potential of pseudo-values in the validation of prediction model, in particular in the assessment and visualisation of calibration performance, are not yet widely established.

Traditional approaches to the plotting of observed survival have involved sub-grouping data [261], which is inefficient and looses information at the individual level [101]. Alternative calibration plots for survival outcomes, based on the pseudo-values for observed risk (as demonstrated in Chapter 6 of this thesis, and the associated publications), could be used in the future to allow smoothed calibration curves across all individuals, giving a more complete view of a model's calibration performance.

External validation of prediction models in clustered data

Chapters 5 and 6 of this thesis have demonstrated the potential for IPD meta-analysis methods to increase the potential for insight into model performance, and in particular the heterogeneity of performance, across different populations[67, 63]. In cases where the available samples for external

validation are small, combinations of records from different sources can vastly boost the available sample size [63]. In such cases, however, it is possible that the increased sample size will not lead to increased precision in summary estimates of predictive performance, due to heterogeneity in model performance across populations. This imprecision in pooled estimates should not be seen as a limitation of the external validation, and instead be recognised as an artefact of differing case-mixes and outcome prevalences within clusters.

When assessing model performance in large populations, for example at a national level, performance summaries across the whole population might hide important variations in the predictive ability of the model across distinct subpopulations. Heterogeneity in model performance across sub-groups, for example across geographical regions defined by GP practice, should be considered and appropriately visualised. Where a model is intended to be applied in practice, adequate and precise estimates of performance *on average* should not be assumed to mean that the model is equally applicable in all sub-populations. Where analysis suggests heterogeneity in model performance across clusters, tailoring of the model to the local setting (for example, through re-estimation of the baseline risk or value of the outcome) should be considered [262, 59].

7.2.2 Clinical applications

While the research contributing to methodological chapters could be beneficial to any clinical area, the particular applied research areas benefiting from this thesis include:

Fetal growth restriction

Chapter 2 revealed a lack of prediction models currently available in the literature for predicting a clinically acceptable definition of FGR. Most binary outcome models simply dichotomised

birthweight, defining their outcome based on either the 10th of 5th percentile as a cut-point. Not only does this definition fail to capture the pregnancy and birth complications indicative of FGR, but is a highly inefficient way to model a continuous outcome [106, 102, 103]. Prediction of birthweight on a continuous scale was found to be rare in the literature, as was justification of authors' choice on how to treat this outcome variable in their modelling.

Where possible, performance of published prediction models on external validation was assessed and compared using IPD from cohorts within the IPPIC collaboration data collection [109]. Of the models identified, none adequately met the binary FGR definition provided by clinical collaborators. Thus, this chapter lends support to calls for better standardisation of the definition of FGR used in a research setting, to more better reflect true growth restriction rather than just normal levels of smallness in a healthy baby.

Of the models predicting birthweight on its continuous scale, 11 were excluded from the external validation analyses due to containing variables (or combinations thereof) that were not present in the available IPD. Given the data included in the IPPIC collaboration data collection was extensive, containing information on around 3 million pregnancies in total from 14 UK and 66 international datasets, it was disappointing that so many of the reported models included predictors were not represented. This is suggestive of models including variables that are not routinely recorded, which would likely negatively impact their usability in practice. Though factors such as biparietal diameter, or birthweight from previous pregnancies, might contribute strongly to predictions on model development, further consideration of more widely measured clinical predictors could be beneficial, especially if models are to be used in resource poor settings.

In the external validation discussed in Chapter 5, the birthweight prediction model was well calibrated on average. However, large amounts of individual-level variation in birthweight were not adequately represented by the model predictions. Thus, it was concluded that the current literature was lacking in an acceptable model for the prediction of FGR, either directly or through use of birthweight as a proxy. Future research, therefore, aimed to build on this model to develop and validate new prediction models for identifying the risk of delivering a growth restricted baby. An updated model for the prediction of birthweight was also considered, in an attempt to improve calibration on the individual level.

Neck and/or low back pain (NLBP)

Chapter 3 demonstrated an application of model development and validation methods for the continuous outcome of future pain intensity, in a NLBP population. This applied example aimed to contribute to improved communication and treatment matching for patients consulting with NLBP, through predictions of pain on both the continuous and a dichotomised scale [91]. This research formed a part of an international body of work, aiming to develop varied digital health technologies to support decision making for those consulting with NLBP [164].

Visualisations based on the prediction models developed within this chapter were incorporated into an online demonstrator, to allow clinicians to see patient predictions for hypothetical individuals and to give their feedback on their usability. Future research will include reporting on the acceptability of the prediction visualisations to both clinicians and patients, to inform improvements to the layout and usability of this decision support tool. At present, however, the prediction models themselves are not yet adequate for clinical use at the point of consultation,

and will not be used in patients. Further research is planned to update and hopefully improve the prediction models through inclusion of non-modifiable risk factors. External validation in additional, larger datasets would also reduce uncertainty in predictive performance estimates and to inform the need for updating to local settings, for example through tailored recalibration [59].

Adverse events following anti-hypertensive medication

Chapter 6 discussed the external validation of a model for the prediction of hospitalisation or death associated with a serious fall [97]. The clinical example formed a part of a larger body of work, the STRAtifying Treatments In the multi-morbid Frail elderly (STRATIFY) project, aiming to develop and externally validate prediction models for adverse events associated with antihypertensive medication in frail older adults [13, 98]. The overall aim of this project was to allow targeted treatment in those with an indication for antihypertensive medication, tailoring prescribed medications to those who are least likely to experience harm from associated adverse events [6].

The STRATIFY-Falls model could be used in future to aid primary care doctors in the assessment of falls risk, using data that is routinely available in electronic health records [263, 264]. For individuals with a high risk of a serious fall but a low risk of cardiovascular disease [265], clinicians might consider whether new or continued antihypertensive treatment is still appropriate. Though miscalibration was noted in higher risk individuals the external validation, under-prediction in this clinical setting was deemed to be of minimal concern. Given the known benefits of antihypertensive medication, clinical collaborators on this project were clear that, were the model to be used to inform treatment changes, doctors would need to be confident that the true risk was at least at

the indicated level, if not higher.

Further research is planned to explore the appropriate treatment decision thresholds to maximise the model's clinical utility and cost effectiveness, and to examine the practicality of recalibration to local settings. Not only could tailoring the model to local settings substantially reduce variation in the calibration performance across populations, but these measures may also eliminate the under-prediction of falls risk seen on external validation [63, 262]. Thus, there is potential for local recalibration to further improve calibration performance in this case, and for future patients to directly benefit from the consideration of more accurate risk estimates to inform their treatment options.

7.3 Further research

Building on the research discussed in this thesis, several areas have been identified as having potential for important further research. Key areas for extensions of both the methodological and applied projects are summarised below.

7.3.1 Methodological approaches

Treatment of continuous outcomes in prediction modelling

The methods review in Chapter 2 aimed to investigate how continuous outcomes are being modelled in the development of prediction models, and what justification is given by researchers for any dichotomisation of outcome variables in practice. In particular, this review focused on the prediction of continuous birthweight as a proxy for binary FGR, thus conclusions on the regularly of dichotomisation and the lack of justification are somewhat restricted to this clinical area.

The current review is restricted in scope, limiting the ability to make conclusions about the treatment of continuous outcomes in other contexts. An extension of this review into a wider variety of clinical areas could provide further understanding of motivations behind dichotomisation decisions across different medical specialities, and would allow scope for conclusions to be more generally applicable. Furthermore, an update into more varied clinical areas may attract attention from a wider audience, with potential for increased impact across the prediction modelling literature in the future.

With regards to dichotomisation of predictions post-modelling, as proposed in Chapter 3, the examples shown in this thesis were generally straightforward in terms of their modelling approaches and outcome distributions. Thus, findings regarding the suitability of post-analysis dichotomisation in reducing model instability may not extend to scenarios involving more complicated analysis methods. Future work will include testing the methods more thoroughly across a range of simulated scenarios, to assess stability [182, 181] and performance of clinical prediction models with dichotomised continuous outcomes, and to identify situations where the suggested transformation to the probability scale might not be appropriate.

Furthermore, the proposed probability generation methods are highly reliant on the underlying assumptions of linear regression [30], and do not allow for alternative methods of predicting outcomes on a continuous scale. In particular, the method stems from the linear regression requirement for normally distributed residuals and for constant error terms across different observed outcome values [30, 191]. Further research is also needed to investigate the suitability of these methods to generate probabilities from a continuous outcome prediction where data do

not meet the assumptions, and where linear regression models are not necessarily appropriate. Understanding how to link the approach to conformal prediction methods would also be important, where the uncertainty of predictions is summarised using distribution-free approaches [266].

Extension of sample size calculations for external validation

Important extensions of the work in Chapter 4 include assessment of the necessary sample size for external validations of prediction models with non-continuous outcomes, building on the work of others in this area [76, 172]. Such calculations were investigated as a part of a wider research team. As mentioned in Chapter 4, subsequently published simulation-based extensions to binary [95] and time-to-event [93] outcome settings, and approximate closed-form solutions for a binary outcome model [94], are now available to help researchers externally validating prediction models with these outcome types. Future work might also consider an extension of the proposed criteria to include precise estimation of calibration curves [79, 188], which was not considered in Chapter 4, nor in subsequent publications.

As with many new statistical methods, ease of implementation can be a barrier to use of the approach in practice. User-friendly software to simplify the application of such approaches therefore facilitates their implementation. Thus, to encouraging adoption of the sample size calculation proposed in Chapter 4, software that is easy to use and freely available is key, and was intended as a follow-up to this chapter. These methods were incorporated instead into the freely available *pmvalsampsize* package in Stata and R [260, 267], along with methods for the calculation of the minimum sample size required to externally validate a prediction model with a binary outcome mentioned above.

In recent years, there have been an increasing number of published prediction models being attributed to machine learning methods though these are often of poor quality [20, 74]. Sample size recommendations for external validation must be suitable for use with models of this type, in addition to more traditional regression-based models [23]. Further extensions to sample size criteria are planned to consider precision in estimates of cut-off-based performance metrics more commonly used in machine learning research (including precision, recall, and F1 score), and whether the amount of data required to precisely estimate these is higher than that required for precise estimation of, for example, calibration.

7.3.2 Clinical applications

Prediction of FGR with complications in pregnancy

Following the literature review in Chapter 2 and external validation of existing models discussed in Chapter 5, it was concluded that the current literature was lacking in an acceptable model for the prediction of FGR, either directly or through use of birthweight as a proxy. The next steps of this research project, therefore, were to develop and externally validate a new prediction model for identifying the risk of delivering a growth restricted baby (as defined in the previous chapters), using data from the IPPIC collaboration data collection.

The planned model development combines IPD meta-analysis methods with multiple imputation, variable selection, and assessment of non-linear predictor-outcome relationships, while an internal-external cross-validation (IECV) approach was used for model validation [67, 109, 56]. The aim of this new model development was to allow predictions conditional on some assumed

gestational age at delivery, allowing assessment of anticipated FGR risk or continuous birthweight if a baby were to be born at various different stages of gestation. Although the observed gestational age at delivery would not be available at the moment of prediction, producing the models in this way allowed for a range of potential gestational ages at delivery to be assessed for each pregnancy, with plots of predictions against gestational age to give a more complete assessment of risk over time.

Before any birthweight or FGR prediction model is deemed suitable for use to identify high risk pregnancies that would benefit from increased monitoring, further assessment will be required to further consider the clinical utility and cost effectiveness of the model, and the implications of its use in practice.

Prediction of adverse events in antihypertensive-eligible populations As mentioned in Chapter 6, the prediction of serious falls risk was just one outcome of interest within a larger body of work, developing and external validating models to predict hospitalisation or death within 10 years of antihypertensive treatment being indicated. The next steps of this project involve further research to develop and externally validate models for other adverse events known to be associated with antihypertensive medications [13].

Thus far, the development and external validation of a model to predict acute kidney injury have been published in the British Journal of General Practice [98]. Further models concerning the risk of fractures, hypotension, and syncope have each been developed and externally validated in UK-based EHR, with a paper reporting these three models currently being under review. The external validation of two further models to predict the risks of different electrolyte abnormalities

is currently underway, with the ability of models to predict risks of both hyperkalaemia and hyponatraemia being assessed across GP practices.

7.4 Recommendations for future research

The work included in this thesis has lead to a number of recommendations that could contribute to improved quality in research involving prediction modelling in a healthcare setting. In particular, these relate to the prediction of a continuous outcomes and to methods for external validation, as discussed below.

- As with the dichotomisation of continuous predictors, the dichotomisation of continuous outcomes prior to modelling is not recommended. Continuous outcomes are better modelled on their continuous scale, to allow predictions of continuous values that retain a more complete picture for informing decisions. In the case of cut-points being desired for clinical decision making, dichotomisation after modelling on the continuous scale would be preferable.

- Accurate estimation of the model's performance on external validation is vital, as imprecise estimation could wrongly influence decisions on whether or not to use the model in new patients. Assessments of whether a prediction model is suitable to inform clinical practice should be based on validation in data that is large enough to precisely estimate all key measures of predictive performance.

- Small samples for validation may not be suitable in isolation, but are highly valuable

if analysed along with other datasets in an IPD meta-analysis of model performance, though heterogeneity in the predictors recorded across studies may hamper intended IPD meta-analyses by limiting which models can be externally validated. Where researchers have insufficient data to accurately assess model performance, they should consider contacting other researchers in the same clinical area to identify additional data resources that could contribute to a more thorough model evaluation. Researchers should also embrace opportunities to contribute their own data to future IPD meta-analyses, by making their own datasets available to others.

- When externally validating a prediction model across different settings, researchers should examine and embrace the heterogeneity in model performance. Clinical prediction models should be expected to perform differently in different populations and among different subgroups, thus heterogeneous performance should not be considered a drawback of the model or of the validation analyses. In particular, researchers should ensure that they thoroughly investigate differential model performance, especially where differences might increase health inequalities. Where a model is intended to be used in a variety of individuals, adequate and precise estimates of performance *on average* should not be assumed to mean that the model is equally applicable in all sub-populations.

7.5 Concluding remarks

Building on previous work in a range of clinical settings, this thesis has explored opportunities to improve the quality of research in the prediction modelling field, in particular relating to the modelling of continuous outcomes and the external validation of prediction models in small or clustered data. Many challenges remain in the field of clinical prediction modelling in healthcare,

and methodologists will continue to play a critical role in addressing them and educating better practice going forwards.

References

[1] L. Wynants, B. Van Calster, M. M. J. Bonten, G. S. Collins, T. P. A. Debray, M. De Vos, M. C. Haller, G. Heinze, K. G. M. Moons, R. D. Riley, E. Schuit, L. J. M. Smits, K. I. E. Snell, E. W. Steyerberg, C. Wallisch, and M. van Smeden, "Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal," *BMJ*, vol. 369, p. m1328, 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/32265220

[2] M. van Smeden, J. B. Reitsma, R. D. Riley, G. S. Collins, and K. G. Moons, "Clinical prediction models: diagnosis versus prognosis," *J Clin Epidemiol*, vol. 132, pp. 142–145, 2021.

[3] H. Hemingway, P. Croft, P. Perel, J. Hayden, K. Abrams, A. Timmis, A. Briggs, R. Udumyan, K. Moons, E. Steyerberg, I. Roberts, S. Schroter, D. Altman, and R. Riley, "Prognosis research strategy (progress) 1: A framework for researching clinical outcomes," *British Medical Journal*, vol. 346, 2013. [Online]. Available: Go to ISI://000314806700016

[4] R. Riley, J. Hayden, E. Steyerberg, K. Moons, K. Abrams, P. Kyzas, N. Malats, A. Briggs, S. Schroter, D. Altman, and H. Hemingway, "Prognosis research strategy (progress) 2: Prognostic factor research," *Plos Medicine*, vol. 10, no. 2, 2013. [Online]. Available: Go to ISI://000315592800002

[5] E. Steyerberg, K. Moons, D. van der Windt, J. Hayden, P. Perel, S. Schroter, R. Riley, H. Hemingway, and D. Altman, "Prognosis research strategy (progress) 3: Prognostic model research," *Plos Medicine*, vol. 10, no. 2, 2013. [Online]. Available: Go to ISI://000315592800003

[6] A. Hingorani, D. van der Windt, R. Riley, K. Abrams, K. Moons, E. Steyerberg, S. Schroter, W. Sauerbrei, D. Altman, and H. Hemingway, "Prognosis research strategy (progress) 4: Stratified medicine research," *British Medical Journal*, vol. 346, 2013. [Online]. Available: Go to ISI://000314806700017

[7] H. M. Krumholz, "Outcomes research: generating evidence for best practice and policies," *Circulation*, vol. 118, no. 3, pp. 309–18, 2008.

[8] C. Allemani, H. K. Weir, H. Carreira, R. Harewood, D. Spika, X. S. Wang, F. Bannon, J. V. Ahn, C. J. Johnson, A. Bonaventure, R. Marcos-Gragera, C. Stiller, G. Azevedo e Silva, W. Q. Chen, O. J. Ogunbiyi, B. Rachet, M. J. Soeberg, H. You, T. Matsuda, M. Bielska-Lasota, H. Storm, T. C. Tucker, and M. P. Coleman, "Global surveillance of cancer survival 1995-2009: analysis of individual data for 25,676,887 patients from 279 population-based registries in 67 countries (concord-2)," *Lancet*, vol. 385, no. 9972, pp. 977–1010, 2015.

[9] A. Wu, L. March, X. Zheng, J. Huang, X. Wang, J. Zhao, F. M. Blyth, E. Smith, R. Buchbinder, and D. Hoy, "Global low back pain prevalence and years lived with disability from 1990 to 2017: estimates from the global burden of disease study 2017," *Annals of Translational Medicine*, vol. 8, no. 6, p. 299, 2020. [Online]. Available: https://atm.amegroups.com/article/view/38037

[10] R. Riley, W. Sauerbrei, and D. Altman, "Prognostic markers in cancer: the evolution of evidence from single studies to meta-analysis, and beyond," *British Journal of Cancer*, vol. 100, no. 8, pp. 1219–1229, 2009. [Online]. Available: WOS:000265575000002

[11] P. Campbell, N. E. Foster, E. Thomas, and K. M. Dunn, "Prognostic indicators of low back pain in primary care: five-year prospective study," *J Pain*, vol. 14, no. 8, pp. 873–83, 2013.

[12] S. G. Parker, S. Mallett, L. Quinn, C. P. J. Wood, R. W. Boulton, S. Jamshaid, M. Erotocritou, S. Gowda, W. Collier, A. A. O. Plumb, A. C. J. Windsor, L. Archer, and S. Halligan, "Identifying predictors of ventral hernia recurrence: systematic review and meta-analysis," *BJS Open*, vol. 5, no. 2, 2021.

[13] A. Albasri, M. Hattle, C. Koshiaris, A. Dunnigan, B. Paxton, S. Fox, M. Smith, L. Archer, B. Levis, R. Payne, R. Riley, N. Roberts, K. Snell, S. Lay-Flurrie, J. Usher-Smith, R. Stevens, F. Hobbs, R. McManus, J. Sheppard, and S. investigators, "Association between antihypertensive treatment and adverse events: systematic review and meta-analysis." *BMJ*, vol. 372:n189, 2021.

[14] A. K. Clift, C. A. C. Coupland, R. H. Keogh, K. Diaz-Ordaz, E. Williamson, E. M. Harrison, A. Hayward, H. Hemingway, P. Horby, N. Mehta, J. Benger, K. Khunti, D. Spiegelhalter, A. Sheikh, J. Valabhji, R. A. Lyons, J. Robson, M. G. Semple, F. Kee, P. Johnson, S. Jebb, T. Williams, and J. Hippisley-Cox, "Living risk prediction algorithm (qcovid) for risk of hospital admission and mortality from coronavirus 19 in adults: national derivation and validation cohort study," *BMJ*, vol. 371, p. m3731, 2020. [Online]. Available: https://www.bmj.com/content/bmj/371/bmj.m3731.full.pdf

[15] N. H. S. Digital, "Covid-19 population risk assessment," 2023. [Online]. Available: https://digital.nhs.uk/services/coronavirus-risk-assessment/population

[16] K. Moons, P. Royston, Y. Vergouwe, D. Grobbee, and D. Altman, "Prognosis and prognostic research: what, why, and how?" *British Medical Journal*, vol. 338, 2009. [Online]. Available: WOS:000264056000016

[17] M. T. Hudda, M. S. Fewtrell, D. Haroun, S. Lum, J. E. Williams, J. C. K. Wells, R. D. Riley, C. G. Owen, D. G. Cook, A. R. Rudnicka, P. H. Whincup, and C. M. Nightingale,

"Development and validation of a prediction model for fat mass in children and adolescents: meta-analysis using individual participant data," *BMJ*, vol. 366, p. l4293, 2019. [Online]. Available: https://www.bmj.com/content/bmj/366/bmj.l4293.full.pdf

[18] A. S. Moriarty, N. Meader, K. I. E. Snell, R. D. Riley, L. W. Paton, S. Dawson, J. Hendon, C. A. Chew-Graham, S. Gilbody, R. Churchill, R. S. Phillips, S. Ali, and D. McMillan, "Predicting relapse or recurrence of depression: systematic review of prognostic models," *Br J Psychiatry*, vol. 221, no. 2, pp. 448–458, 2022.

[19] E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, and B. Van Calster, "A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models," *J Clin Epidemiol*, vol. 110, pp. 12–22, 2019.

[20] P. Dhiman, J. Ma, C. L. Andaur Navarro, B. Speich, G. Bullock, J. A. A. Damen, L. Hooft, S. Kirtley, R. D. Riley, B. Van Calster, K. G. M. Moons, and G. S. Collins, "Methodological conduct of prognostic prediction models developed using machine learning in oncology: a systematic review," *BMC Medical Research Methodology*, vol. 22, no. 1, p. 101, 2022. [Online]. Available: https://doi.org/10.1186/s12874-022-01577-x

[21] D. M. Kent, J. K. Paulus, D. van Klaveren, R. D'Agostino, S. Goodman, R. Hayward, J. P. A. Ioannidis, B. Patrick-Lake, S. Morton, M. Pencina, G. Raman, J. S. Ross, H. P. Selker, R. Varadhan, A. Vickers, J. B. Wong, and E. W. Steyerberg, "The predictive approaches to treatment effect heterogeneity (path) statement," *Ann Intern Med*, vol. 172, no. 1, pp. 35–45, 2020.

[22] R. D. Riley, T. P. Debray, D. Fisher, M. Hattle, N. Marlin, J. Hoogland, F. Gueyffier, J. A. Staessen, J. Wang, K. G. Moons, J. B. Reitsma, and J. Ensor, "Individual participant data meta-analysis to examine interactions between treatment effect and

participant-level covariates: Statistical recommendations for conduct and planning," *Statistics in Medicine*, vol. 39, no. 15, pp. 2115–2137, 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.8516

[23] P. Dhiman, J. Ma, C. Qi, G. Bullock, J. C. Sergeant, R. D. Riley, and G. S. Collins, "Sample size requirements are not being considered in studies developing prediction models for binary outcomes: a systematic review," *BMC Medical Research Methodology*, vol. 23, no. 1, p. 188, 2023. [Online]. Available: https://doi.org/10.1186/s12874-023-02008-1

[24] P. Dhiman, J. Ma, V. N. Gibbs, A. Rampotas, H. Kamal, S. S. Arshad, S. Kirtley, C. Doree, M. F. Murphy, G. S. Collins, and A. J. R. Palmer, "Systematic review highlights high risk of bias of clinical prediction models for blood transfusion in patients undergoing elective surgery," *J Clin Epidemiol*, vol. 159, pp. 10–30, 2023.

[25] P. Dhiman, J. Ma, C. A. Navarro, B. Speich, G. Bullock, J. A. Damen, S. Kirtley, L. Hooft, R. D. Riley, B. Van Calster, K. G. M. Moons, and G. S. Collins, "Reporting of prognostic clinical prediction models based on machine learning methods in oncology needs to be improved," *J Clin Epidemiol*, vol. 138, pp. 60–72, 2021.

[26] M. T. Hudda, L. Archer, M. van Smeden, K. G. M. Moons, G. S. Collins, E. W. Steyerberg, C. Wahlich, J. B. Reitsma, R. D. Riley, B. Van Calster, and L. Wynants, "Minimal reporting improvement after peer review in reports of covid-19 prediction models: systematic review," *J Clin Epidemiol*, vol. 154, pp. 75–84, 2023.

[27] P. Perel, M. Arango, T. Clayton, P. Edwards, E. Komolafe, S. Poccock, I. Roberts, H. Shakur, E. Steyerberg, and S. Yutthakasemsunt, "Predicting outcome after traumatic brain injury: practical prognostic models based on large cohort of international patients," *BMJ*, vol. 336, no. 7641, pp. 425–9, 2008.

[28] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American Statistical Association*, vol. 53, no. 282, pp. 457–481, 1958. [Online]. Available: Go to ISI://WOS:A1958WX09300012

[29] D. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society Series B-Statistical Methodology*, vol. 34, no. 2, pp. 187–220, 1972. [Online]. Available: WOS:A1972N572600003

[30] F. Harrell, *Cox Proportional Hazards Regression Model*. New York: Springer-Verlag, 2001, book section 19, pp. 465–507.

[31] P. Royston, "Flexible parametric alternatives to the cox model, and more," *Stata Journal*, vol. 1, no. 1, pp. 1–28, 2001.

[32] P. Royston and P. Lambert, *Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model*. College Station, Texas: Stata Press, 2011.

[33] G. S. Collins, E. O. Ogundimu, J. A. Cook, Y. L. Manach, and D. G. Altman, "Quantifying the impact of different approaches for handling continuous predictors on the performance of a prognostic model," *Stat Med*, vol. 35, no. 23, pp. 4124–35, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/27193918

[34] P. Royston, D. G. Altman, and W. Sauerbrei, "Dichotomizing continuous predictors in multiple regression: a bad idea," *Stat Med*, vol. 25, no. 1, pp. 127–41, 2006.

[35] P. Royston, G. Ambler, and W. Sauerbrei, "The use of fractional polynomials to model continuous risk variables in epidemiology," *Int J Epidemiol*, vol. 28, no. 5, pp. 964–74, 1999.

[36] P. Royston and W. Sauerbrei, "Multivariable modeling with cubic regression splines: A principled approach," *Stata Journal*, vol. 7, no. 1, pp. 45–70, 2007. [Online]. Available: Go to ISI/WOS:000244769500003

[37] ——, "Building multivariable regression models with continuous covariates in clinical epidemiology - with an emphasis on fractional polynomials," *Methods of Information in Medicine*, vol. 44, no. 4, pp. 561–571, 2005. [Online]. Available: Go to ISI://000232490100012

[38] ——, *Multivariable model-building - A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables.* Chichester: John Wiley, 2008.

[39] W. Sauerbrei and P. Royston, "Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials," *Journal of the Royal Statistical Society Series A-Statistics in Society*, vol. 162, pp. 71–94, 1999. [Online]. Available: WOS:000078538300006

[40] P. Royston and W. Sauerbrei, "A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials," *Stat Med*, vol. 23, no. 16, pp. 2509–25, 2004. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/15287081

[41] W. Sauerbrei, P. Royston, and K. Zapien, "Detecting an interaction between treatment and a continuous covariate: A comparison of two approaches," *Computational Statistics & Data Analysis*, vol. 51, no. 8, pp. 4054–4063, 2007. [Online]. Available: WOS:000246128500034

[42] P. Royston and W. Sauerbrei, "Two techniques for investigating interactions between treatment and continuous covariates in clinical trials," *Stata Journal*, vol. 9, no. 2, pp. 230–251, 2009.

[43] R. Whittle, K. L. Royle, K. P. Jordan, R. D. Riley, C. D. Mallen, and G. Peat, "Prognosis research ideally should measure time-varying predictors at their intended moment of use," *Diagn Progn Res*, vol. 1, p. 1, 2017.

[44] D. Rizopoulos and J. J. Takkenberg, "Tools & techniques–statistics: Dealing with time-varying covariates in survival analysis–joint models versus cox models," *EuroIntervention*, vol. 10, no. 2, pp. 285–8, 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24952063

[45] H. v. Houwelingen and H. Putter, *Dynamic Prediction in Clinical Survival Analysis*. CRC Press, Inc., 2011.

[46] A. C. Justice, K. E. Covinsky, and J. A. Berlin, "Assessing the generalizability of prognostic information," *Ann Intern Med*, vol. 130, no. 6, pp. 515–24, 1999. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/10075620

[47] E. W. Steyerberg and Y. Vergouwe, "Towards better clinical prediction models: seven steps for development and an abcd for validation," *Eur Heart J*, vol. 35, no. 29, pp. 1925–31, 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24898551

[48] K. G. Moons, A. P. Kengne, M. Woodward, P. Royston, Y. Vergouwe, D. G. Altman, and D. E. Grobbee, "Risk prediction models: I. development, internal validation, and assessing the incremental value of a new (bio)marker," *Heart*, vol. 98, no. 9, pp. 683–90, 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22397945

[49] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society Series B-Methodological*, vol. 58, no. 1, pp. 267–288, 1996.

[50] J. Van Houwelingen, "Shrinkage and penalized likelihood as methods to improve predictive accuracy," *Statistica Neerlandica*, vol. 55, no. 1, pp. 17–34, 2001.

[51] E. W. Steyerberg, "Validation in prediction research: the waste by data splitting," *J Clin Epidemiol*, vol. 103, pp. 131–133, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/30063954

[52] E. Steyerberg, F. Harrell, G. Borsboom, M. Eijkemans, Y. Vergouwe, and J. Habbema, "Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis," *Journal of Clinical Epidemiology*, vol. 54, no. 8, pp. 774–781, 2001. [Online]. Available: WOS:000170019900003

[53] P. C. Austin and E. W. Steyerberg, "Events per variable (epv) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models," *Statistical Methods in Medical Research*, vol. 26, no. 2, pp. 796–808, 2017. [Online]. Available: https://journals.sagepub.com/doi/abs/10.1177/0962280214558972

[54] J. Van Houwelingen and S. Le Cessie, "Predictive value of statistical models," *Stat Med*, vol. 9, no. 11, 1990.

[55] J. Copas, "Regression, prediction and shrinkage," *Journal of the Royal Statistical Society Series B-Methodological*, vol. 45, no. 3, pp. 311–354, 1983. [Online]. Available: Go to ISI://A1983RY02400001

[56] E. W. Steyerberg and J. Harrell, F. E., "Prediction models need appropriate internal, internal-external, and external validation," *J Clin Epidemiol*, vol. 69, pp. 245–7, 2016.

[57] D. G. Altman, Y. Vergouwe, P. Royston, and K. G. M. Moons, "Prognosis and prognostic research: validating a prognostic model," *BMJ*, vol. 338, p. b605, 2009.

[58] S. E. Bleeker, H. A. Moll, E. W. Steyerberg, A. R. Donders, G. Derksen-Lubsen, D. E. Grobbee, and K. G. Moons, "External validation is necessary in prediction research: a clinical example," *J Clin Epidemiol*, vol. 56, no. 9, pp. 826–32, 2003. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/14505766

[59] K. G. Moons, A. P. Kengne, D. E. Grobbee, P. Royston, Y. Vergouwe, D. G. Altman, and M. Woodward, "Risk prediction models: Ii. external validation, model updating, and impact assessment," *Heart*, vol. 98, no. 9, pp. 691–8, 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22397946

[60] E. Steyerberg, *Clinical prediction models: A practical approach to development, validation, and updating*. New York, NY: Springer, 2009.

[61] T. P. Debray, Y. Vergouwe, H. Koffijberg, D. Nieboer, E. W. Steyerberg, and K. G. Moons, "A new framework to enhance the interpretation of external validation studies of clinical prediction models," *J Clin Epidemiol*, vol. 68, no. 3, pp. 279–89, 2015. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/25179855

[62] D. Altman and P. Royston, "What do we mean by validating a prognostic model?" *Statistics in Medicine*, vol. 19, no. 4, pp. 453–473, 2000. [Online]. Available: WOS:000085613900005

[63] R. D. Riley, J. Ensor, K. I. E. Snell, T. P. A. Debray, D. G. Altman, K. G. M. Moons, and G. S. Collins, "External validation of clinical prediction models using big datasets from e-health records or ipd meta-analysis: opportunities and challenges," *BMJ*, vol. 353, p. i3140, 2016. [Online]. Available: https://www.bmj.com/content/bmj/353/bmj.i3140.full.pdf

[64] T. P. Debray, J. A. Damen, R. D. Riley, K. Snell, J. B. Reitsma, L. Hooft, G. S. Collins, and K. G. Moons, "A framework for meta-analysis of prediction model studies with binary and

time-to-event outcomes," *Stat Methods Med Res*, vol. 28, no. 9, pp. 2768–2786, 2018.

[65] T. P. Debray, J. A. Damen, K. I. Snell, J. Ensor, L. Hooft, J. B. Reitsma, R. D. Riley, and K. G. Moons, "A guide to systematic review and meta-analysis of prediction model performance," *BMJ*, vol. 356, p. i6460, 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28057641

[66] I. Ahmed, T. P. Debray, K. G. Moons, and R. D. Riley, "Developing and validating risk prediction models in an individual participant data meta-analysis," *BMC Med Res Methodol*, vol. 14, p. 3, 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24397587

[67] T. P. Debray, R. D. Riley, M. M. Rovers, J. B. Reitsma, K. G. Moons, and I. P. D. M.-a. M. g. Cochrane, "Individual participant data (ipd) meta-analyses of diagnostic and prognostic modeling studies: guidance on their use," *PLoS Med*, vol. 12, no. 10, p. e1001886, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/26461078

[68] F. Harrell, K. Lee, and D. Mark, "Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors," *Statistics in Medicine*, vol. 15, no. 4, pp. 361–387, 1996. [Online]. Available: Go to ISI://A1996TY77400003

[69] A. J. Vickers, A. M. Cronin, E. B. Elkin, and M. Gonen, "Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers," *BMC Med Inform Decis Mak*, vol. 8, p. 53, 2008. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/19036144

[70] A. J. Vickers and E. B. Elkin, "Decision curve analysis: a novel method for evaluating prediction models," *Med Decis Making*, vol. 26, no. 6, pp. 565–74, 2006. [Online]. Available:

https://www.ncbi.nlm.nih.gov/pubmed/17099194

[71] A. J. Vickers, B. Van Calster, and E. W. Steyerberg, "Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests," *BMJ*, vol. 352:i6, 2016.

[72] K. I. Snell, J. Ensor, T. P. Debray, K. G. Moons, and R. D. Riley, "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*, p. 962280217705678, 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28480827

[73] L. Wynants, R. D. Riley, D. Timmerman, and B. Van Calster, "Random-effects meta-analysis of the clinical utility of tests and prediction models," *Stat Med*, vol. 37, no. 12, pp. 2034–2052, 2018.

[74] C. L. Andaur Navarro, J. A. A. Damen, T. Takada, S. W. J. Nijman, P. Dhiman, J. Ma, G. S. Collins, R. Bajpai, R. D. Riley, K. G. M. Moons, and L. Hooft, "Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review," *BMJ*, vol. 375, p. n2281, 2021. [Online]. Available: https://www.bmj.com/content/bmj/375/bmj.n2281.full.pdf

[75] R. D. Riley, K. I. E. Snell, J. Ensor, D. L. Burke, F. E. Harrell Jr, K. G. M. Moons, and G. S. Collins, "Minimum sample size for developing a multivariable prediction model: Part i – continuous outcomes," *Statistics in Medicine*, vol. 38, no. 7, pp. 1262–1275, 2019. [Online]. Available: https://doi.org/10.1002/sim.7993

[76] R. D. Riley, K. I. E. Snell, J. Ensor, D. L. Burke, J. Harrell, F. E., K. G. Moons, and G. S. Collins, "Minimum sample size for developing a multivariable prediction model: Part ii - binary and time-to-event outcomes," *Stat Med*, vol. 38, no. 7, pp. 1276–1296, 2019.

[77] M. van Smeden, K. G. Moons, J. A. de Groot, G. S. Collins, D. G. Altman, M. J. Eijkemans, and J. B. Reitsma, "Sample size for binary logistic prediction models: Beyond events per variable criteria," *Stat Methods Med Res*, p. 962280218784726, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29966490

[78] R. D. Riley, J. Ensor, K. I. E. Snell, J. Harrell, F. E., G. P. Martin, J. B. Reitsma, K. G. M. Moons, G. Collins, and M. van Smeden, "Calculating the sample size required for developing a clinical prediction model," *BMJ*, vol. 368, p. m441, 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/32188600

[79] B. Van Calster, D. Nieboer, Y. Vergouwe, B. De Cock, M. J. Pencina, and E. W. Steyerberg, "A calibration hierarchy for risk models was defined: from utopia to empirical data," *J Clin Epidemiol*, vol. 74, pp. 167–76, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/26772608

[80] B. Van Calster, D. J. McLernon, M. van Smeden, L. Wynants, E. W. Steyerberg, t. Topic Group 'Evaluating diagnostic, and S. i. prediction models' of the, "Calibration: the achilles heel of predictive analytics," *BMC Med*, vol. 17, no. 1, p. 230, 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/31842878

[81] P. Dhiman, J. Ma, C. L. Andaur Navarro, B. Speich, G. Bullock, J. A. A. Damen, L. Hooft, S. Kirtley, R. D. Riley, B. Van Calster, K. G. M. Moons, and G. S. Collins, "Overinterpretation of findings in machine learning prediction model studies in oncology: a systematic review," *J Clin Epidemiol*, vol. 157, pp. 120–133, 2023.

[82] C. L. Andaur Navarro, J. A. A. Damen, T. Takada, S. W. J. Nijman, P. Dhiman, J. Ma, G. S. Collins, R. Bajpai, R. D. Riley, K. G. M. Moons, and L. Hooft, "Systematic review

finds "spin" practices and poor reporting standards in studies on machine learning-based prediction models," *J Clin Epidemiol*, vol. 158, pp. 99–110, 2023.

[83] C. L. Andaur Navarro, J. A. A. Damen, M. Ghannad, P. Dhiman, M. van Smeden, J. B. Reitsma, G. S. Collins, R. D. Riley, K. G. M. Moons, and L. Hooft, "Spin-pm: a consensus framework to evaluate the presence of spin in studies on prediction models," *J Clin Epidemiol*, vol. 170, p. 111364, 2024.

[84] G. S. Collins, J. B. Reitsma, D. G. Altman, K. G. Moons, and T. g. for the members of the, "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): The tripod statement," *Eur Urol*, 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/25572824 http://www.sciencedirect.com/science/article/pii/S0302283814011993

[85] K. G. Moons, D. G. Altman, J. B. Reitsma, and et al, "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): Explanation and elaboration," *Annals of Internal Medicine*, vol. 162, no. 1, pp. W1–W73, 2015. [Online]. Available: https://www.acpjournals.org/doi/abs/10.7326/M14-0698 (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) Statement includes a 22-item checklist, which aims to improve the reporting of studies developing, validating, or updating a prediction model, whether for diagnostic or prognostic purposes. The TRIPOD Statement aims to improve the transparency of the reporting of a prediction model study regardless of the study methods used. This explanation and elaboration document describes the rationale; clarifies the meaning of each item; and discusses why transparent reporting is important, with a view to assessing risk of bias and clinical usefulness of the prediction model. Each checklist item of the TRIPOD

Statement is explained in detail and accompanied by published examples of good reporting. The document also provides a valuable reference of issues to consider when designing, conducting, and analyzing prediction model studies. To aid the editorial process and help peer reviewers and, ultimately, readers and systematic reviewers of prediction model studies, it is recommended that authors include a completed checklist in their submission. The TRIPOD checklist can also be downloaded from www.tripod-statement.org.

[86] T. P. A. Debray, G. S. Collins, R. D. Riley, K. I. E. Snell, B. Van Calster, J. B. Reitsma, and K. G. M. Moons, "Transparent reporting of multivariable prediction models developed or validated using clustered data: Tripod-cluster checklist," *BMJ*, vol. 380, p. e071018, 2023. [Online]. Available: https://www.bmj.com/content/bmj/380/bmj-2022-071018.full.pdf

[87] K. I. E. Snell, B. Levis, J. A. A. Damen, P. Dhiman, T. P. A. Debray, L. Hooft, J. B. Reitsma, K. G. M. Moons, G. S. Collins, and R. D. Riley, "Transparent reporting of multivariable prediction models for individual prognosis or diagnosis: checklist for systematic reviews and meta-analyses (tripod-srma)," *BMJ*, vol. 381, p. e073538, 2023. [Online]. Available: https://www.bmj.com/content/bmj/381/bmj-2022-073538.full.pdf

[88] G. S. Collins, K. G. M. Moons, P. Dhiman, R. D. Riley, A. L. Beam, B. Van Calster, M. Ghassemi, X. Liu, J. B. Reitsma, M. van Smeden, A.-L. Boulesteix, J. C. Camaradou, L. A. Celi, S. Denaxas, A. K. Denniston, B. Glocker, R. M. Golub, H. Harvey, G. Heinze, M. M. Hoffman, A. P. Kengne, E. Lam, N. Lee, E. W. Loder, L. Maier-Hein, B. A. Mateen, M. D. McCradden, L. Oakden-Rayner, J. Ordish, R. Parnell, S. Rose, K. Singh, L. Wynants, and P. Logullo, "Tripod+ai statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods," *BMJ*, vol. 385, p. e078378, 2024. [Online]. Available: https://www.bmj.com/content/bmj/385/bmj-2023-078378.full.pdf

[89] C. L. Andaur Navarro, J. A. A. Damen, T. Takada, S. W. J. Nijman, P. Dhiman, J. Ma, G. S. Collins, R. Bajpai, R. D. Riley, K. G. M. Moons, and L. Hooft, "Completeness of reporting of clinical prediction models developed using supervised machine learning: a systematic review," *BMC Med Res Methodol*, vol. 22, no. 1, p. 12, 2022.

[90] G. S. Collins, R. Whittle, G. S. Bullock, P. Logullo, P. Dhiman, J. A. de Beyer, R. D. Riley, and M. M. Schlussel, "Open science practices need substantial improvement in prognostic model studies in oncology using machine learning," *J Clin Epidemiol*, vol. 165, p. 111199, 2024.

[91] L. Archer, K. I. E. Snell, S. Stynes, I. Axén, K. M. Dunn, N. E. Foster, G. Wynne-Jones, D. A. van der Windt, and J. C. Hill, "Development and external validation of individualized prediction models for pain intensity outcomes in patients with neck pain, low back pain, or both in primary care settings," *Phys Ther*, vol. 103, no. 11, 2023.

[92] L. Archer, K. I. Snell, J. Ensor, M. T. Hudda, G. S. Collins, and R. D. Riley, "Minimum sample size for external validation of a clinical prediction model with a continuous outcome," *Stat Med*, vol. 40, no. 1, pp. 133–146, 2020.

[93] R. Riley, G. Collins, J. Ensor, L. Archer, S. Booth, S. Mozumder, M. Rutherford, M. van Smeden, P. Lambert, and K. Snell, "Minimum sample size calculations for external validation of a clinical prediction model with a time-to-event outcome," *Statistics in Medicine*, vol. 41, no. 7, p. 1280–1295, 2022.

[94] R. D. Riley, T. P. A. Debray, G. S. Collins, L. Archer, J. Ensor, M. van Smeden, and K. I. E. Snell, "Minimum sample size for external validation of a clinical prediction model with a binary outcome," *Statistics in Medicine*, vol. 40, no. 19, pp. 4230–4251, 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.9025

[95] K. I. E. Snell, L. Archer, J. Ensor, L. J. Bonnett, T. P. Debray, B. Phillips, G. S. Collins, and R. D. Riley, "External validation of clinical prediction models: simulation-based sample size calculations were more reliable than rules-of-thumb," *J Clin Epidemiol*, vol. S0895-4356(21)00048-2, 2021.

[96] R. D. Riley, K. I. E. Snell, L. Archer, J. Ensor, T. P. A. Debray, B. van Calster, M. van Smeden, and G. S. Collins, "Evaluation of clinical prediction models (part 3): calculating the sample size required for an external validation study," *BMJ*, vol. 384, p. e074821, 2024. [Online]. Available: https://www.bmj.com/content/bmj/384/bmj-2023-074821.full.pdf

[97] L. Archer, C. Koshiaris, S. Lay-Flurrie, K. I. E. Snell, R. D. Riley, R. Stevens, A. Banerjee, J. A. Usher-Smith, A. Clegg, R. A. Payne, F. D. R. Hobbs, R. J. McManus, and J. P. Sheppard, "Development and external validation of a risk prediction model for falls in patients with an indication for antihypertensive treatment: retrospective cohort study," *BMJ*, vol. 379, p. e070918, 2022.

[98] C. Koshiaris, L. Archer, S. Lay-Flurrie, K. I. Snell, R. D. Riley, R. J. Stevens, A. Banerjee, J. A. Usher-Smith, A. P. Clegg, R. A. Payne, M. Ogden, F. R. Hobbs, R. McManus, and J. Sheppard, "Predicting the risk of acute kidney injury: Derivation and validation of stratify-aki," *British Journal of General Practice*, p. BJGP.2022.0389, 2023. [Online]. Available: https://bjgp.org/content/bjgp/early/2023/01/31/BJGP.2022.0389.full.pdf

[99] L. Archer, S. D. Relton, A. Akbari, K. Best, M. Bucknall, S. Conroy, M. Hattle, J. Hollinghurst, S. Humphrey, R. A. Lyons, S. Richards, K. Walters, R. West, D. van der Windt, R. D. Riley, and A. Clegg, "Development and external validation of the efalls tool: a multivariable prediction model for the risk of ed attendance or hospitalisation with a fall or fracture in older adults," *Age Ageing*, vol. 53, no. 3, 2024.

[100] L. Wynants, M. van Smeden, D. J. McLernon, D. Timmerman, E. W. Steyerberg, B. Van Calster, t. on behalf of the Topic Group 'Evaluating diagnostic, and S. i. prediction models' of the, "Three myths about risk thresholds for prediction models," *BMC Medicine*, vol. 17, no. 1, p. 192, 2019. [Online]. Available: https://doi.org/10.1186/s12916-019-1425-3

[101] D. Altman and P. Royston, "Statistics notes - the cost of dichotomising continuous variables," *British Medical Journal*, vol. 332, no. 7549, pp. 1080–1080, 2006. [Online]. Available: Go to ISI://000237518400022

[102] S. Senn, "Individual response to treatment: is it a valid assumption?" *BMJ*, vol. 329, no. 7472, pp. 966–968, 2004. [Online]. Available: https://www.bmj.com/content/bmj/329/7472/966.full.pdf

[103] S. Senn and S. Julious, "Measurement in clinical trials: a neglected issue for statisticians?" *Stat Med*, vol. 28, no. 26, pp. 3189–209, 2009.

[104] G. S. Collins, S. Mallett, O. Omar, and L. M. Yu, "Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting," *BMC Med*, vol. 9, p. 103, 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21902820

[105] C. Chilvers, M. Dewey, K. Fielding, V. Gretton, P. Miller, B. Palmer, D. Weller, R. Churchill, I. Williams, N. Bedi, C. Duggan, A. Lee, and G. Harrison, "Antidepressant drugs and generic counselling for treatment of major depression in primary care: randomised trial with patient preference arms," *BMJ*, vol. 322, no. 7289, p. 772, 2001. [Online]. Available: https://www.bmj.com/content/bmj/322/7289/772.full.pdf

[106] R. D. Riley, T. J. Cole, J. Deeks, J. J. Kirkham, J. Morris, R. Perera, A. Wade, and G. S. Collins, "On the 12th day of christmas, a statistician sent to me," *BMJ*, vol. 379, p. e072883,

2022. [Online]. Available: https://www.bmj.com/content/bmj/379/bmj-2022-072883.full.pdf

[107] C. Vayssière, L. Sentilhes, A. Ego, C. Bernard, D. Cambourieu, C. Flamant, G. Gascoin, A. Gaudineau, G. Grangé, V. Houfflin-Debarge, B. Langer, V. Malan, P. Marcorelles, J. Nizard, F. Perrotin, L. Salomon, M. V. Senat, A. Serry, V. Tessier, P. Truffert, V. Tsatsaris, C. Arnaud, and B. Carbonne, "Fetal growth restriction and intra-uterine growth restriction: guidelines for clinical practice from the french college of gynaecologists and obstetricians," *Eur J Obstet Gynecol Reprod Biol*, vol. 193, pp. 10–8, 2015.

[108] J. Deeks, J. Higgins, D. Altman, and (editors), *Chapter 9: Analysing data and undertaking meta-analyses*, The Cochrane Collaboration, 2011, pp. Available from www.cochrane–handbook.org. [Online]. Available: www.cochrane-handbook.org

[109] J. Allotey, S. Thangaratinam, J. Zamora, A. Khalil, B. Thilaganathan, R. Riley, K. Snell, G. Smith, L. Chappell, J. Myers, R. Morris, A. Papageorghiou, S. Gordijn, W. Ganzevoort, B. Mol, V. Flenady, L. Askie, A. P. B. Lazaga, H. Mistry, and J. Dodds, "Prediction of fetal growth restriction and complications: individual participant data (ipd) meta-analysis with decision curve analysis. international prediction of complications in pregnancy: Fetal growth restriction (ippic-fgr)." *PROSPERO CRD42019135045*, 2019.

[110] R. F. Wolff, K. G. M. Moons, R. D. Riley, P. F. Whiting, M. Westwood, G. S. Collins, J. B. Reitsma, J. Kleijnen, S. Mallett, and P. Groupdagger, "Probast: A tool to assess the risk of bias and applicability of prediction model studies," *Ann Intern Med*, vol. 170, no. 1, pp. 51–58, 2019. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/30596875

[111] K. G. M. Moons, R. F. Wolff, R. D. Riley, P. F. Whiting, M. Westwood, G. S. Collins, J. B. Reitsma, J. Kleijnen, and S. Mallett, "Probast: A tool to assess risk of bias and applicability

of prediction model studies: Explanation and elaboration," *Ann Intern Med*, vol. 170, no. 1, pp. W1–W33, 2019. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/30596876

[112] H. Yadav and N. Lee, "Maternal factors in predicting low birth weight babies," *Med J Malaysia*, vol. 68, no. 1, pp. 44–7, 2013.

[113] D. A. Doherty, I. R. James, and J. P. Newnham, "Estimation of the doppler ultrasound umbilical maximal waveform envelope: Ii. prediction of fetal distress," *Ultrasound Med Biol*, vol. 28, no. 10, pp. 1261–70, 2002.

[114] A. Sharp, R. Jackson, C. Cornforth, J. Harrold, M. A. Turner, L. Kenny, P. N. Baker, E. D. Johnstone, A. Khalil, P. von Dadelszen, A. T. Papageorghiou, and Z. Alfirevic, "A prediction model for short-term neonatal outcomes in severe early-onset fetal growth restriction," *Eur J Obstet Gynecol Reprod Biol*, vol. 241, pp. 109–118, 2019.

[115] L. M. Bachmann, K. S. Khan, J. Ogah, and P. Owen, "Multivariable analysis of tests for the diagnosis of intrauterine growth restriction," *Ultrasound Obstet Gynecol*, vol. 21, no. 4, pp. 370–4, 2003.

[116] C. Lesmes, D. M. Gallo, J. Panaiotova, L. C. Poon, and K. H. Nicolaides, "Prediction of small-for-gestational-age neonates: screening by fetal biometry at 19-24 weeks," *Ultrasound Obstet Gynecol*, vol. 46, no. 2, pp. 198–207, 2015.

[117] A. Ciobanu, A. Rouvali, A. Syngelaki, R. Akolekar, and K. H. Nicolaides, "Prediction of small for gestational age neonates: screening by maternal factors, fetal biometry, and biomarkers at 35-37 weeks' gestation," *Am J Obstet Gynecol*, vol. 220, no. 5, pp. 486.e1–486.e11, 2019.

[118] A. Ciobanu, C. Formuso, A. Syngelaki, R. Akolekar, and K. H. Nicolaides, "Prediction of small-for-gestational-age neonates at 35-37-weeks' gestation: contribution of maternal factors

and growth velocity between 20 and 36-weeks," *Ultrasound Obstet Gynecol*, vol. 53, no. 4, pp. 488–495, 2019.

[119] A. Sotiriadis, F. Figueras, M. Eleftheriades, G. K. Papaioannou, G. Chorozoglou, K. Dinas, and N. Papantoniou, "First-trimester and combined first- and second-trimester prediction of small-for-gestational age and late fetal growth restriction," *Ultrasound Obstet Gynecol*, vol. 53, no. 1, pp. 55–61, 2019.

[120] W. Snidvongs, S. Bhongsvej, P. Witoonpanich, P. Thaitumyanond, D. Charoenvidhya, V. Wiswasukmongkol, Y. Tannirandorn, and D. Trisukosol, "Intrauterine growth retardation: incidence, screening results, pregnancy outcome," *J Med Assoc Thai*, vol. 72, no. 7, pp. 387–94, 1989.

[121] F. de Caunes, G. R. Alexander, C. Berchel, J. P. Guengant, and E. Papiernik, "Anamnestic pregnancy risk assessment," *Int J Gynaecol Obstet*, vol. 33, no. 3, pp. 221–7, 1990.

[122] C. Kienast, W. Moya, O. Rodriguez, A. Jijón, and A. Geipel, "Predictive value of angiogenic factors, clinical risk factors and uterine artery doppler for pre-eclampsia and fetal growth restriction in second and third trimester pregnancies in an ecuadorian population," *J Matern Fetal Neonatal Med*, vol. 29, no. 4, pp. 537–43, 2016.

[123] M. L. E. Hendrix, J. A. P. Bons, R. R. G. Snellings, O. Bekers, S. M. J. van Kuijk, M. E. A. Spaanderman, and S. Al-Nasiry, "Can fetal growth velocity and first trimester maternal biomarkers improve the prediction of small-for-gestational age and adverse neonatal outcome?" *Fetal Diagn Ther*, vol. 46, no. 4, pp. 274–284, 2019.

[124] D. Anggraini, M. Abdollahian, and K. Marion, "Foetal weight prediction models at a given gestational age in the absence of ultrasound facilities: application in indonesia," *BMC*

Pregnancy Childbirth, vol. 18, no. 1, p. 436, 2018.

[125] J. Miranda, M. Rodriguez-Lopez, S. Triunfo, M. Sairanen, H. Kouru, M. Parra-Saavedra, F. Crovetto, F. Figueras, F. Crispi, and E. Gratacós, "Prediction of fetal growth restriction using estimated fetal weight vs a combined screening model in the third trimester," *Ultrasound Obstet Gynecol*, vol. 50, no. 5, pp. 603–611, 2017.

[126] R. Allen and J. Aquilina, "Prospective observational study to determine the accuracy of first-trimester serum biomarkers and uterine artery dopplers in combination with maternal characteristics and arteriography for the prediction of women at risk of preeclampsia and other adverse pregnancy outcomes," *J Matern Fetal Neonatal Med*, vol. 31, no. 21, pp. 2789–2806, 2018.

[127] L. M. McCowan, J. M. Thompson, R. S. Taylor, P. N. Baker, R. A. North, L. Poston, C. T. Roberts, N. A. Simpson, J. J. Walker, J. Myers, and L. C. Kenny, "Prediction of small for gestational age infants in healthy nulliparous women using clinical and ultrasound risk factors combined with early pregnancy biomarkers," *PLoS One*, vol. 12, no. 1, p. e0169311, 2017.

[128] S. M. Kim, H. G. Yun, R. Y. Kim, Y. H. Chung, J. Y. Cheon, J. H. Wie, J. Y. Kwon, H. S. Ko, Y. H. Kim, E. H. Han, J. H. Park, H. J. Kim, M. S. Kim, J. C. Shin, and I. Y. Park, "Maternal serum placental growth factor combined with second trimester aneuploidy screening to predict small-for-gestation neonates without preeclampsia," *Taiwan J Obstet Gynecol*, vol. 56, no. 6, pp. 801–805, 2017.

[129] E. Litwińska, M. Litwińska, P. Oszukowski, K. Szaflik, and P. Kaczmarek, "Combined screening for early and late pre-eclampsia and intrauterine growth restriction by maternal history, uterine artery doppler, mean arterial pressure and biochemical markers," *Adv Clin Exp Med*, vol. 26, no. 3, pp. 439–448, 2017.

[130] N. L. González-González, E. González-Dávila, L. González Marrero, E. Padrón, J. R. Conde, and W. Plasencia, "Value of placental volume and vascular flow indices as predictors of intrauterine growth retardation," *Eur J Obstet Gynecol Reprod Biol*, vol. 212, pp. 13–19, 2017.

[131] F. Crovetto, S. Triunfo, F. Crispi, V. Rodriguez-Sureda, C. Dominguez, F. Figueras, and E. Gratacos, "Differential performance of first-trimester screening in predicting small-for-gestational-age neonate or fetal growth restriction," *Ultrasound Obstet Gynecol*, vol. 49, no. 3, pp. 349–356, 2017.

[132] F. Crovetto, S. Triunfo, F. Crispi, V. Rodriguez-Sureda, E. Roma, C. Dominguez, E. Gratacos, and F. Figueras, "First-trimester screening with specific algorithms for early- and late-onset fetal growth restriction," *Ultrasound Obstet Gynecol*, vol. 48, no. 3, pp. 340–8, 2016.

[133] C. Lesmes, D. M. Gallo, R. Gonzalez, L. C. Poon, and K. H. Nicolaides, "Prediction of small-for-gestational-age neonates: screening by maternal serum biochemical markers at 19-24 weeks," *Ultrasound Obstet Gynecol*, vol. 46, no. 3, pp. 341–9, 2015.

[134] C. Lesmes, D. M. Gallo, Y. Saiid, L. C. Poon, and K. H. Nicolaides, "Prediction of small-for-gestational-age neonates: screening by uterine artery doppler and mean arterial pressure at 19-24 weeks," *Ultrasound Obstet Gynecol*, vol. 46, no. 3, pp. 332–40, 2015.

[135] C. Fadigas, Y. Saiid, R. Gonzalez, L. C. Poon, and K. H. Nicolaides, "Prediction of small-for-gestational-age neonates: screening by fetal biometry at 35-37 weeks," *Ultrasound Obstet Gynecol*, vol. 45, no. 5, pp. 559–65, 2015.

[136] C. Fadigas, L. Guerra, S. Garcia-Tizon Larroca, L. C. Poon, and K. H. Nicolaides, "Prediction of small-for-gestational-age neonates: screening by uterine artery doppler and mean arterial pressure at 35-37 weeks," *Ultrasound Obstet Gynecol*, vol. 45, no. 6, pp. 715–21, 2015.

[137] S. Bakalis, G. Peeva, R. Gonzalez, L. C. Poon, and K. H. Nicolaides, "Prediction of small-for-gestational-age neonates: screening by biophysical and biochemical markers at 30-34 weeks," *Ultrasound Obstet Gynecol*, vol. 46, no. 4, pp. 446–51, 2015.

[138] S. Bakalis, D. M. Gallo, O. Mendez, L. C. Poon, and K. H. Nicolaides, "Prediction of small-for-gestational-age neonates: screening by maternal biochemical markers at 30-34 weeks," *Ultrasound Obstet Gynecol*, vol. 46, no. 2, pp. 208–15, 2015.

[139] S. Bakalis, M. Silva, R. Akolekar, L. C. Poon, and K. H. Nicolaides, "Prediction of small-for-gestational-age neonates: screening by fetal biometry at 30-34 weeks," *Ultrasound Obstet Gynecol*, vol. 45, no. 5, pp. 551–8, 2015.

[140] C. Macdonald-Wallis, R. J. Silverwood, B. L. de Stavola, H. Inskip, C. Cooper, K. M. Godfrey, S. Crozier, A. Fraser, S. M. Nelson, D. A. Lawlor, and K. Tilling, "Antenatal blood pressure for prediction of pre-eclampsia, preterm birth, and small for gestational age babies: development and validation in two general population cohorts," *Bmj*, vol. 351, p. h5948, 2015.

[141] I. Papastefanou, A. P. Souka, M. Eleftheriades, A. Pilalis, C. Chrelias, and D. Kassanos, "Predicting fetal growth deviation in parous women: combining the birth weight of the previous pregnancy and third trimester ultrasound scan," *J Perinat Med*, vol. 43, no. 4, pp. 485–92, 2015.

[142] V. Seravalli, D. M. Block-Abraham, O. M. Turan, L. E. Doyle, J. N. Kopelman, R. O. Atlas, C. B. Jenkins, M. G. Blitzer, and A. A. Baschat, "First-trimester prediction of small-for-gestational age neonates incorporating fetal doppler parameters and maternal characteristics," *Am J Obstet Gynecol*, vol. 211, no. 3, pp. 261.e1–8, 2014.

[143] V. Seravalli, D. M. Block-Abraham, O. M. Turan, L. E. Doyle, M. G. Blitzer, and A. A. Baschat, "Second-trimester prediction of delivery of a small-for-gestational-age neonate: integrating sequential doppler information, fetal biometry, and maternal characteristics," *Prenat Diagn*, vol. 34, no. 11, pp. 1037–43, 2014.

[144] N. Schwartz, M. D. Sammel, R. Leite, and S. Parry, "First-trimester placental ultrasound and maternal serum markers as predictors of small-for-gestational-age infants," *Am J Obstet Gynecol*, vol. 211, no. 3, pp. 253.e1–8, 2014.

[145] I. Boucoiran, S. Thissier-Levy, Y. Wu, S. Q. Wei, Z. C. Luo, E. Delvin, W. D. Fraser, and F. Audibert, "Risks for preeclampsia and small for gestational age: predictive values of placental growth factor, soluble fms-like tyrosine kinase-1, and inhibin a in singleton and multiple-gestation pregnancies," *Am J Perinatol*, vol. 30, no. 7, pp. 607–12, 2013.

[146] L. C. Poon, A. Syngelaki, R. Akolekar, J. Lai, and K. H. Nicolaides, "Combined screening for preeclampsia and small for gestational age at 11-13 weeks," *Fetal Diagn Ther*, vol. 33, no. 1, pp. 16–27, 2013.

[147] S. Singh, U. Verma, K. Shrivastava, S. Khanduri, N. Goel, F. T. Zahra, and K. L. Shirivastava, "Role of color doppler in the diagnosis of intra uterine growth restriction (iugr)," *International journal of reproduction, contraception, obstetrics and gynecology*, vol. 2, pp. 566–572, 2013.

[148] P. T. Seed, L. C. Chappell, M. A. Black, K. K. Poppe, Y. C. Hwang, N. Kasabov, L. McCowan, A. H. Shennan, S. H. Wu, L. Poston, and R. A. North, "Prediction of preeclampsia and delivery of small for gestational age babies based on a combination of clinical risk factors in high-risk women," *Hypertens Pregnancy*, vol. 30, no. 1, pp. 58–73, 2011.

[149] L. C. Poon, G. Karagiannis, I. Staboulidou, A. Shafiei, and K. H. Nicolaides, "Reference range of birth weight with gestation and first-trimester prediction of small-for-gestation neonates," *Prenat Diagn*, vol. 31, no. 1, pp. 58–65, 2011.

[150] G. Karagiannis, R. Akolekar, R. Sarquis, D. Wright, and K. H. Nicolaides, "Prediction of small-for-gestation neonates from biophysical and biochemical markers at 11-13 weeks," *Fetal Diagn Ther*, vol. 29, no. 2, pp. 148–54, 2011.

[151] C. De Paco, N. Kametas, G. Rencoret, I. Strobl, and K. H. Nicolaides, "Maternal cardiac output between 11 and 13 weeks of gestation in the prediction of preeclampsia and small for gestational age," *Obstet Gynecol*, vol. 111, no. 2 Pt 1, pp. 292–300, 2008.

[152] N. Onwudiwe, C. K. Yu, L. C. Poon, I. Spiliopoulos, and K. H. Nicolaides, "Prediction of pre-eclampsia by a combination of maternal history, uterine artery doppler and mean arterial pressure," *Ultrasound Obstet Gynecol*, vol. 32, no. 7, pp. 877–83, 2008.

[153] C. M. Liu, S. D. Chang, and P. J. Cheng, "Prediction of fetal birthweight in taiwanese women with pre-eclampsia and gestational hypertension using an equation based on maternal characteristics," *J Obstet Gynaecol Res*, vol. 34, no. 4, pp. 480–6, 2008.

[154] W. Plasencia, N. Maiz, L. Poon, C. Yu, and K. H. Nicolaides, "Uterine artery doppler at 11 + 0 to 13 + 6 weeks and 21 + 0 to 24 + 6 weeks in the prediction of pre-eclampsia," *Ultrasound Obstet Gynecol*, vol. 32, no. 2, pp. 138–46, 2008.

[155] A. Pilalis, A. P. Souka, P. Antsaklis, G. Daskalakis, N. Papantoniou, S. Mesogitis, and A. Antsaklis, "Screening for pre-eclampsia and fetal growth restriction by uterine artery doppler and papp-a at 11-14 weeks' gestation," *Ultrasound Obstet Gynecol*, vol. 29, no. 2, pp. 135–40, 2007.

[156] N. Mamelle, V. Cochet, and O. Claris, "Definition of fetal growth restriction according to constitutional growth potential," *Biol Neonate*, vol. 80, no. 4, pp. 277–85, 2001.

[157] C. P. Weiner, R. E. Sabbagha, N. Vaisrub, and M. L. Socol, "Ultrasonic fetal weight prediction: role of head circumference and femur length," *Obstet Gynecol*, vol. 65, no. 6, pp. 812–7, 1985.

[158] B. J. Ingui and M. A. Rogers, "Searching for clinical prediction rules in medline," *J Am Med Inform Assoc*, vol. 8, no. 4, pp. 391–7, 2001. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/11418546

[159] G. J. Geersing, W. Bouwmeester, P. Zuithoff, R. Spijker, M. Leeflang, and K. G. Moons, "Search filters for finding prognostic and diagnostic prediction studies in medline to enhance systematic reviews," *PLoS One*, vol. 7, no. 2, p. e32844, 2012.

[160] S. S. Wong, N. L. Wilczynski, R. B. Haynes, and R. Ramkissoonsingh, "Developing optimal search strategies for detecting sound clinical prediction studies in medline," *AMIA Annu Symp Proc*, vol. 2003, pp. 728–32, 2003.

[161] R. Buchbinder, M. Underwood, J. Hartvigsen, and C. G. Maher, "The lancet series call to action to reduce low value care for low back pain: an update," *PAIN*, vol. 161, pp. S57–S64, 2020.

[162] J. M. Stevans, A. Delitto, S. S. Khoja, C. G. Patterson, C. N. Smith, M. J. Schneider, J. K. Freburger, C. M. Greco, J. A. Freel, G. A. Sowa, A. D. Wasan, G. P. Brennan, S. J. Hunter, K. I. Minick, S. T. Wegener, P. L. Ephraim, M. Friedman, J. M. Beneciuk, S. Z. George, and R. B. Saper, "Risk factors associated with transition from acute to chronic low back pain in us patients seeking primary care," *JAMA Network Open*, vol. 4, no. 2, pp. e2 037 371–e2 037 371, 2021. [Online]. Available: https://doi.org/10.1001/jamanetworkopen.2020.37371

[163] A. C. Traeger, N. Henschke, M. Hübscher, C. M. Williams, S. J. Kamper, C. G. Maher, G. L. Moseley, and J. H. McAuley, "Estimating the risk of chronic pain: Development and validation of a prognostic model (pickup) for patients with acute low back pain," *PLOS Medicine*, vol. 13, no. 5, p. e1002019, 2016. [Online]. Available: https://doi.org/10.1371/journal.pmed.1002019

[164] Back-UP, "Personalised prognostic models to improve well-being and return to work after neck and low back pain," 2020. [Online]. Available: http://backup-project.eu/

[165] K. Dunn, P. Campbell, M. Lewis, J. Hill, D. van der Windt, E. Afolabi, J. Protheroe, S. Wathall, S. Jowett, R. Oppong, C. Mallen, E. Hay, and N. Foster, "Refinement and validation of a tool for stratifying patients with musculoskeletal pain," *European Journal of Pain*, 2021.

[166] J. Hill, S. Garvin, Y. Chen, V. Cooper, S. Wathall, B. Saunders, M. Lewis, J. Protheroe, A. Chudyk, K. Dunn, E. Hay, D. van der Windt, C. Mallen, and N. Foster, "Stratified primary care versus non-stratified care for musculoskeletal pain: findings from the start msk feasibility and pilot cluster randomized controlled trial," *BMC Family Practice*, vol. 21, no. 30, 2020.

[167] J. C. Hill, S. Garvin, K. Bromley, B. Saunders, J. Kigozi, V. Cooper, M. Lewis, J. Protheroe, S. Wathall, A. Chudyk, K. M. Dunn, H. Birkinshaw, S. Jowett, E. M. Hay, D. van der Windt,

C. Mallen, and N. E. Foster, "Risk-based stratified primary care for common musculoskeletal pain presentations (start msk): a cluster-randomised, controlled trial," *The Lancet Rheumatology*, 2022. [Online]. Available: https://doi.org/10.1016/S2665-9913(22)00159-X

[168] C. Aun, Y. Lam, and B. Collect, "Evaluation of the use of visual analogue scale in chinese patients," *Pain*, vol. 25, pp. 215–21, 1986.

[169] M. Von Korff, J. Ormel, F. Keefe, and S. Dworkin, "Grading the severity of chronic pain," *Pain*, vol. 50, p. 133–149, 1992.

[170] R. D. Riley, K. I. E. Snell, J. Ensor, D. L. Burke, J. Harrell, F. E., K. G. M. Moons, and G. S. Collins, "Minimum sample size for developing a multivariable prediction model: Part i - continuous outcomes," *Stat Med*, vol. 38, no. 7, pp. 1262–1275, 2019. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/30347470

[171] J. Hill, D. Whitehurst, M. Lewis, S. Bryan, K. Dunn, N. Foster, and et al, "Comparison of stratified primary care management for low back pain with current best practice (start back): a randomised controlled trial," *Lancet*, vol. 378, no. 9802, pp. 1560–1571, 2011.

[172] G. S. Collins, E. O. Ogundimu, and D. G. Altman, "Sample size considerations for the external validation of a multivariable prognostic model: a resampling study," *Stat Med*, vol. 35, no. 2, pp. 214–26, 2016.

[173] I. White, P. Royston, and A. Wood, "Multiple imputation using chained equations: Issues and guidance for practice," *Statistics in Medicine*, vol. 30, no. 4, pp. 377–399, 2011. [Online]. Available: Go to ISI://000287106200008

[174] Y. Vergouwe, P. Royston, K. Moons, and D. Altman, "Development and validation of a prediction model with missing predictor data: a practical approach," *Journal of Clinical Epidemiology*, vol. 63, no. 2, pp. 205–214, 2010. [Online]. Available: WOS:000274062400017

[175] J. Hardt, M. Herke, and R. Leonhart, "Auxiliary variables in multiple imputation in regression with missing x: a warning against including too many in small sample research," *BMC Med Res Methodol*, vol. 12, p. 184, 2012.

[176] E. Kontopantelis, I. White, M. Sperrin, and et al, "Outcome-sensitive multiple imputation: a simulation study," *BMC Med Res Methodol*, vol. 17, no. 2, 2017.

[177] D. Rubin, *Multiple Imputation for Nonresponse in Surveys*, 1987.

[178] A. Marshall, D. Altman, R. Holder, and P. Royston, "Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines," *Bmc Medical Research Methodology*, vol. 9, 2009. [Online]. Available: WOS:000269644400001

[179] P. Royston, K. G. M. Moons, D. G. Altman, and Y. Vergouwe, "Prognosis and prognostic research: Developing a prognostic model," *BMJ*, vol. 338, p. b604, 2009.

[180] K. G. Moons, D. G. Altman, J. B. Reitsma, J. P. Ioannidis, P. Macaskill, E. W. Steyerberg, A. J. Vickers, D. F. Ransohoff, and G. S. Collins, "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): explanation and elaboration," *Ann Intern Med*, vol. 162, no. 1, pp. W1–73, 2015. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/25560730

[181] R. D. Riley and G. S. Collins, "Stability of clinical prediction models developed using statistical or machine learning methods," *Biom J*, p. e2200302, 2023.

[182] R. D. Riley, A. Pate, P. Dhiman, L. Archer, G. P. Martin, and G. S. Collins, "Clinical prediction models and the multiverse of madness," *BMC Medicine*, vol. 21, no. 1, p. 502, 2023. [Online]. Available: https://doi.org/10.1186/s12916-023-03212-y

[183] T. da Silva, P. Macaskill, K. Mills, and et al, "Predicting recovery in patients with acute low back pain: A clinical prediction model," *Eur J Pain*, vol. 21, no. 4, p. 716-726, 2017.

[184] B. L. Myhrvold, A. Kongsted, P. Irgens, H. S. Robinson, M. Thoresen, and N. K. Vøllestad, "Broad external validation and update of a prediction model for persistent neck pain after 12 weeks," *SPINE*, vol. 44, no. 22, pp. E1298–E1310, 2019.

[185] R. Burgess, G. Mansell, A. Bishop, M. Lewis, and J. Hill, "Predictors of functional outcome in musculoskeletal healthcare: An umbrella review," *European Journal of Pain*, vol. 24, no. 1, pp. 51–70, 2019.

[186] R. D. Riley, D. A. Van Der Windt, P. Croft, and K. G. M. Moons, *Prognosis Research in Healthcare: Concepts, Methods, and Impact*. Oxford, UK: Oxford University Press, 2019.

[187] G. S. Collins, J. A. de Groot, S. Dutton, O. Omar, M. Shanyinde, A. Tajar, M. Voysey, R. Wharton, L. M. Yu, K. G. Moons, and D. G. Altman, "External validation of multivariable prediction models: a systematic review of methodological conduct and reporting," *BMC Med Res Methodol*, vol. 14, p. 40, 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24645774

[188] P. C. Austin and E. W. Steyerberg, "Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers," *Stat Med*, vol. 33, no. 3, pp. 517–35, 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/24002997

[189] J. B. Copas, "Using regression models for prediction: shrinkage and regression to the mean," *Stat Methods Med Res*, vol. 6, no. 2, pp. 167–83, 1997.

[190] C. Stein, "Inadmissibility of the usual estimator of the mean of a multivariate normal distribution," *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1956.

[191] J. Kirchner, "Data analysis toolkit n10: Simple linear regression," 1996.

[192] L. Tan, "Confidence intervals for comparison of the squared multiple correlation coefficients of non-nested models," *Electronic Thesis and Dissertation Repository (Paper 384)*, 2012.

[193] J. Wishart, "The mean and second moment coefficient of the multiple correlation coefficient in samples from a normal population," *Biometrika*, vol. 22, pp. 353–361, 1931.

[194] K. Kelley, "Confidence intervals for standardized effect sizes: Theory, application, and implementation," *Journal of Statistical Software*, vol. 20, no. 8, pp. 1 – 24, 2007. [Online]. Available: https://www.jstatsoft.org/index.php/jss/article/view/v020i08

[195] ——, "Methods for the behavioral, educational, and social sciences: an r package," *Behav Res Methods*, vol. 39, no. 4, pp. 979–84, 2007.

[196] ——, "Mbess (version 4.0.0 and higher) [computer software and manual]," vol. Accessible from https://CRAN.R-project.org/package=MBESS, 2017.

[197] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis (Third Edition)*. New York: Wiley, 2001.

[198] F. Harrell, *Regression Modeling Strategies with Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York: Springer, 2001.

[199] A. Boyd, J. Golding, J. Macleod, D. A. Lawlor, A. Fraser, J. Henderson, L. Molloy, A. Ness, S. Ring, and G. Davey Smith, "Cohort profile: The 'children of the 90s'—the index offspring of the avon longitudinal study of parents and children," *International Journal of Epidemiology*, vol. 42, no. 1, pp. 111–127, 2012. [Online]. Available: https://doi.org/10.1093/ije/dys064

[200] A. Fraser, C. Macdonald-Wallis, K. Tilling, A. Boyd, J. Golding, G. Davey Smith, and et al, "Cohort profile: the avon longitudinal study of parents and children: Alspac mothers cohort." *Int J Epidemiol*, vol. 42, pp. 97–110, 2013.

[201] X. Wan, W. Wang, J. Liu, and T. Tong, "Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range," *BMC Medical Research Methodology*, vol. 14, no. 1, p. 135, 2014. [Online]. Available: https://doi.org/10.1186/1471-2288-14-135

[202] K. G. M. Moons, R. F. Wolff, R. D. Riley, P. F. Whiting, M. Westwood, G. S. Collins, J. B. Reitsma, J. Kleijnen, and S. Mallett, "Probast: A tool to assess risk of bias and applicability of prediction model studies: Explanation and elaboration," *Ann Intern Med*, vol. 170, no. 1, pp. W1–W33, 2019. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/30596876

[203] M. T. Hudda, J. C. K. Wells, L. S. Adair, J. R. A. Alvero-Cruz, M. N. Ashby-Thompson, M. N. Ballesteros-Vásquez, J. Barrera-Exposito, B. Caballero, E. A. Carnero, G. J. Cleghorn, P. S. W. Davies, M. Desmond, D. Devakumar, D. Gallagher, E. V. Guerrero-Alcocer, F. Haschke, M. Horlick, H. B. Jemaa, A. I. Khan, A. Mankai, M. A. Monyeki, H. L. Nashandi, L. Ortiz-Hernandez, G. Plasqui, F. F. Reichert, A. E. Robles-Sardin, E. Rush, R. J. Shypailo, J. G. Sobiecki, G. A. t. Hoor, J. Valdés, V. P. Wickramasinghe, W. W. Wong, R. D. Riley, C. G. Owen, P. H. Whincup, and C. M. Nightingale, "External

validation of a prediction model for estimating fat mass in children and adolescents in 19 countries: individual participant data meta-analysis," *BMJ*, vol. 378, p. e071185, 2022. [Online]. Available: https://www.bmj.com/content/bmj/378/bmj-2022-071185.full.pdf

[204] K. I. Snell, J. Ensor, T. P. Debray, K. G. Moons, and R. D. Riley, "Meta-analysis of prediction model performance across multiple studies: Which scale helps ensure between-study normality for the c-statistic and calibration measures?" *Statistical Methods in Medical Research*, vol. 27, no. 11, pp. 3505–3522, 2018. [Online]. Available: https://journals.sagepub.com/doi/abs/10.1177/0962280217705678

[205] K. I. E. Snell, J. Allotey, M. Smuk, R. Hooper, C. Chan, A. Ahmed, L. C. Chappell, P. Von Dadelszen, M. Green, L. Kenny, A. Khalil, K. S. Khan, B. W. Mol, J. Myers, L. Poston, B. Thilaganathan, A. C. Staff, G. C. S. Smith, W. Ganzevoort, H. Laivuori, A. O. Odibo, J. Arenas Ramírez, J. Kingdom, G. Daskalakis, D. Farrar, A. A. Baschat, P. T. Seed, F. Prefumo, F. da Silva Costa, H. Groen, F. Audibert, J. Masse, R. B. Skråstad, K. Salvesen, C. Haavaldsen, C. Nagata, A. R. Rumbold, S. Heinonen, L. M. Askie, L. J. M. Smits, C. A. Vinter, P. Magnus, K. Eero, P. M. Villa, A. K. Jenum, L. B. Andersen, J. E. Norman, A. Ohkuchi, A. Eskild, S. Bhattacharya, F. M. McAuliffe, A. Galindo, I. Herraiz, L. Carbillon, K. Klipstein-Grobusch, S. A. Yeo, J. L. Browne, K. G. M. Moons, R. D. Riley, and S. Thangaratinam, "External validation of prognostic models predicting pre-eclampsia: individual participant data meta-analysis," *BMC Med*, vol. 18, no. 1, p. 302, 2020.

[206] J. Allotey, R. Whittle, K. I. E. Snell, M. Smuk, R. Townsend, P. von Dadelszen, A. E. P. Heazell, L. Magee, G. C. S. Smith, J. Sandall, B. Thilaganathan, J. Zamora, R. D. Riley, A. Khalil, and S. Thangaratinam, "External validation of prognostic models to predict stillbirth using international prediction of pregnancy complications (ippic) network database:

individual participant data meta-analysis," *Ultrasound Obstet Gynecol*, vol. 59, no. 2, pp. 209–219, 2022.

[207] A. J. Vickers, B. van Calster, and E. W. Steyerberg, "A simple, step-by-step guide to interpreting decision curve analysis," *Diagn Progn Res*, vol. 3, p. 18, 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/31592444

[208] L. Poon and et al, "Reference range of birth weight with gestation and first-trimester prediction of small-for-gestation neonates," *Prenat Diagn*, vol. 31, no. 1, pp. 58–65, 2011.

[209] S. Jolani, T. P. Debray, H. Koffijberg, S. van Buuren, and K. G. Moons, "Imputation of systematically missing predictors in an individual participant data meta-analysis: a generalized approach using mice," *Stat Med*, vol. 34, no. 11, pp. 1841–63, 2015. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/25663182

[210] M. Resche-Rigon and I. R. White, "Multiple imputation by chained equations for systematically and sporadically missing multilevel data," *Stat Methods Med Res*, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/27647809

[211] R. Riley, J. Tierney, and L. E. Stewart, *Individual Participant Data Meta-Analysis: A Handbook for Healthcare Research*. Chicester: Wiley, 2021.

[212] T. R. Sullivan, A. B. Salter, P. Ryan, and K. J. Lee, "Bias and precision of the "multiple imputation, then deletion" method for dealing with missing outcome data," *Am J Epidemiol*, vol. 182, no. 6, pp. 528–34, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/26337075

[213] S. Burgess, I. R. White, M. Resche-Rigon, and A. M. Wood, "Combining multiple imputation and meta-analysis with individual participant data," *Stat Med*, vol. 32, p. 4499–4514, 2013.

[214] J. Higgins, S. Thompson, and D. Spiegelhalter, "A re-evaluation of random-effects meta-analysis," *Journal of the Royal Statistical Society Series A-Statistics in Society*, vol. 172, pp. 137–159, 2009. [Online]. Available: Go to ISI://000261962600009

[215] R. Riley, J. Higgins, and J. Deeks, "Interpretation of random effects meta-analyses," *British Medical Journal*, vol. 342, pp. 964–967, 2011.

[216] D. Burke, J. Ensor, and R. Riley, "Meta-analysis using individual participant data: one-stage and two-stage approaches, and why they may differ," *Stat Med*, vol. 36, no. 5, pp. 855–875, 2017.

[217] C. Röver, G. Knapp, and T. Friede, "Hartung-knapp-sidik-jonkman approach and its modification for random-effects meta-analysis with few studies," *BMC Medical Research Methodology*, vol. 15, pp. 99–105, 2015.

[218] A. Jenum, L. Sletner, N. Voldner, S. Vangen, K. Morkrid, L. Andersen, and et al, "The stork groruddalen research programme: A population-based cohort study of gestational diabetes, physical activity, and obesity in pregnancy in a multiethnic population. rationale, methods, study population, and participation rates," *Scand J Public Health*, vol. 38, pp. 60–70, 2010.

[219] R. Allen, J. Zamora, D. Arroyo-Manzano, L. Velauthar, J. Allotey, S. Thangaratinam, and et al, "External validation of preexisting first trimester preeclampsia prediction models," *Eur J Obstet Gynecol Reprod Biol*, vol. 217, pp. 119–25, 2017.

[220] A. Odibo, Y. Zhong, K. Goetzinger, L. Odibo, J. Bick, C. Bower, and et al, "First-trimester placental protein 13, papp-a, uterine artery doppler and maternal characteristics in the prediction of pre-eclampsia," *Placenta*, vol. 32, pp. 598–602, 2011.

[221] A. Baschat, L. Magder, L. Doyle, R. Atlas, C. Jenkins, and M. Blitzer, "Prediction of preeclampsia utilizing the first trimester screening examination," *Am J Obstet Gynecol*, vol. 211:514, pp. e1–7, 2014.

[222] A. Rumbold, C. Crowther, R. Haslam, G. Dekker, J. Robinson, and A. Group, "Vitamins c and e and the risks of preeclampsia and perinatal complications," *N Engl J Med*, vol. 354, pp. 1796–806, 2006.

[223] U. Sovio, I. White, A. Dacey, D. Pasupathy, and G. Smith, "Screening for fetal growth restriction with universal third trimester ultrasonography in nulliparous women in the pregnancy outcome prediction (pop) study: a prospective cohort study," *The Lancet*, vol. 386, pp. 2089–97, 2015.

[224] V. Jaddoe, C. van Duijn, O. Franco, A. van der Heijden, M. van Iizendoorn, J. de Jongste, and et al, "The generation r study: design and cohort update 2012," *Eur J Epidemiol*, vol. 27, pp. 739–56, 2012.

[225] "Japan society of obstetrics and gynecology." [Online]. Available: http://www.jsog.or.jp/

[226] N. H. S. Digital, "Health survey for england 2019," 2018. [Online]. Available: https://digital.nhs.uk/data-and-information/publications/statistical/health-survey-for-england/health-survey-for-england-2016

[227] D. Ettehad, C. A. Emdin, A. Kiran, S. G. Anderson, T. Callender, J. Emberson, J. Chalmers, A. Rodgers, and K. Rahimi, "Blood pressure lowering for prevention of cardiovascular disease and death: a systematic review and meta-analysis," *Lancet*, vol. 387, no. 10022, pp. 957–967, 2016.

[228] L. J. Seppala, E. M. M. van de Glind, J. G. Daams, K. J. Ploegmakers, M. de Vries, A. Wermelink, and N. van der Velde, "Fall-risk-increasing drugs: A systematic review and meta-analysis: Iii. others," *J Am Med Dir Assoc*, vol. 19, no. 4, pp. 372.e1–372.e8, 2018.

[229] R. M. Leipzig, R. G. Cumming, and M. E. Tinetti, "Drugs and falls in older people: a systematic review and meta-analysis: Ii. cardiac and analgesic drugs," *J Am Geriatr Soc*, vol. 47, no. 1, pp. 40–50, 1999.

[230] P. K. Whelton, R. M. Carey, W. S. Aronow, J. Casey, D. E., K. J. Collins, C. Dennison Himmelfarb, S. M. DePalma, S. Gidding, K. A. Jamerson, D. W. Jones, E. J. MacLaughlin, P. Muntner, B. Ovbiagele, J. Smith, S. C., C. C. Spencer, R. S. Stafford, S. J. Taler, R. J. Thomas, S. Williams, K. A., J. D. Williamson, and J. Wright, J. T., "2017 acc/aha/aapa/abc/acpm/ags/apha/ash/aspc/nma/pcna guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: A report of the american college of cardiology/american heart association task force on clinical practice guidelines," *J Am Coll Cardiol*, vol. 71, no. 19, pp. e127–e248, 2018.

[231] J. P. Fine and R. J. Gray, "A proportional hazards model for the subdistribution of a competing risk," *Journal of the American Statistical Association*, vol. 94, no. 446, pp. 496–509, 1999. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/01621459.1999.10474144

[232] A. Wolf, D. Dedman, J. Campbell, H. Booth, D. Lunn, J. Chapman, and P. Myles, "Data resource profile: Clinical practice research datalink (cprd) aurum," *Int J Epidemiol*, vol. 48, no. 6, pp. 1740–1740g, 2019.

[233] D. Kleinbaum and M. Klein, *Competing risks survival analysis*, 2nd ed. New York, NY: Springer, 2005, book section 9, pp. 391–461.

[234] M. Wolbers, M. T. Koller, J. C. Witteman, and E. W. Steyerberg, "Prognostic models with competing risks: methods and application to coronary risk prediction," *Epidemiology*, vol. 20, no. 4, pp. 555–61, 2009. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/19367167

[235] M. T. Koller, B. Schaer, M. Wolbers, C. Sticherling, H. C. Bucher, and S. Osswald, "Death without prior appropriate implantable cardioverter-defibrillator therapy: a competing risk study," *Circulation*, vol. 117, no. 15, pp. 1918–26, 2008.

[236] P. K. Andersen, R. B. Geskus, T. de Witte, and H. Putter, "Competing risks in epidemiology: possibilities and pitfalls," *Int J Epidemiol*, vol. 41, no. 3, pp. 861–70, 2012.

[237] M. Wolbers, M. T. Koller, V. S. Stel, B. Schaer, K. J. Jager, K. Leffondré, and G. Heinze, "Competing risks analyses: objectives and approaches," *Eur Heart J*, vol. 35, no. 42, pp. 2936–41, 2014.

[238] G. L. Grunkemeier, R. Jin, M. J. Eijkemans, and J. J. Takkenberg, "Actual and actuarial probabilities of competing risks: apples and lemons," *Ann Thorac Surg*, vol. 83, no. 5, pp. 1586–92, 2007.

[239] H. Putter, M. Fiocco, and R. B. Geskus, "Tutorial in biostatistics: competing risks and multi-state models," *Stat Med*, vol. 26, no. 11, pp. 2389–430, 2007.

[240] M. J. Zhang and J. Fine, "Summarizing differences in cumulative incidence functions," *Stat Med*, vol. 27, no. 24, pp. 4939–49, 2008.

[241] P. C. Austin, E. W. Steyerberg, and H. Putter, "Fine-gray subdistribution hazard models to simultaneously estimate the absolute risk of different event types: Cumulative total failure probability may exceed 1," *Stat Med*, vol. 40, no. 19, pp. 4200–4212, 2021.

[242] P. Andersen and M. Perme, "Pseudo-observations in survival analysis," *Stat Methods Med Res*, vol. 19, no. 1, pp. 71–99, 2010.

[243] P. K. Andersen, J. P. Klein, and S. Rosthøj, "Generalised linear models for correlated pseudo-observations, with applications to multi-state models," *Biometrika*, vol. 90, no. 1, pp. 15–27, 2003. [Online]. Available: https://doi.org/10.1093/biomet/90.1.15

[244] P. K. Andersen and H. Ravn, *Models for Multi-State Survival Data: Rates, Risks, and Pseudo-Values*. CRC Press, 2023.

[245] J. P. Klein and P. K. Andersen, "Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function," *Biometrics*, vol. 61, no. 1, pp. 223–9, 2005.

[246] M. J. Zhang, X. Zhang, and T. H. Scheike, "Modeling cumulative incidence function for competing risks data," *Expert Rev Clin Pharmacol*, vol. 1, no. 3, pp. 391–400, 2008.

[247] M. Overgaard, E. T. Parner, and J. Pedersen, "Pseudo-observations under covariate-dependent censoring," *Journal of Statistical Planning and Inference*, vol. 202, pp. 112–122, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378375819300175

[248] E. T. Parner, P. K. Andersen, and M. Overgaard, "Regression models for censored time-to-event data using infinitesimal jack-knife pseudo-observations, with applications to left-truncation," *Lifetime Data Anal*, vol. 29, no. 3, pp. 654–671, 2023.

[249] I. White and P. Royston, "Imputing missing covariate values for the cox model," *Statistics in Medicine*, vol. 28, no. 15, pp. 1982–1998, 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19452569

[250] B. Laua and C. Leskoa, "Missingness in the setting of competing risks: from missing values to missing potential outcomes," *Curr Epidemiol Rep*, vol. 5, no. 2, p. 153–159, 2018.

[251] F. Graw, T. Gerds, and M. Schumacher, "On pseudo-values for regression analysis in competing risks models," *Lifetime Data Anal*, vol. 15, no. 2, pp. 241–55, 2009.

[252] P. Blanche, M. W. Kattan, and T. A. Gerds, "The c-index is not proper for the evaluation of t-year predicted risks," *Biostatistics.*, vol. 20, no. 2, p. 347–357, 2019.

[253] P. Royston and W. Sauerbrei, "A new measure of prognostic separation in survival data," *Statistics in Medicine*, vol. 23, no. 5, pp. 723–748, 2004. [Online]. Available: WOS:000189296500005

[254] J. Beyersmann, A. Latouche, A. Buchholz, and M. Schumacher, "Simulating competing risks data in survival analysis," *Stat Med*, vol. 28, no. 6, pp. 956–71, 2009.

[255] H. C. van Houwelingen and H. Putter, "Dynamic predicting by landmarking as an alternative for multi-state modeling: an application to acute lymphoid leukemia data," *Lifetime Data Anal*, vol. 14, no. 4, pp. 447–63, 2008.

[256] D. Toll, K. Janssen, Y. Vergouwe, and K. Moons, "Validation, updating and impact of clinical prediction rules: A review," *Journal of Clinical Epidemiology*, vol. 61, no. 11, pp. 1085–1094, 2008. [Online]. Available: WOS:000259955800003

[257] G. S. Collins, P. Dhiman, J. Ma, M. M. Schlussel, L. Archer, B. Van Calster, F. E. Harrell, G. P. Martin, K. G. M. Moons, M. van Smeden, M. Sperrin, G. S. Bullock, and R. D. Riley, "Evaluation of clinical prediction models (part 1): from development to external validation," *BMJ*, vol. 384, p. e074819, 2024. [Online]. Available: https://www.bmj.com/content/bmj/384/bmj-2023-074819.full.pdf

[258] R. D. Riley, L. Archer, K. I. E. Snell, J. Ensor, P. Dhiman, G. P. Martin, L. J. Bonnett, and G. S. Collins, "Evaluation of clinical prediction models (part 2): how to undertake an external validation study," *BMJ*, vol. 384, p. e074820, 2024. [Online]. Available: https://www.bmj.com/content/bmj/384/bmj-2023-074820.full.pdf

[259] L. Archer, G. Peat, K. I. E. Snell, J. C. Hill, K. M. Dunn, N. E. Foster, A. Bishop, D. van der Windt, and G. Wynne-Jones, "Musculoskeletal health and work: Development and internal–external cross-validation of a model to predict risk of work absence and presenteeism in people seeking primary healthcare," *Journal of Occupational Rehabilitation*, 2024. [Online]. Available: https://doi.org/10.1007/s10926-024-10223-w

[260] J. Ensor, "Pmvalsampsize: Stata module to calculate the minimum sample size required for external validation of a multivariable prediction model," 2023. [Online]. Available: https://EconPapers.repec.org/RePEc:boc:bocode:s459226

[261] H. van Houwelingen, "Validation, calibration, revision and combination of prognostic survival models," *Statistics in Medicine*, vol. 19, no. 24, pp. 3401–3415, 2000. [Online]. Available: Go to ISI://000166395500009

[262] K. Janssen, K. Moons, C. Kalkman, D. Grobbee, and Y. Vergouwe, "Updating methods improved the performance of a clinical prediction model in new patients," *Journal of Clinical Epidemiology*, vol. 61, no. 1, pp. 76–86, 2008. [Online]. Available: WOS:000252044900009

[263] E. Herrett, A. M. Gallagher, K. Bhaskaran, H. Forbes, R. Mathur, T. van Staa, and L. Smeeth, "Data resource profile: Clinical practice research datalink (cprd)," *Int J Epidemiol*, vol. 44, no. 3, pp. 827–36, 2015.

[264] A. Wolf, D. Dedman, J. Campbell, H. Booth, D. Lunn, J. Chapman, and P. Myles, "Data resource profile: Clinical practice research datalink (cprd) aurum," *Int J Epidemiol*, vol. 48, no. 6, pp. 1740–1740g, 2019.

[265] J. Hippisley-Cox, C. Coupland, and P. Brindle, "Development and validation of qrisk3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study," *BMJ*, vol. 357, p. j2099, 2017.

[266] J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman, "Distribution-free predictive inference for regression," *Journal of the American Statistical Association*, vol. 113, no. 523, pp. 1094–1111, 2018. [Online]. Available: https://doi.org/10.1080/01621459.2017.1307116

[267] J. Ensor, "pmvalsampsize: Sample size for external validation of a prediction model," 2023. [Online]. Available: https://github.com/JoieEnsor/pmvalsampsize

[268] T. P. Debray, K. G. Moons, I. Ahmed, H. Koffijberg, and R. D. Riley, "A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis," *Statistics in Medicine*, vol. 32, no. 18, pp. 3158–3180, 2013. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.5732

[269] P. Royston, M. K. Parmar, and R. Sylvester, "Construction and validation of a prognostic model across several studies, with an application in superficial bladder cancer," *Stat Med*, vol. 23, no. 6, pp. 907–26, 2004.

APPENDICES

8 Appendices

Appendix I - Chapter 2, model characteristics summary

Table 8.1: Table of study characteristics

Model reference	Publication	Study Title	Outcome definition	Type of Predictors	Type of model
1	Ciobanu 2019a [117]	Prediction of small for gestational age neonates: screening by maternal factors, fetal biometry, and biomarkers at 35–37 weeks' gestation	SGA <10th Percentile	UA-PI, MCA-PI, MA, PiGF, sFLT	Logistic regression
2	Ciobanu 2019a [117]	Prediction of small for gestational age neonates: screening by maternal factors, fetal biometry, and biomarkers at 35–37 weeks' gestation	SGA <3rd percentile	UA-PI, MCA-PI, MA, PiGF, sFLT	Logistic regression
3	Ciobanu 2019a [117]	Prediction of small for gestational age neonates: screening by maternal factors, fetal biometry, and biomarkers at 35–37 weeks' gestation	SGA <10th Percentile	UA-PI, MCA-PI, MA, PiGF, sFLT	Logistic regression
4	Ciobanu 2019b [118]	Prediction of small-for-gestational-age neonates at 35–37 weeks' gestation: contribution of maternal factors and growth velocity between 20 and 36 weeks	SGA (within two weeks) <10th percentile	Maternal risk factors, EFW score(HC, AC, FL), AC growth velocity	Logistic regression
5	Ciobanu 2019b [118]	Prediction of small-for-gestational-age neonates at 35–37 weeks' gestation: contribution of maternal factors and growth velocity between 20 and 36 weeks	SGA (within two weeks) <3rd percentile	Maternal risk factors, EFW score(HC, AC, FL), AC growth velocity	Logistic regression
6	Ciobanu 2019b [118]	Prediction of small-for-gestational-age neonates at 35–37 weeks' gestation: contribution of maternal factors and growth velocity between 20 and 36 weeks	SGA <10th percentile	Maternal risk factors, EFW score(HC, AC, FL), AC growth velocity	Logistic regression
7	Ciobanu 2019b [118]	Prediction of small-for-gestational-age neonates at 35–37 weeks' gestation: contribution of maternal factors and growth velocity between 20 and 36 weeks	SGA <3rd percentile	Maternal risk factors, EFW score(HC, AC, FL), AC growth velocity	Logistic regression
8	Sharp 2019 [114]	A prediction model for short-term neonatal outcomes in severe early-onset fetal growth restriction	Severe early onset FGR	Maternal demographics, fetal biometric measurements, doppler measurements, maternal angiogenic markers	Linear regression
9	Sotiriadis 2019 [119]	First-trimester and combined first- and second-trimester prediction of small-for-gestational age and late fetal growth restriction	Overall late FGR	Conception method, smoking status, maternal height, PAPP-A and UtA- PI or CPR	Logistic regression
10	Sotiriadis 2019 [120]	First-trimester and combined first- and second-trimester prediction of small-for-gestational age and late fetal growth restriction	prenatally detected late FGR	Conception method, smoking status, maternal height, PAPP-A and UtA- PI or CPR	Logistic regression

Model reference	Publication	Study Title	Outcome definition	Type of Predictors	Type of model
11	Sotiriadis 2019 [113]	First-trimester and combined first- and second-trimester prediction of small-for-gestational age and late fetal growth restriction	SGA	Conception method, smoking status, maternal height, PAPP-A and UtA- PI or CPR	Logistic regression
12	Sotiriadis 2019 [121]	First-trimester and combined first- and second-trimester prediction of small-for-gestational age and late fetal growth restriction	Overall late FGR	Conception method, smoking status, maternal height, PAPP-A and UtA- PI, second trimester EFW, HC/AC ratio a	Logistic regression
13	Sotiriadis 2019 [115]	First-trimester and combined first- and second-trimester prediction of small-for-gestational age and late fetal growth restriction	prenatally detected FGR	Conception method, smoking status, maternal height, PAPP-A and UtA- PI, second trimester EFW, HC/AC ratio a	Logistic regression
14	Sotiriadis 2019 [122]	First-trimester and combined first- and second-trimester prediction of small-for-gestational age and late fetal growth restriction	SGA	Conception method, smoking status, maternal height, PAPP-A and UtA- PI, second trimester EFW, HC/AC ratio a	Logistic regression
15	Sharp 2019 [114]	A prediction model for short-term neonatal outcomes in severe early-onset fetal growth restriction	Birthweight	EFW, Previous pregnancy, gestation at randomisation, sENG, SFlt	Linear regression
16	Hendrix 2018 [123]	Can Fetal Growth Velocity and First Trimester Maternal Biomarkers Improve the Prediction of Small-for-Gestational Age and Adverse Neonatal Outcome?	SGA	PAPP- A, B-jcG, PiGF, SFlt, AC, Biparietal diameter (BPD), HC, FL	ANOVA
17	Anggraini 2018 [124]	Foetal weight prediction models at a given gestational age in the absence of ultrasound facilities: application in Indonesia, 2018	Birthweight	GA, fundal height, foetal head engagement/foetal station, EFW, Actual birth weight, neonatal head circumference (HC) and neonatal abdominal circumference (AC)	Linear regression
18	Miranda 2017 [125]	Prediction of fetal growth restriction using estimated fetal weight vs a combined screening model in the third trimester	SGA <10th percentile	A priori maternal risk, EFWc, PiGF, Lipocalin-2, Estriol, Map, Uta-Pi, CPR	Logistic regression
19	Miranda 2017 [125]	Prediction of fetal growth restriction using estimated fetal weight vs a combined screening model in the third trimester	SGA <10th percentile	A priori maternal risk, EFWc, PiGF, Lipocalin-2, Estriol	Logistic regression

Model reference	Publication	Study Title	Outcome definition	Type of Predictors	Type of model
20	Allen 2017 [126]	Prospective observational study to determine the accuracy of first-trimester serum biomarkers and uterine artery Dopplers in combination with maternal characteristics and arteriography for the prediction of women at risk of preeclampsia and other adverse pregnancy outcomes	SGA < 5th percentile	Uterine artery doppler, PiGF, AFP, PAPP-A and B-HCG	Logistic regression
21	Allen 2017 [126]	Prospective observational study to determine the accuracy of first-trimester serum biomarkers and uterine artery Dopplers in combination with maternal characteristics and arteriography for the prediction of women at risk of preeclampsia and other adverse pregnancy outcomes	SGA <10th percentile	Uterine artery doppler, PiGF, AFP, PAPP-A and B-HCG	Logistic regression
22	McCowan 2017 [127]	Prediction of Small for Gestational Age Infants in Healthy Nulliparous Women Using Clinical and Ultrasound Risk Factors Combined with Early Pregnancy Biomarkers.	SGA <10th customised centile	Clinical risk factors, biomarkers and ultrasound data	Logistic regression
23	McCowan 2017 [127]	Prediction of Small for Gestational Age Infants in Healthy Nulliparous Women Using Clinical and Ultrasound Risk Factors Combined with Early Pregnancy Biomarkers.	Normosensitive SGA (SGA with a normosensitive mother)	Clinical risk factors, biomarkers and ultrasound data	Logistic regression
24	McCowan 2017 [127]	Prediction of Small for Gestational Age Infants in Healthy Nulliparous Women Using Clinical and Ultrasound Risk Factors Combined with Early Pregnancy Biomarkers.	Hypersensitive SGA (SGA with a hypersensitive mother)	Clinical risk factors, biomarkers and ultrasound data	Logistic regression
25	Kim 2017 [128]	Maternal serum placental growth factor combined with second trimester aneuploidy screening to predict small-for-gestation neonates without preeclampsia	SGA	E3, HCG, PlGF	Mann whitney and logistic regression
26	Litwinska 2017 [129]	Combined screening for early and late pre-eclampsia and intrauterine growth restriction by maternal history, uterine artery Doppler, mean arterial pressure and biochemical markers	IUGR	maternal risk factors, UtA-PI and PlGF concentrations	Logistic regression
27	González González 2017 [130]	Value of placental volume and vascular flow indices as predictors of intrauterine growth retardation	IUGR	MC, FTSA, Uterine Artery PI, Placental volume, vascular indicies	Linear regression

Model reference	Publication	Study Title	Outcome definition	Type of Predictors	Type of model
28	González González 2017 [130]	Value of placental volume and vascular flow indices as predictors of intrauterine growth retardation	IUGR	MC, FTSA, Uterine Artery PI, Placental volume, vascular indicies	Logistic regression
29	Crovetto 2016 [131]	Differential performance of first-trimester screening in predicting small-for-gestational-age neonate or fetal growth restriction.	SGA		Logistic regression
30	Crovetto 2016a [132]	First-trimester screening with specific algorithms for early- and late-onset fetal growth restriction.	IUGR		Logistic regression
31	Crovetto 2016a [132]	First-trimester screening with specific algorithms for early- and late-onset fetal growth restriction.	early IUGR		Logistic regression
32	Crovetto 2016a [132]	First-trimester screening with specific algorithms for early- and late-onset fetal growth restriction.	late IUGR		Logistic regression
33	Lesmes 2015a [133]	Prediction of small-for-gestational-age neonates: screening by maternal serum biochemical markers at 19–24 weeks	SGA	Maternal Factors, Z scores for fetal head circumference, abdominal circumference, femur length, concentrations of placenta growth factor serums (PIGF), soluble fms-like tyrosine kinase-1 (sFlt-1), pregnancy-associated plasma protein-A (PAPP-A), freeβ-human chorionicgonadotropin (β-hCG) and α-fetoprotein (AFP)	Logistic regression
34	Lesmes 2015a [133]	Prediction of small-for-gestational-age neonates: screening by maternal serum biochemical markers at 19–24 weeks	SGA <10th percentile <32 weeks	Maternal Factors, Z scores for fetal head circumference, abdominal circumference, femur length, concentrations of placenta growth factor serums (PIGF), soluble fms-like tyrosine kinase-1 (sFlt-1), pregnancy-associated plasma protein-A (PAPP-A), freeβ-human chorionicgonadotropin (β-hCG) and α-fetoprotein (AFP)	Logistic regression

Model reference	Publication	Study Title	Outcome definition	Type of Predictors	Type of model
35	Lesmes 2015a [133]	Prediction of small-for-gestational-age neonates: screening by maternal serum biochemical markers at 19–24 weeks	SGA <10th percentile 32- 36 weeks	Maternal Factors, Z scores for fetal head circumference, abdominal circumference, femur length, concentrations of placenta growth factor serums (PIGF), soluble fms-like tyrosine kinase-1 (sFlt-1), pregnancy-associated plasma protein-A (PAPP-A), freeβ-human chorionicgonadotropin (β-hCG) and α-fetoprotein (AFP)	Logistic regression
36	Lesmes 2015a [133]	Prediction of small-for-gestational-age neonates: screening by maternal serum biochemical markers at 19–24 weeks	SGA <10th percentile ¿37weeks	Maternal Factors, Z scores for fetal head circumference, abdominal circumference, femur length, concentrations of placenta growth factor serums (PIGF), soluble fms-like tyrosine kinase-1 (sFlt-1), pregnancy-associated plasma protein-A (PAPP-A), freeβ-human chorionicgonadotropin (β-hCG) and α-fetoprotein (AFP)	Logistic regression
37	Lesmes 2015a [133]	Prediction of small-for-gestational-age neonates: screening by maternal serum biochemical markers at 19–24 weeks	SGA <5th percentile <32 weeks	Maternal Factors, Z scores for fetal head circumference, abdominal circumference, femur length, concentrations of placenta growth factor serums (PIGF), soluble fms-like tyrosine kinase-1 (sFlt-1), pregnancy-associated plasma protein-A (PAPP-A), freeβ-human chorionicgonadotropin (β-hCG) and α-fetoprotein (AFP)	Logistic regression

Model reference	Publication	Study Title	Outcome definition	Type of Predictors	Type of model
38	Lesmes 2015a [133]	Prediction of small-for-gestational-age neonates: screening by maternal serum biochemical markers at 19–24 weeks	SGA <5th percentile 32- 36 weeks	Maternal Factors, Z scores for fetal head circumference, abdominal circumference, femur length, concentrations of placenta growth factor serums (PIGF), soluble fms-like tyrosine kinase-1 (sFlt-1), pregnancy-associated plasma protein-A (PAPP-A), freeβ-human chorionicgonadotropin (β-hCG) and α-fetoprotein (AFP)	Logistic regression
39	Lesmes 2015a [133]	Prediction of small-for-gestational-age neonates: screening by maternal serum biochemical markers at 19–24 weeks	SGA < 5th percentile ¿37 weeks	Maternal Factors, Z scores for fetal head circumference, abdominal circumference, femur length, concentrations of placenta growth factor serums (PIGF), soluble fms-like tyrosine kinase-1 (sFlt-1), pregnancy-associated plasma protein-A (PAPP-A), freeβ-human chorionicgonadotropin (β-hCG) and α-fetoprotein (AFP)	Logistic regression
40	Lesmes 2015a [133]	Prediction of small-for-gestational-age neonates: screening by maternal serum biochemical markers at 19–24 weeks	SGA <3rd percentile <32 weeks	Maternal Factors, Z scores for fetal head circumference, abdominal circumference, femur length, concentrations of placenta growth factor serums (PIGF), soluble fms-like tyrosine kinase-1 (sFlt-1), pregnancy-associated plasma protein-A (PAPP-A), freeβ-human chorionicgonadotropin (β-hCG) and α-fetoprotein (AFP)	Logistic regression

Model reference	Publication	Study Title	Outcome definition	Type of Predictors	Type of model
41	Lesmes 2015a [133]	Prediction of small-for-gestational-age neonates: screening by maternal serum biochemical markers at 19–24 weeks	SGA <3rd percentile 32- 36 weeks	Maternal Factors, Z scores for fetal head circumference, abdominal circumference, femur length, concentrations of placenta growth factor serums (PIGF), soluble fms-like tyrosine kinase-1 (sFlt-1), pregnancy-associated plasma protein-A (PAPP-A), freeβ-human chorionicgonadotropin (β-hCG) and α-fetoprotein (AFP)	Logistic regression
42	Lesmes 2015a [133]	Prediction of small-for-gestational-age neonates: screening by maternal serum biochemical markers at 19–24 weeks	SGA < 3rd percentile ¿37 weeks	Maternal Factors, Z scores for fetal head circumference, abdominal circumference, femur length, concentrations of placenta growth factor serums (PIGF), soluble fms-like tyrosine kinase-1 (sFlt-1), pregnancy-associated plasma protein-A (PAPP-A), freeβ-human chorionicgonadotropin (β-hCG) and α-fetoprotein (AFP)	Logistic regression
43	Lesmes 2015b [134]	Prediction of small-for-gestational-age neonates: screening by uterine artery Doppler and mean arterial pressure at 19–24 weeks	SGA	Combination maternal factors, fetal head circumference (HC), Abdominal circumference (AC), Femur Length (FL), Uterine artery pulasatiruty Index (UtA-PI) and Mean Arterial Pressure (MAP)	Logistic regression
44	Lesmes 2015b [134]	Prediction of small-for-gestational-age neonates: screening by uterine artery Doppler and mean arterial pressure at 19–24 weeks	SGA <10th percentile <32 weeks	Combination maternal factors, fetal head circumference (HC), Abdominal circumference (AC), Femur Length (FL), Uterine artery pulasatiruty Index (UtA-PI) and Mean Arterial Pressure (MAP)	Logistic regression

Model reference	Publication	Study Title	Outcome definition	Type of Predictors	Type of model
45	Lesmes 2015b [134]	Prediction of small-for-gestational-age neonates: screening by uterine artery Doppler and mean arterial pressure at 19–24 weeks	SGA <10th percentile 32- 36 weeks	Combination maternal factors, fetal head circumference (HC), Abdominal circumference (AC), Femur Length (FL), Uterine artery pulsatiruty Index (UtA-PI) and Mean Arterial Pressure (MAP)	Logistic regression
46	Lesmes 2015b [134]	Prediction of small-for-gestational-age neonates: screening by uterine artery Doppler and mean arterial pressure at 19–24 weeks	SGA <10th percentile ¿37weeks	Combination maternal factors, fetal head circumference (HC), Abdominal circumference (AC), Femur Length (FL), Uterine artery pulsatiruty Index (UtA-PI) and Mean Arterial Pressure (MAP)	Logistic regression
47	Lesmes 2015b [134]	Prediction of small-for-gestational-age neonates: screening by uterine artery Doppler and mean arterial pressure at 19–24 weeks	SGA <5th percentile <32 weeks	Combination maternal factors, fetal head circumference (HC), Abdominal circumference (AC), Femur Length (FL), Uterine artery pulsatiruty Index (UtA-PI) and Mean Arterial Pressure (MAP)	Logistic regression
48	Lesmes 2015b [134]	Prediction of small-for-gestational-age neonates: screening by uterine artery Doppler and mean arterial pressure at 19–24 weeks	SGA <5th percentile 32- 36 weeks	Combination maternal factors, fetal head circumference (HC), Abdominal circumference (AC), Femur Length (FL), Uterine artery pulsatiruty Index (UtA-PI) and Mean Arterial Pressure (MAP)	Logistic regression
49	Lesmes 2015b [134]	Prediction of small-for-gestational-age neonates: screening by uterine artery Doppler and mean arterial pressure at 19–24 weeks	SGA <5th percentile ¿37 weeks	Combination maternal factors, fetal head circumference (HC), Abdominal circumference (AC), Femur Length (FL), Uterine artery pulsatiruty Index (UtA-PI) and Mean Arterial Pressure (MAP)	Logistic regression

Model reference	Publication	Study Title	Outcome definition	Type of Predictors	Type of model
50	Lesmes 2015b [134]	Prediction of small-for-gestational-age neonates: screening by uterine artery Doppler and mean arterial pressure at 19–24 weeks	SGA <3rd percentile <32 weeks	Combination maternal factors, fetal head circumference (HC), Abdominal circumference (AC), Femur Length (FL), Uterine artery pulsatiruty Index (UtA-PI) and Mean Arterial Pressure (MAP)	Logistic regression
51	Lesmes 2015b [134]	Prediction of small-for-gestational-age neonates: screening by uterine artery Doppler and mean arterial pressure at 19–24 weeks	SGA <3rd percentile 32- 36 weeks	Combination maternal factors, fetal head circumference (HC), Abdominal circumference (AC), Femur Length (FL), Uterine artery pulsatiruty Index (UtA-PI) and Mean Arterial Pressure (MAP)	Logistic regression
52	Lesmes 2015b [134]	Prediction of small-for-gestational-age neonates: screening by uterine artery Doppler and mean arterial pressure at 19–24 weeks	SGA <3rd percentile ¿37 weeks	Combination maternal factors, fetal head circumference (HC), Abdominal circumference (AC), Femur Length (FL), Uterine artery pulsatiruty Index (UtA-PI) and Mean Arterial Pressure (MAP)	Logistic regression
53	Fadigas 2015a [135]	Prediction of small-for-gestational-age neonates: screening by fetal biometry at 35–37 weeks	SGA <5th percentile	Maternal factors, HC, AC, FL and EFW	Logistic regression
54	Fadigas 2015c [136]	Prediction of small-for-gestational-age neonates: screening by uterine artery Doppler and mean arterial pressure at 35–37 weeks	SGA <5th percentile	maternal characteristics, medical history, EFW, Uterine Artery pulsality index (Uta-pi) and MAP	Logistic regression
55	Bakalis 2015a [137]	Prediction of small-for-gestational-age neonates: screening by biophysical and biochemical markers at 30–34 weeks	SGA	Maternal Factors, EFW, UtA-PI, MAP, Placental Growth Factor (PiGF) and sFlt-1	Logistic regression
56	Bakalis 2015b [138]	Prediction of small-for-gestational-age neonates: screening by maternal biochemical markers at 30–34 weeks	SGA <10th Percentile	GiGF, sfLt-1, PAPP-A, B-hCG, AFP	Logistic regression
57	Bakalis 2015b [138]	Prediction of small-for-gestational-age neonates: screening by maternal biochemical markers at 30–34 weeks	SGA <5th percentile	GiGF, sfLt-1, PAPP-A, B-hCG, AFP	Logistic regression
58	Bakalis 2015b [138]	Prediction of small-for-gestational-age neonates: screening by maternal biochemical markers at 30–34 weeks	SGA <3rd percentile	GiGF, sfLt-1, PAPP-A, B-hCG, AFP	Logistic regression

Model reference	Publication	Study Title	Outcome definition	Type of Predictors	Type of model
59	Bakalis 2015d [139]	Prediction of small-for-gestational-age neonates: screening by fetal biometry at 30–34 weeks	SGA delivery <5 weeks after assessment		Logistic regression
60	Kienast 2015 [122]	Predictive value of angiogenic factors, clinical risk factors and uterine artery Doppler for pre-eclampsia and fetal growth restriction in second and third trimester pregnancies in an Ecuadorian population	IUGR		Logistic regression
61	Lesmes 2015c [116]	Prediction of small for gestational age neonates: screening by fetal biometry at 19 - 24 weeks	SGA <5th percentile (delivering <37 weeks)	MC, HC, AC and FL	Logistic regression
62	MacdonaldWallis 2015 [140]	Antenatal blood pressure for prediction of pre-eclampsia, preterm birth, and small for gestational age babies: development and validation in two general population cohorts	PE and its adverse outcomes: pre-term birth and SGA		
63	Papastefanou 2015 [141]	Predicting fetal growth deviation in parous women: combining the birth weight of the previous pregnancy and third trimester ultrasound scan	Birthweight	Conception method, maternal weight, BW z-score (previous and current pregnancy), smoking	Linear regression
64	Seravalli 2014a [142]	First-trimester prediction of small-for-gestational age neonates incorporating fetal Doppler parameters and maternal characteristics	SGA	Maternal risk factors, UtA, fetal Umbilical artery (UA) and ductus venosus (DV)	Logistic regression
65	Seravalli 2014b [143]	Second-trimester prediction of delivery of a small-for-gestational-age neonate: integrating sequential Doppler information, fetal biometry, and maternal characteristics.	SGA < 3rd percentile	Maternal characteristics, UtA, PI, fetal biometry, (UA)-PI	Logistic regression
66	Seravalli 2014b [143]	Second-trimester prediction of delivery of a small-for-gestational-age neonate: integrating sequential Doppler information, fetal biometry, and maternal characteristics.	SGA 3rd -10th percentile	Maternal characteristics, UtA, PI, fetal biometry, (UA)-PI	Logistic regression
67	Schwartz 2014 [144]	First-trimester placental ultrasound and maternal serum markers as predictors of small-for-gestational-age infants	SGA		None
68	Boucoiran 2013 [145]	Risks for Preeclampsia and Small for Gestational Age: Predictive Values of Placental Growth Factor, Soluble fms-like Tyrosine Kinase-1, and Inhibin A in Singleton and Multiple-Gestation Pregnancies	SGA	PiGF, dFlt-1, inhibin A	Wilcoxon and Chi- square

Model reference	Publication	Study Title	Outcome definition	Type of Predictors	Type of model
69	Boucoiran 2013 [145]	Risks for Preeclampsia and Small for Gestational Age: Predictive Values of Placental Growth Factor, Soluble fms-like Tyrosine Kinase-1, and Inhibin A in Singleton and Multiple-Gestation Pregnancies	SGA <10th Percentile (Visit 1 - 12 - 18 wks)	PiGF, dFlt-1, inhibin A	Wilcoxon and Chi- square
70	Boucoiran 2013 [145]	Risks for Preeclampsia and Small for Gestational Age: Predictive Values of Placental Growth Factor, Soluble fms-like Tyrosine Kinase-1, and Inhibin A in Singleton and Multiple-Gestation Pregnancies	SGA <10th Percentile (Visit 2 - 24 - 26 wks)	PiGF, dFlt-1, inhibin A	Wilcoxon and Chi- square
71	Poon 2013 [146]	Combined Screening for Preeclampsia and Small for Gestational Age at 11- 13 Weeks.	SGA	Maternal characteristics, Uterine Artery, PI ,MAP, PiGF MAP, PAPP-A, PiGF	Linear regression
72	Poon 2013 [146]	Combined Screening for Preeclampsia and Small for Gestational Age at 11- 13 Weeks.	SGA (preterm)	Maternal characteristics, Uterine Artery, PI ,MAP, PiGF MAP, PAPP-A, PiGF	Linear regression
73	Poon 2013 [146]	Combined Screening for Preeclampsia and Small for Gestational Age at 11- 13 Weeks.	SGA (term)	Maternal characteristics, Uterine Artery, PI ,MAP, PiGF MAP, PAPP-A, PiGF	Linear regression
74	Singh 2013 [147]	Role of color doppler in the diagnosis of intra uterine growth restriction (IUGR)	IUGR	PI, Resistive Index (RI), umbilical artery and middle cerbeal artery	None
75	Yadav 2013 [112]	Maternal factors in predicting low birth weight babies.	Low birth weight	maternal age, pre preganncy BMI, fathers BMI, parity, ethnicity, per capita monthly income, maternal blood pressure during preganancy	Logistic regression
76	Seed 2011 [148]	Prediction of preeclampsia and delivery of small for gestational age babies based on a combination of clinical risk factors in high-risk women	SGA		Logistic regression
77	Seed 2011 [148]	Prediction of preeclampsia and delivery of small for gestational age babies based on a combination of clinical risk factors in high-risk women	SGA with adverse perinatal outcomes		Logistic regression
78	Poon 2011 [149]	Reference range of birth weight with gestation and first-trimester prediction of small-for-gestation neonates	SGA		Logistic regression
79	Karagiannis 2011 [150]	Prediction of small-for-gestation neonates from biophysical and biochemical markers at 11-13 weeks	SGA		Logistic regression
80	Karagiannis 2011 [150]	Prediction of small-for-gestation neonates from biophysical and biochemical markers at 11-13 weeks	SGA with delivery <37wks		Logistic regression
81	Karagiannis 2011 [150]	Prediction of small-for-gestation neonates from biophysical and biochemical markers at 11-13 weeks	SGA with delivery >37wks		Logistic regression

Model reference	Publication	Study Title	Outcome definition	Type of Predictors	Type of model
82	Poon 2011 [149]	Reference range of birth weight with gestation and first-trimester prediction of small-for-gestation neonates	Birthweight	Gestatational age, weight, height, age, parous, smoking, ethnicity, hypertension, diabetes, conception method	Linear regression
83	Poon 2011 [149]	Reference range of birth weight with gestation and first-trimester prediction of small-for-gestation neonates	Birthweight	Maternal factors, NT,$\beta-$ hCG, PAPP-A	Linear regression
84	De Paco 2008 [151]	Maternal cardiac output between 11 and 13 weeks of gestation in the prediction of preeclampsia and small for gestational age	SGA		Logistic regression
85	Onwudiwe 2008 [152]	Prediction of pre-eclampsia by a combination of maternal history, uterine artery Doppler and mean arterial pressure	SGA		Logistic regression
86	Liu 2008 [153]	Prediction of fetal birthweight in Taiwanese women with pre-eclampsia and gestational hypertension using an equation based on maternal characteristics	Birthweight	preterm delivery, referral status, DBP, GA of hypertensive detection, SBP, parioty	Linear regression
87	Plasencia 2007 [154]	Uterine artery Doppler at 11 + 0 to 13 + 6 weeks in the prediction of pre-eclampsia	SGA		Bayesian Logistic regression
88	Pilalis 2007 [155]	Screening for pre-eclampsia and fetal growth restriction by uterine artery Doppler and PAPP-A at 11-14 weeks' gestation	SGA		Logistic regression
89	Bachman 2003 [115]	Multivariable analysis of tests for the diagnosis of intrauterine growth restriction	IUGR		Logistic regression
90	Doherty 2002 [113]	Estimation of the Doppler ultrasound umbilical maximal waveform envelope, II: prediction of fetal distress	IUGR		None
91	Mamelle 2001 [156]	Definition of fetal growth restriction according to constitutional growth potential	Birthweight	Gestational age, sex, birthrank, pre pregnancy weight, height	Linear regression
92	de Caunes 1990 [121]	Anamnestic pregnancy risk assessment	IUGR		Unclear
93	de Caunes 1990 [121]	Anamnestic pregnancy risk assessment	Low-Birthweight		Unclear
94	Snidvongs 1989 [120]	Intrauterine growth retardation: incidence, screening results, pregnancy outcome	IUGR		Logistic regression
95	Weiner 1985 [157]	Ultrasonic fetal weight prediction: role of head circumference and femur length	Birthweight	Head circumference, Abdominal circumference, fetal length	Linear regression
96	Weiner 1985 [157]	Ultrasonic fetal weight prediction: role of head circumference and femur length	Birthweight	Head circumference, Abdominal circumference, biparietal diameter	Linear regression

Model reference	Publication	Study Title	Outcome definition	Type of Predictors	Type of model
97	Weiner 1985 [157]	Ultrasonic fetal weight prediction: role of head circumference and femur length	Birthweight	Head circumference, Abdominal circumference, fetal length	Linear regression
98	Weiner 1985 [157]	Ultrasonic fetal weight prediction: role of head circumference and femur length	Birthweight	Head circumference, Abdominal circumference, fetal length	Linear regression
99	Weiner 1985 [157]	Ultrasonic fetal weight prediction: role of head circumference and femur length	Birthweight	Head circumference, Abdominal circumference, biparietal diameter	Linear regression

Appendix II - Chapter 3

Appendix IIa - Associated publication

PTJ: Physical Therapy & Rehabilitation Journal | *Physical Therapy*, 2023;103:1–12
https://doi.org/10.1093/ptj/pzad128
Advance access publication date September 26, 2023
Original Research

Development and External Validation of Individualized Prediction Models for Pain Intensity Outcomes in Patients With Neck Pain, Low Back Pain, or Both in Primary Care Settings

Lucinda Archer, MSc, MPH[1,2,3], Kym I.E. Snell, PhD[1,2,3], Siobhán Stynes, PhD[1,4], Iben Axén, PhD[5], Kate M. Dunn, PhD[1], Nadine E. Foster, PhD[1,6], Gwenllian Wynne-Jones, PhD[1], Daniëlle A. van der Windt, PhD[1], Jonathan C. Hill, PhD[1,*]

[1]School of Medicine, Keele University, Keele, Staffordshire, UK
[2]Institute for Applied Health Research, University of Birmingham, Birmingham, UK
[3]National Institute for Health and Care Research (NIHR) Birmingham Biomedical Research Centre, UK
[4]Midlands Partnership Foundation NHS Trust, North Staffordshire Musculoskeletal Interface Service, Haywood Hospital, Staffordshire, UK
[5]Unit of Intervention and Implementation Research for Worker Health, Institute of Environmental Medicine, Karolinska Institutet, Nobels väg 13, Stockholm, Sweden
[6]Surgical Treatment and Rehabilitation Service (STARS) Education and Research Alliance, The University of Queensland and Metro North Hospital and Health Service, Queensland, Australia

*Address all correspondence to Prof. Hill at: j.hill@keele.ac.uk

Abstract

Objective. The purpose of this study was to develop and externally validate multivariable prediction models for future pain intensity outcomes to inform targeted interventions for patients with neck or low back pain in primary care settings.
Methods. Model development data were obtained from a group of 679 adults with neck or low back pain who consulted a participating United Kingdom general practice. Predictors included self-report items regarding pain severity and impact from the STarT MSK Tool. Pain intensity at 2 and 6 months was modeled separately for continuous and dichotomized outcomes using linear and logistic regression, respectively. External validation of all models was conducted in a separate group of 586 patients recruited from a similar population with patients' predictor information collected both at point of consultation and 2 to 4 weeks later using self-report questionnaires. Calibration and discrimination of the models were assessed separately using STarT MSK Tool data from both time points to assess differences in predictive performance.
Results. Pain intensity and patients reporting their condition would last a long time contributed most to predictions of future pain intensity conditional on other variables. On external validation, models were reasonably well calibrated on average when using tool measurements taken 2 to 4 weeks after consultation (calibration slope = 0.848 [95% CI = 0.767 to 0.928] for 2-month pain intensity score), but performance was poor using point-of-consultation tool data (calibration slope for 2-month pain intensity score of 0.650 [95% CI = 0.549 to 0.750]).
Conclusion. Model predictive accuracy was good when predictors were measured 2 to 4 weeks after primary care consultation, but poor when measured at the point of consultation. Future research will explore whether additional, nonmodifiable predictors improve point-of-consultation predictive performance.
Impact. External validation demonstrated that these individualized prediction models were not sufficiently accurate to recommend their use in clinical practice. Further research is required to improve performance through inclusion of additional nonmodifiable risk factors.

Keywords: Back Pain, External Validation, Neck Pain, Prediction Model Development, Targeted Interventions

Received: October 11, 2022. **Revised:** May 5, 2023. **Accepted:** July 14, 2023

Introduction

Despite substantial research focused on improving patient outcomes in those with neck and/or low back pain (NLBP), the impact of these conditions persists and they are among the top 10 reasons for overall disease burden in terms of disability-adjusted-life-years.[1] Although most NLBP episodes are not long-lasting, the proportion of people for whom these symptoms develop into disabling problems is growing.[2] The task of improving first contact treatment for such a large public health problem is an international priority,[3] particularly among low-income and middle-income countries where the disease burden is rising fastest.[4]

This research builds on epidemiological studies, which have consistently highlighted that the transition from acute NLBP episodes into persistent NLBP can be predicted.[5,6] In addition, the use of risk stratification tools to discriminate between risk subgroups to better match initial treatment[7,8] has demonstrated advantages to first contact treatment decision-making[3,9] and is, therefore, recommended by international guidelines.[10–12]

A key next step is not only to stratify individuals into subgroups based on prognostic information, but to also develop and validate individual patient prediction models and to produce communication aids for consultations, including clear visualizations of predictions. Accurately predicting an individual's future pain intensity scores may allow for the development of clinician decision support tools that enable more tailored, individualized clinical care. Existing tools, such as the Keele STarT Back risk stratification tool,[8] determine risk subgroups but do not predict an individual patient's future pain intensity outcomes, which could be used to shape patient and clinician expectations and lead to more personalized health care.

Existing prediction tools for patients with NLBP have been developed using data collected through self-report questionnaires or interviews, often after first presentation in primary care, rather than during the consultation where most of these tools are intended to be used.[7,13,14] It is important to understand to what extent predictions based on such data align with predictions based on information collected by clinicians during a routine consultation.

This research forms part of a larger body of work, developing digital health technology for first contact consultations to support clinical decision-making for patients with NLBP based on individual outcome predictions, as part of a Horizon 2020 European research program (http://backup-project.eu/).[15] This incorporates the Keele STarT MSK Tool (www.keele.ac.uk/startmsk), which predicts poor outcomes in patients in primary care settings who are consulting due to musculoskeletal pain,[16] alongside a set of recommended risk-matched treatment options developed through consensus.[9,17,18]

In the present study, we used predictor items that were agreed to be clinically relevant during the forming of the Keele STarT MSK Tool to develop new models to predict an individual's future pain intensity. We report on the development, internal validation, and external validation of prognostic models to predict 2- and 6- month pain intensity score, which was also modeled when dichotomized as low/moderate-high pain intensity. We explored the predictive performance of these models in external data, using predictor information collected at 2 distinct time points: during the consultation and through patient self-report questionnaires collected 2 to 4 weeks afterwards.

Methods

Source of Data

For this study, secondary analysis of data from 2 existing datasets, the Keele Aches and Pains Study (KAPS)[16] and the STarT MSK Pilot Trial (STarT MSK-pilot),[9] was combined for model development and internal validation, while external validation of the prediction models was conducted in patients from a third existing dataset: the STarT MSK Main Trial (STarT MSK-MT).[18] Eligible patients were defined in the same way in all 3 datasets: those aged 18 and over, consulting at a participating general practice with musculoskeletal pain. The present study included only the subset of patients who consulted with NLBP. Further details of the datasets used in these analyses are included in Supplementary Appendix I.

Outcome Definitions

The outcome was future pain intensity score, which was measured at 2- and 6-month follow-up, through participants' self-reported response to the question "*How intense was your pain, on average, over the last 2 weeks? [Responses on a 0-10 scale, where 0 is 'no pain' and 10 is 'worst pain ever'*]." This score was modeled continuously and separately as a binary outcome, dichotomized as 0 to 4 (low pain intensity) versus 5 to 10 (moderate-high pain intensity). A cut-off of 5 on a 0–10 numerical rating scale to indicate at least moderate pain intensity has been reported previously in the literature[19–21] and was considered clinically meaningful by the physical therapists in the research team (ie, was considered to be the most appropriate cut-point to group patients into those with good pain intensity outcomes [score of 0 to 4] and those with poor pain intensity outcomes [5 or more]).

Predictors

The 10 items from the Keele STarT MSK Tool were considered as predictors in all models (Tab. 1). These were pain intensity (on a scale from 0 to 10), pain self-efficacy, pain impact, walking short distances only, pain elsewhere, thinking their condition will last a long time, other important health problems, emotional well-being, fear of pain-related movement, and pain duration.[16] No statistical selection was conducted, as these predictor variables had all been considered clinically important during the development of the Keele STarT MSK Tool.[16] We included the additional predictor of primary pain site (back or neck pain) as this was identified through discussion with the wider research team as being potentially clinically important for both accurate prediction and face validity.

Predictor information was collected through a postal questionnaire sent to patients within a few days of their general practice consultation (KAPS, STarT MSK-pilot, and STarT MSK-MT). For STarT MSK-MT (the external validation data), predictor information was also collected at the time of general practice consultation in patients in the intervention arm, and these data were used for additional validation analyses.

Further detail on all candidate predictors is given in Supplementary Appendix IV (Tab. S1).

Table 1. Baseline Predictor Responses[a]

	Model Development			External Validation	
	STarT MSK-Pilot *n* = 214	KAPS *n* = 465	Total *n* = 679	POC *n* = 275	2–4 Wk *n* = 586
Age (at consultation, y), mean (SD)	58.2 (15.9)	54.6 (17)	55.7 (16.7)	56.6 (15.5)	58.5 (16.0)
Sex, female	134 (62.6)	267 (57.4)	401 (59.1)	167 (60.7)	353 (60.2)
Primary pain site, neck	59 (27.6)	57 (12.3)	116 (17.1)	61 (22.2)	129 (22.0)
Pain duration					
<3 mo	67 (31.3)	116 (25.0)	183 (27.0)	63 (22.9)	140 (23.9)
3–6 mo	33 (15.4)	67 (14.4)	100 (14.7)	41 (14.9)	91 (15.5)
7–12 mo	29 (13.6)	33 (7.1)	62 (9.1)	33 (12.0)	70 (12.0)
Over 1 y	85 (39.7)	240 (51.6)	325 (47.9)	134 (48.7)	278 (47.4)
Comorbidities (self-reported)					
Diabetes	19 (8.9)	41 (8.8)	60 (8.8)	23 (8.4)	59 (10.0)
Respiratory problem	37 (17.3)	67 (14.4)	104 (15.3)	48 (17.5)	102 (17.4)
Heart problem	65 (30.4)	118 (25.4)	183 (27.0)	75 (27.3)	171 (29.2)
Chronic fatigue	14 (6.5)	11 (2.4)	25 (3.7)	16 (5.8)	43 (7.3)
Anxiety/depression	48 (22.4)	97 (20.9)	145 (21.4)	61 (22.2)	147 (25.1)
Other	46 (21.5)	121 (26.0)	167 (24.6)	71 (25.8)	132 (22.5)
EQ5D—Usual activities					
No problem	44 (20.6)	138 (29.7)	182 (26.8)	40 (14.6)	96 (16.4)
Slight problem	80 (37.4)	139 (29.9)	219 (32.3)	95 (34.6)	206 (35.2)
Moderate problem	52 (24.3)	106 (22.8)	158 (23.3)	86 (31.3)	183 (31.2)
Severe problem	28 (13.1)	60 (12.9)	88 (13.0)	38 (13.8)	70 (12.0)
Unable/extreme problem	10 (4.7)	15 (3.2)	25 (3.7)	12 (4.6)	23 (3.9)
Help to read instructions					
Never	162 (75.7)	354 (76.1)	516 (76.0)	221 (80.4)	476 (82.2)
Rarely	33 (15.4)	48 (10.3)	81 (11.9)	22 (8.0)	51 (8.7)
Sometimes	9 (4.2)	36 (7.7)	45 (6.6)	19 (6.9)	32 (5.5)
Often	2 (0.9)	14 (3.0)	16 (2.4)	9 (3.3)	19 (3.2)
Always	3 (1.4)	13 (2.8)	16 (2.4)	1 (0.4)	2 (0.3)
NDI—Baseline score, mean (SD)	16.1 (8)			17.5 (8.7)	16.6 (8.7)
NDI %—Baseline score, mean (SD)	32.2 (16)			35.1 (17.4)	33.2 (17.4)
RMDQ—Baseline score, median (IQR)	9 (5–13)			10 (5–15)	9 (5–14)
STarT MSK Tool score (baseline), median (IQR)	6 (4–7)	7 (4–9)	7 (4–9)	7 (5–8)	7 (5–9)
STarT MSK Tool score subgroup (baseline)					
High risk	29 (13.6)	118 (25.4)	147 (21.6)	95 (34.6)	140 (23.9)
Medium risk	105 (49.1)	176 (37.9)	281 (41.4)	120 (43.6)	273 (46.6)
Low risk	63 (29.4)	125 (26.9)	188 (27.7)	40 (14.6)	101 (17.4)
1) Pain intensity, median (IQR) *On average, how intense was your pain [where 0 is "no pain" and 10 is "pain as bad as it could be"]?*	7 (5–8)	7 (4–8)	7 (5–8)	7 (6–8)	7 (5–8)
2) Pain self-efficacy *Do you often feel unsure about how to manage your pain condition?*	144 (67.3)	225 (48.4)	369 (54.3)	163 (59.3)	279 (47.6)
3) Pain impact *Over the last 2 wk, have you been bothered a lot by your pain?*	113 (52.8)	330 (71.0)	443 (65.2)	228 (82.9)	462 (78.8)
4) Walking short distances only *Have you only been able to walk short distances because of your pain?*	117 (54.7)	248 (53.3)	365 (53.8)	143 (52.0)	344 (58.7)
5) Pain elsewhere *Have you had troublesome joint or muscle pain in more than 1 part of your body?*	147 (68.7)	304 (65.4)	451 (66.4)	126 (45.8)	398 (67.9)
6) Thinking their condition will last a long time *Do you think your condition will last a long time?*	154 (72.0)	315 (67.7)	469 (69.1)	211 (76.7)	477 (81.4)
7) Other important health problems *Do you have other important health problems?*	81 (37.9)	183 (39.4)	264 (38.9)	88 (32.0)	242 (41.3)
8) Emotional well-being *Has pain made you feel down or depressed in the last 2 wk?*	132 (61.7)	284 (61.1)	416 (61.3)	170 (61.8)	388 (66.2)
9) Fear of pain-related movement *Do you feel it is unsafe for a person with a condition like yours to be physically active?*	56 (26.2)	133 (28.6)	189 (27.8)	144 (52.4)	322 (55.0)
10) Pain duration *Have you had your current pain problem for 6 months or more?*	114 (53.3)	219 (47.1)	333 (49.0)	121 (44.0)	345 (58.9)

[a]Baseline characteristics and responses to the Keele STarT MSK tool items ("As you answer these questions, think about how you have been over the last two weeks:") in model development and external validation data sets. Values are *n* (%) unless otherwise stated. EQ5D = EuroQol 5 Dimension; IQR = interquartile range; KAPS = Keele Aches and Pains Study; NDI = Neck Disability Index; POC = point of consultation; RMDQ = Roland Morris Disability Questionnaire; SD = standard deviation.

Statistical Analysis

Sample Size

The sample size for all analyses was fixed due to the size of the available datasets. We compared the available number of participants for each analysis (Fig. 1) to sample size recommendations for developing[22,23] and externally validating[24–26] clinical prediction models. Further details on the sample size calculations are included in Supplementary Appendix II.

Based on the anticipated inclusion of 11 predefined predictor parameters (1 continuous predictor, modeled linearly, and 10 binary predictors), we required 311 participants for

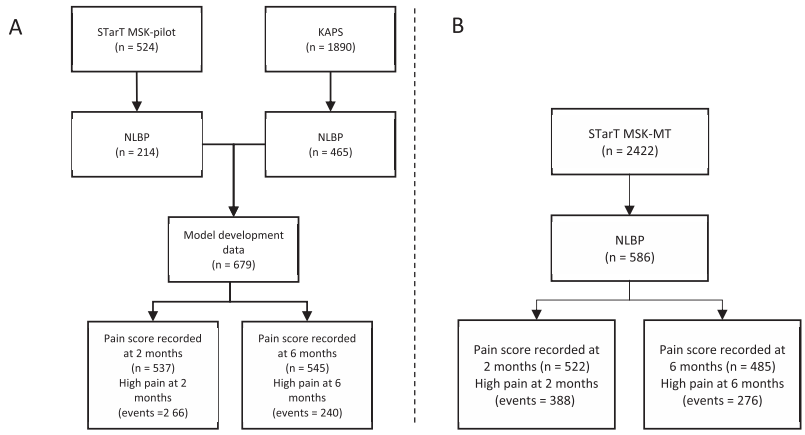


Figure 1. Patient flow summary at model development (A) and at external validation (B). KAPS = Keele Aches and Pains Study; NLBP = neck and/or low back pain.

the development of models for continuous pain intensity score, and at least 824 participants (with 412 "moderate-high pain intensity" events) for binary pain intensity outcomes. Thus, our available data exceeded the requirements for the continuous pain intensity score model but was not enough for the binary pain intensity outcome.

For precise estimation of model performance, we required 892 and 1946 participants to externally validate the continuous and binary outcome models, respectively; thus, estimates of predictive performance are subject to some uncertainty.

Missing Data

Multiple imputation by chained equations was used to account for missing data in both predictor and outcome measurements, under the assumption that data were missing at random.[27,28] Multiple imputation was performed separately for each dataset to allow for the clustering of individuals within that dataset, with the number of imputations chosen to exceed the percentage of incomplete cases.[28] Preliminary checks for associations between missingness and predictor values were conducted to check for obvious violations of the missing at random assumption. Results of analyses were pooled across imputations using Rubin rules where appropriate.[27]

Model Development

Continuous pain intensity score outcomes were modeled using random-effects linear regression, while binary outcomes of moderate-high pain intensity were modeled using random-effects logistic regression.[29] Outcomes were modeled using multilevel mixed-effects models to account for heterogeneity across the 2 model development datasets, resulting in average model intercepts across the KAPS and STarT MSK-pilot datasets.[30] Models were fitted using restricted maximum likelihood (REML), with an unstructured variance–covariance for the random effects on the intercept term.[31] Continuous predictors were modeled linearly on their continuous scale, and all predictors were forced into all models.[32,33]

Internal Validation

Predictive performance of the developed models was assessed through calibration for the continuous outcome models, and through calibration and discrimination for the binary outcome models.[34] Calibration was assessed using the calibration slope, calibration in the- large (CITL), and the ratio of observed to expected cases (O/E, for binary outcome models only). Discrimination was assessed through the C statistic. The proportion of variance in the outcome explained by the predictors in each model was determined using the adjusted R^2 (or pseudo R^2 for binary outcomes, using Nagelkerke and Cox-Snell approaches).

Internal validation was conducted simultaneously for all models, using bootstrapping with 1000 samples to provide optimism-adjusted estimates of predictive performance.[27,29,35]

The optimism-adjusted calibration slope was used as an estimate of the uniform shrinkage factor for each model, with regression coefficients multiplied by this shrinkage factor to correct for overfitting.[29,35,36]

External Validation

Model equations for the 4 prediction models were applied for the participants in the STarT MSK-MT data to calculate the prediction values from each model. Predictive performance measures were calculated, as described for the internal validation, including measures of calibration (calibration slope, CITL, O/E ratio) and discrimination (C statistic), and measures of overall model fit (R^2 or Nagelkerke pseudo R^2 for binary outcomes). Model performance was also assessed within subgroups to check consistency in performance across age ranges, sex, treatment group (matched treatment or usual care), and pain durations prior to presentation.

Predictors in the STarT MSK-MT data were recorded for each patient at 2 time points. Data on predictors were available for each participant when assessed (i) within the general practice consultation, and (ii) after the consultation using a self-reported questionnaire, which was returned by post around 2 to 4 weeks after the consultation.

Table 2. Prognostic Models[a]

	Coefficients for Continuous Outcome, Pain Score (β)		Odds Ratios for Binary Outcome, High Pain	
	2 mo	6 mo	2 mo	6 mo
Pain intensity	0.236	0.269	1.23	1.26
Pain self-efficacy	0.526	0.212	1.38	1.20
Pain impact	0.859	0.632	2.33	1.40
Walking short distances only	0.447	0.934	1.57	2.37
Pain elsewhere	0.49	0.278	1.51	1.33
Thinking their condition will last a long time	1.22	1.673	2.51	3.61
Other important health problems	0.783	0.578	1.73	1.38
Emotional well-being	−0.032	−0.002	0.91	1.28
Fear of pain-related movement	0.041	−0.434	1.07	0.63
Pain duration	1.029	1.129	2.82	2.19
Primary pain site	−0.171	0.515	0.69	1.02
Intercept	−0.304	−1.153	−4.010	−4.324
Var (intercept)	0.237	0.132	0.030	0.041
Shrinkage factor	0.975	0.982	0.930	0.938

[a]Prognostic models after optimism adjustment. Numbers are intercepts (α) and coefficients (β) and for continuous outcome models, intercepts (α), and odds ratios (exp[β]) and for binary outcome models. Uniform shrinkage factors for each model were obtained through bootstrapping with 1000 replications.

The timing of risk predictors collected 2 to 4 weeks after consultation was more similar to the recording of the predictor variables in the model development data, while predictor variable collection at the point of consultation better reflects the models' intended future use. We therefore tested model performance for both data collection time points to assess the validity of our assumption that the developed models could be used in practice at point of consultation.

Extended statistical methods are given in Supplementary Appendix III. All analyses were performed using Stata MP Version 16 (StataCorp). This paper adheres to the TRIPOD checklist for the transparent reporting of multivariable prediction models, see Supplementary Appendix VI.[37]

Role of the Funding Source

The funders played no role in the design, conduct, or reporting of this study.

Results

Study Population

Development Data

Across the 2 model development datasets, 679 patients with NLBP were available for inclusion in the analysis (Fig. 1A). Patients predominantly presented with back pain (83%), with a much smaller group presenting with neck pain (17%), and had a median baseline pain intensity score of 7 (interquartile range = 5–8) out of 10. Patient demographics across the 2 datasets are given in Table 1.

Table 1 also shows a summary of predictor responses. When summarized across both model development datasets, most patients ($n = 451$, 66.4%) had troublesome musculoskeletal pain in more than 1 part of their bodies, with 69.1% ($n = 469$) thinking their condition would last a long time. A further 27.8% ($n = 189$) participants were reported a "fear of pain-related movement."

Few predictors' variables showed substantial differences (larger than 10%) in distribution between the STarT MSK-pilot and KAPS datasets, as can be seen in Table 1. Notable differences in predictor variable distributions included those for pain self-efficacy ("*unsure about how to manage* [their] *pain condition*"; STarT MSK-pilot 67.3%, KAPS 48.4%); and for pain impact ("*bothered a lot by* [their] *pain*" in the preceding 2 weeks; STarT MSK-pilot 52.8%, KAPS 71.0%).

External Validation Data

The STarT MSK-MT data included 586 patients with NLBP for the external validation of the above models (Fig. 1B). Patients predominantly presented with back pain (78%) and had a median baseline pain intensity score of 7 (Interquartile range [IQR] = 6–8) out of 10 reported at point of consultation, and a median of 7 (IQR = 5–8) when reporting in the data collected 2 to 4 weeks after consultation.

The majority of patients reported experiencing moderate-high pain intensity at 2 months, with a prevalence of moderate-high pain intensity at 57.7% (slightly higher than the 49.5% prevalence seen in the development data). This dropped to 47.1% at 6 months follow-up (44.0% for the development data). Baseline pain intensity scores on a scale from 0 to 10 were reported consistently across the data collected 2 to 4 weeks after consultation (median = 7; IQR = 5–8) and at the point of consultation (median = 7; IQR = 6–8), with a slightly narrower spread of scores recorded at point of consultation.

Predictors regarding pain impact and pain self-efficacy were both reported as present in a higher proportion of patients when collected at point of consultation, while the remaining predictor items showed a higher prevalence when recorded 2 to 4 weeks after consultation. The largest difference was seen for the "pain elsewhere" item, with 68% answering "yes" 2 to 4 weeks after consultation compared to only 46% answering "yes" at point of consultation.

Model Development and Internal Validation

The final models for predicting pain intensity scores in patients with NLBP, after optimism adjustment, are presented in Table 2, along with the shrinkage factor estimated via bootstrapping. Detailed results from the internal validation

Demonstration of equations for predicting 6-month pain in individual patients with NLBP

6-month pain score = – 1.153 + 0.269 x (Pain intensity) + 0.212 x (Pain self-efficacy) + 0.632 x (Pain impact) + 0.934 x (Walking short distances only) + 0.278 x (Pain elsewhere) + 1.673 x (Thinking their condition will last a long time) + 0.578 x (Other important health problems) – 0.002 x (Emotional well-being) – 0.434 x (fear of pain-related movement) + 1.129 x (Pain duration) + 0.515 x (Primary pain site)

$$\text{Probability of high pain in 6 months} = \frac{\exp(LP)}{1+\exp(LP)}$$

LP = – 4.324 + 0.231 x (Pain intensity) + 0.182 x (Pain self-efficacy) + 0.336 x (Pain impact) + 0.863 x (Walking short distances only) + 0.285 x (Pain elsewhere) + 1.284 x (Thinking their condition will last a long time) + 0.322 x (Other important health problems) + 0.247 x (Emotional well-being) – 0.462 x (fear of pain-related movement) + 0.784 x (Pain duration) + 0.020 x (Primary pain site)

Where:

- exp is the exponential function
- Pain intensity is scored from 0 to 10, where 0 is "no pain" and 10 is "pain as bad as it could be"
- Primary pain site is scored as 1 for patients with their worst pain in their back, and 0 for neck
- All other variables are scored as 1 if the patient answered "yes" to that question, and 0 otherwise

Example 1

For a patient with back pain with a pain intensity of 7, pain elsewhere, who thinks their condition will last a long time and has pain that has lasted for more than 6 months, pain in 6 months' time would be estimated as:

6-month pain score = – 1.153 + 0.269 x (7) + 0.212 x (0) + 0.632 x (0) + 0.934 x (0) + 0.278 x (1) + 1.673 x (1) + 0.578 x (0) – 0.002 x (0) – 0.434 x (0) + 1.129 x (1) + 0.515 x (1)
= – 1.153 + (0.269 x 7) + 0.278 + 1.673 + 1.129 + 0.515
= 4.3 out of 10

LP = – 4.324 + 0.231 x (7) + 0.182 x (0) + 0.336 x (0) + 0.863 x (0) + 0.285 x (1) + 1.284 x (1) + 0.322 x (0) + 0.247 x (0) – 0.462 x (0) + 0.784 x (1) + 0.020 x (1)
= – 4.324 + (0.231 x 7) + 0.285 + 1.284 + 0.784 + 0.020
= – 0.33

So

$$\text{Probability of high pain in 6 months} = \frac{\exp(-0.33)}{1+\exp(-0.33)} = 42\%$$

Figure 2. Demonstration of prediction calculation. LP = linear predictor; NLBP = neck and/or low back pain.

can be seen in Supplementary Appendix IV (Tab. S2 and Fig. S1).

Conditional on other variables in the model, baseline pain intensity and thinking their condition would last a long time were the strongest predictors of pain intensity at both follow-up time points, with higher baseline pain intensity and expecting the condition to last a long time both being associated with higher pain intensity scores at follow-up. Episode duration (whether the patient had experienced pain for longer than 6 months at the time of their general practice consultation) was also an important predictor, associated with higher pain intensity at both 2 and 6 months. Figure 2 gives a demonstration of how the models for pain intensity could be used to calculate predictions for pain intensity score and the

Table 3. External Validation, Predictive Performance [a]

Time	Outcome	Measure	Point of Consultation	2–4 Wk After Consultation
2 mo	Pain score	Calibration slope	0.650 (0.549 to 0.750)	0.848 (0.767 to 0.928)
		CITL	0.649 (0.506 to 0.792)	0.378 (0.249 to 0.507)
		R^2 median (IQR)	11.1% (10.0% to 12.1%)	25.3% (24.9% to 26.0%)
	High pain	Calibration slope	0.436 (0.338 to 0.535)	0.657 (0.556 to 0.758)
		CITL	1.081 (0.951 to 1.21)	0.798 (0.665 to 0.931)
		O/E	1.599 (1.552 to 1.647)	1.369 (1.329 to 1.41)
		C statistic	0.649 (0.618 to 0.679)	0.726 (0.696 to 0.753)
		Pseudo R^2 median (IQR)	12.1% (10.5% to 13.3%)	27.1% (26.1% to 27.9%)
6 mo	Pain score	Calibration slope	0.593 (0.499 to 0.688)	0.735 (0.656 to 0.815)
		CITL	−0.93 (−1.088 to −0.773)	−1.262 (−1.408 to −1.116)
		R^2 median (IQR)	10.4% (9.4% to 11.4%)	20.8% (20.2% to 21.3%)
	High pain	Calibration slope	0.526 (0.417 to 0.635)	0.71 (0.598 to 0.823)
		CITL	0.028 (−0.101 to 0.157)	−0.307 (−0.438 to −0.176)
		O/E	1.015 (0.984 to 1.046)	0.88 (0.854 to 0.908)
		C statistic	0.663 (0.632 to 0.693)	0.721 (0.692 to 0.749)
		Pseudo R^2 median (IQR)	8.7% (7.6% to 10.2%)	22.1% (21.1% to 23.1%)

[a]Predictive performance of models for pain on external validation in the STarT MSK Main Trial data. CITL = calibration in the large; IQR = interquartile range; O/E = observed/expected ratio.

probability of moderate-high pain intensity for individual patients at 6 months.

External Validation

Details of the model performance on external validation are given in Table 3, while calibration plots for all models can be seen in Figure 3.

Predictions for continuous pain intensity score generated using point-of-consultation data were poorly calibrated when compared to observed values, with calibration slopes of 0.65 (0.55–0.75) and 0.59 (0.50–0.69) at 2 and 6 months respectively. However, the predictions generated in data from 2 to 4 weeks after consultation showed better calibration at both time points, with calibration slopes of 0.85 (0.77–0.93) and 0.74 (0.66–0.82).

Estimates of CITL suggest that the 2-month pain intensity score model systematically underpredicted patients 2-month pain intensity scores by an average of 0.65 pain intensity points in the point-of-consultation predictions (95% CI = 0.51 to 0.79), and by an average of 0.38 pain intensity points (95% CI = 0.25 to 0.51) in predictions generated from predictor responses 2 to 4 weeks after consultation. The pain intensity score model at 6 months systematically overpredicted patients' pain intensity scores for both sources of predictor data, with predicted pain intensity scores being an average of 0.93 points too high in the point-of-consultation data (95% CI = 0.77 to 1.09), and 1.20 points too high in the data from 2 to 4 weeks after consultation (95% CI = 1.12 to 1.41).

The model to predict pain intensity score at 2 months performed better than the 6-month model by all predictive performance measures, in both sources of predictor variables at point of consultation and 2 to 4 weeks after consultation, as can be seen in Table 3.

Calibration performance was poor for the models predicting the binary outcome of moderate-high pain intensity. In the point-of-consultation data, the calibration slope was 0.44 (95% CI = 0.34 to 0.54) for predicting high pain intensity at 2 months and 0.53 (95% CI = 0.42 to 0.64) at 6 months, indicating predictions were too high in those at low risk of high pain intensity and were too low in those at high risk of moderate-high pain. As with the continuous outcome pain intensity models, calibration slopes indicated better calibration performance for predictions generated using predictors collected 2 to 4 weeks after consultation.

Discrimination performance was consistent for models predicting moderate-high pain intensity at both time points. C statistics of 0.65 (0.62–0.68) and 0.66 (0.63–0.69) at 2 and 6 months, respectively, suggest that around 65% and 66% of concordant pairs were correctly identified by the models for these outcome time points, based on predictors recorded at point of consultation. Again, using predictor values from 2 to 4 weeks after consultation to generate predictions gave better discriminative performance, with 73% of concordant pairs correctly identified at 2 months and 72% identified at 6 months.

Analyses to assess model performance across different subgroups (shown in Suppl. Appendix V) suggest that model performance was reasonably consistent across age ranges, sex, treatment group, and pain duration prior to presentation.

Discussion

We have developed and externally validated new individualized prediction models for pain intensity outcomes at 2- and 6-month follow-up in patients consulting with NLBP in general practice, based on the Keele STarT MSK Tool items.

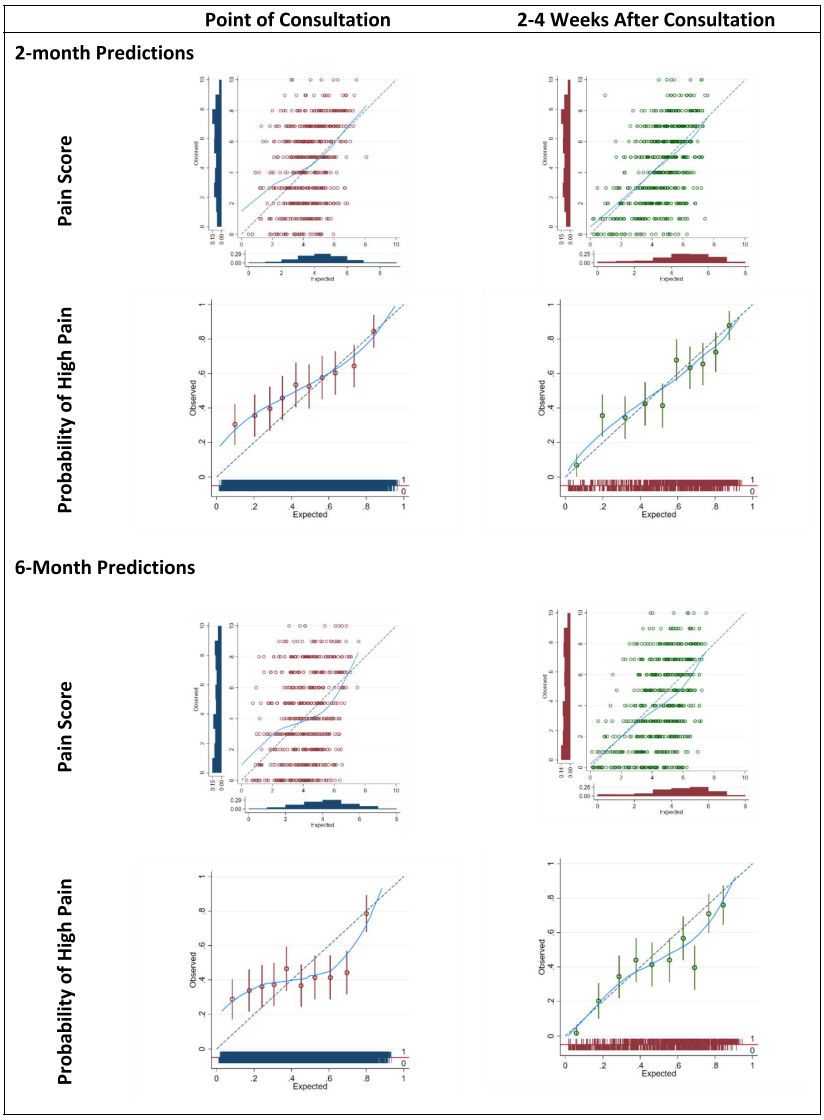


Figure 3. Model calibration in external validation data.

The findings from our external validation demonstrated that models applied in the predictor data collected 2 to 4 weeks after the consultation achieved better predictive performance than data collected at the point of consultation. For example, in terms of calibration performance, the model predicting pain intensity score at 2 months had a calibration slope of 0.848 (0.767–0.928) in data from 2 to 4 weeks after consultation, but only 0.650 (0.549–0.750) at point of consultation. The discriminative ability of the models for predicting 6-month binary outcomes was more consistent between validations in data from 2 to 4 weeks after consultation and point of consultation, with a reasonable C statistic of 0.66 for

moderate–high pain intensity when using data collected at point of consultation compared with 0.72 for predictions based on data from 2 to 4 weeks after consultation. The continuous outcome models showed a systematic underprediction of pain intensity scores at 2 month, and systematic overprediction at 6 months. When predictions were generated using data collected at point of consultation, this overprediction was by around 0.9 points on the 0–10 pain intensity scale, which is not trivial.

The models presented here were based on known predictors for outcomes in patients with NLBP from the team's previous risk stratification work.[16] The choice of candidate predictors was therefore limited to our previously validated Keele STarT MSK Tool items: A decision that proved in hindsight to be a limitation to exploring options for greater accuracy in our predictions of pain intensity outcomes. In understanding reasons for the disappointing performance of these prediction models, it is important to recognize that the Keele STarT MSK Tool was designed for risk-stratification to inform clinicians about risk-matched treatment options. The Keele STarT MSK Tool, therefore, contains mainly items considered to be treatment modifiable risk factors (with pain duration being the only nonmodifiable item). Other known non-modifiable risk factors that were not considered during the model development process presented here include factors such as employment status and other socioeconomic indicators, previous surgery, comorbidities, and previous pain episodes.[38] The next step to improve model performance for individualized predictions will be to explore whether adding such nonmodifiable factors to our models improves their predictive accuracy.

Comparison With Other Studies

Other prediction models currently exist to predict various outcomes in those with NLBP, such as time to recovery for patients with acute low back pain,[13] global perceived effect for patients with persistent neck pain,[39] or disability in patients undergoing surgery for lumbar degeneration.[40] However, as far as we are aware, this is the first time that models have specifically been developed to predict an individual patient's levels of pain intensity at future time points. Prediction models in the field which have previously been updated and externally validated[13,41] often performed suboptimally in external samples. It is not uncommon for risk prediction models to need updating depending on their specific purpose or clinical population, as we have found here.

Several studies have previously shown baseline pain intensity to be a reliable predictor of future pain intensity among patients with low back pain, suggesting that initial pain levels could be a useful indicator of long-term pain outcomes and poor recovery.[42–44] Our findings are also consistent with recent data from Brazil showing that the Keele STarT Back Tool is more predictive of outcomes when collected a few weeks after a physical therapist consultation than at the consultation itself.[28] Our results further agree with a study in United Kingdom and Dutch primary care which suggests that additional assessment of pain intensity after 4 to 6 weeks results in better predictions than when using baseline pain intensity alone.[45] However, although predictor–outcome associations are known to be weaker for time-varying predictors (such as pain intensity) when measured at consultation than when measured 2 to 4 weeks after consultation, it is still recommended for such predictors to be measured at the time of intended model use to improve the applicability of the model in practice.[46]

In patients with neck pain, expectations and previous clinical course of symptoms were found to predict global perceived effect.[41] In patients with low back pain of short duration (<4 weeks) and a pain intensity ≥2/10, the duration of the current episode, pain intensity, and depression was found to predict time to recovery from pain.[14] Despite some differences in populations, we also found that pain intensity, episode duration, and thinking their condition will last a long time to predict future pain intensity (conditional on other model variables). Thus, some predictive factors seem consistent across the literature.

Our findings of the limited predictive performance of low mood, however, contrast with the results of an umbrella review of systematic reviews in this area.[38] Differences in the predictive ability of mood may be related to differences in the populations used for analysis, or due to other variables in our models (such as bothersomeness or pain intensity) assimilating the prognostic impact of mood.

It should be noted that all variables included in our models were self-report measures, with no variables arising from clinical examination, existing electronic medical record data, or imaging results. This decision was made due to a lack of a standardized clinical examination for patients consulting with NLBP, and the expectation that using self-report items could overcome the wide variation in general practitioner clinical examinations. Furthermore, general practitioners rarely have imaging results on which to base a treatment decision. Indeed, previous work suggests that clinical examination and MRI scan results add little to outcome predictions in patients with low back pain over-and-above predictors such as younger age, attitudes and beliefs regarding pain, or depression.[47,48]

Strengths and Limitations

We acknowledge that to achieve a comprehensive understanding of a patient's health status over time, a variety of outcomes are needed. Therefore, a key limitation here is that our new prognostic model focusses solely on predicting pain intensity. Although pain intensity is undoubtedly an essential aspect of measuring a patient's pain experience, it does not reflect the complexity of pain and its wider impact on an individual's physical, emotional, and social well-being. We are, therefore, planning to similarly publish prognostic models for a broader range of outcomes, including physical function (restriction in usual activities) and time off work. Collectively, these models will provide useful information that could inform treatment decisions and guide patient care.

A barrier to implementation in practice is the complexity for clinicians to calculate outcome predictions for individual patients, which may require a prebuilt calculator. Such a calculator has been incorporated into the Back-UP first contact WebApp dashboard,[15] however, reflecting on the relatively low Cstatistic seen within the external validation at the point of consultation; however, further research is needed to improve the discriminative performance of these models before we could recommend use in clinical practice.

Evaluating the performance of predictions at the point of consultation is a clear strength to our validation, with predictive performance assessment at the point where the models are intended to be used in practice. However, small sample sizes for this external validation resulted in some uncertainty around performance estimates in this population, where all

models performed less well than when used in data from 2 to 4 weeks after consultation. Although further external validation in a larger dataset would reduce our uncertainty in predictive performance estimates, the current validation gives a good indication that for these models to perform well at point of consultation, where they would be used in practice, it is likely that updating (for example, to incorporate nonmodifiable risk factors) or recalibration would be needed.

A strength of the external validation is in the representativeness of the sample used. An anonymized medical record audit of all patients with MSK conditions in primary care settings suggested that there was no evidence of selection bias in baseline pain intensity or risk severity in the participants within the STarT MSK MT population. Therefore, we are confident that the sample used was representative of patients consulting to primary care in the United Kingdom.

It was not possible to produce separate prediction models for patients with neck pain and lower back pain, due to the limited number of patients available with the neck as their primary pain site. Rather than neglecting to include these patients with neck pain in our analyses, we instead combined the patient populations and introduced "primary pain site" as a predictor in the model. For this reason, predictor effects are likely to be highly weighted toward the effects experienced by patients with low back pain as their primary pain site, and thus future validation in separate neck pain and low back pain populations would be required to assess the extent to which our conclusions (and models) can be applied to a population with neck pain only.

Implications

Predictions of binary pain intensity outcomes, as included here, provide clinicians and patients with a simple assessment of expected pain intensity at future time points. These offer results that are easy to interpret and can contribute quickly to decision making in clinical settings. However, binary pain intensity outcomes do not provide detailed information about the magnitude or severity of pain, which may limit their usefulness for monitoring changes in pain intensity over time. In contrast, pain intensity predictions on the continuous scale, which this study also presents, give a clear indication of the change in pain intensity score over time, allowing identification of clinically meaningful, smaller changes in expected pain intensity. Analyzing pain intensity outcomes on the continuous scale maximizes the available information for detecting predictor–outcome associations and allows these models to have greater flexibility to be used in different contexts or locations, where a different dichotomy of pain intensity score might be preferred.

The initial individualized prediction models developed within this study were incorporated into an online demonstrator, the Back-UP First Contact web app (http://backup-project.eu/?p=767).[15] For the first time, clinicians were able to see individualized patient predictions on hypothetical patients and give their feedback to the research team about the usability of individual predictions and visualizations to inform treatment decision-making. A future paper will report on the acceptability of the prediction visualizations to both clinicians and patients. The findings of this study, however, suggest that, at present, the prediction models we have are not yet adequate for clinical use for prediction purposes at the point of consultation. Further research is therefore required to improve the prediction models.

Conclusion

We have developed and externally validated models to predict pain intensity outcomes for individual patients consulting in primary care with NLBP. The variables included within the risk prediction models were limited to the existing Keele STarT MSK Tool items. External validation demonstrated that these individualized prediction models, particularly when evaluated at the point of consultation, were not sufficiently accurate to recommend their use in clinical practice. Further research is therefore required to improve the prediction models through inclusion of additional nonmodifiable risk factors.

Author Contributions

Lucinda Archer (Formal analysis, Methodology, Writing—original draft, Writing—review & editing), Kym I.E. Snell (Methodology, Supervision, Writing—original draft, Writing—review & editing), Siobhán Stynes (Conceptualization, Data curation, Funding acquisition, Project administration, Writing—original draft, Writing—review & editing), Iben Axén (Conceptualization, Funding acquisition, Writing—original draft, Writing—review & editing), Kate M. Dunn (Conceptualization, Data curation, Funding acquisition, Writing—original draft, Writing—review & editing), Nadine E. Foster (Conceptualization, Data curation, Funding acquisition, Writing—original draft, Writing—review & editing), Gwenllian Wynne-Jones (Conceptualization, Funding acquisition, Writing—original draft, Writing—review & editing), Danielle A. van der Windt (Conceptualization, Funding acquisition, Writing—original draft, Writing—review & editing), and Jonathan C. Hill (Conceptualization, Data curation, Funding acquisition, Project administration, Writing—original draft, Writing—review & editing)

Acknowledgments

The authors thank the patients who gave their time to participate in the 3 original studies generating the data for this study, as well as the participating general practices.

Funding

This study presents work conducted as part of a project funded by the European Horizon 2020 Research and Innovation Program (Grant Agreement No. 777090). This study uses data collected as part of independent research funded by the National Institute for Health Research under its Programme Grants for Applied Research scheme (STarT MSK program RP-PG-1211-20010) as well as by Centre of Excellence funding from Versus Arthritis (Grant No. 20202). L. Archer and K. Snell were supported by funding from the Evidence Synthesis Working Group, which was funded by the National Institute for Health and Care Research School for Primary Care Research (Project No. 390) and by funding from the National Institute for Health and Care Research Birmingham Biomedical Research Centre at the University Hospitals Birmingham NHS Foundation Trust and the University of Birmingham. K. Snell was funded by the National Institute for Health and Care Research School for Primary Care Research (NIHR SPCR Launching Fellowship). N. Foster was a National Institute for Health and Care Research senior investigator and supported through a National Institute for Health and Care Research Research Professorship (NIHR-RP-011-015).

Data Availability

The datasets analyzed during the current study are available from the corresponding author on reasonable request.

Disclosures and Presentations

The authors completed the ICMJE Form for Disclosure of Potential Conflicts of Interest and reported no conflicts of interest. The views expressed in this study are those of the authors and not necessarily those of the European Union, the National Health Service, the National Institute for Health and Care Research, the funding bodies, or the Department of Health and Social Care. This study reflects only the views of the authors, and the European Commission is not liable for any use that may be made of its contents. The information in this paper is provided "as is," without warranty of any kind, and we accept no liability for loss or damage suffered by any person using this information. Portions of this study were presented at the 41st International Society for Clinical Biostatistics, August 23–27, 2020, Krakow, Poland, and the 18th International Association for the Study of Pain World Congress, August 4–8, 2020, Amsterdam, the Netherlands.

References

1. Vos T, Abajobir A, Abate K, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990–2016: a systematic analysis for the global burden of disease study 2016. *Lancet*. 2017;390:1211–1259. https://doi.org/10.1016/S0140-6736(17)32154-2.
2. Buchbinder R, Underwood M, Hartvigsen J, Maher CG. The lancet series call to action to reduce low value care for low back pain: an update. *Pain*. 2020;161:S57–S64. https://doi.org/10.1097/j.pain.0000000000001869.
3. Foster N, Mullis R, Hill J, et al. Effect of stratified care for low Back pain in family practice (IMPaCT Back): a prospective population-based sequential comparison. *Ann Fam Med*. 2014;12:102–111. https://doi.org/10.1370/afm.1625.
4. Wu A, March L, Zheng X, et al. Global low back pain prevalence and years lived with disability from 1990 to 2017: estimates from the global burden of disease study 2017. *Ann Transl Med*. 2020;8:299. https://doi.org/10.21037/atm.2020.02.175.
5. Stevans JM, Delitto A, Khoja SS, et al. Risk factors associated with transition from acute to chronic low back pain in US patients seeking primary care. *JAMA Netw Open*. 2021;4:e2037371. https://doi.org/10.1001/jamanetworkopen.2020.37371.
6. Traeger AC, Henschke N, Hübscher M, et al. Estimating the risk of chronic pain: development and validation of a prognostic model (PICKUP) for patients with acute low Back pain. *PLoS Med*. 2016;13:e1002019. https://doi.org/10.1371/journal.pmed.1002019.
7. Hill J, Whitehurst D, Lewis M, et al. Comparison of stratified primary care management for low Back pain with current best practice (STarT Back): a randomised controlled trial. *Lancet*. 2011;378:1560–1571. https://doi.org/10.1016/S0140-6736(11)60937-9.
8. Hill J, Dunn K, Lewis M, et al. A primary care back pain screening tool: identifying patient subgroups for initial treatment. *Arthritis Rheum*. 2008;59:632–641. https://doi.org/10.1002/art.23563.
9. Hill J, Garvin S, Chen Y, et al. Stratified primary care versus non-stratified care for musculoskeletal pain: findings from the STarT MSK feasibility and pilot cluster randomized controlled trial. *BMC Fam Pract*. 2020;21:30. https://doi.org/10.1186/s12875-019-1074-9.
10. National Institute for Health and Care Excellence. *Low Back Pain and Sciatica in Over 16s: Assessment and Management (NG59)*. NICE; 2016. Accessed October 13, 2023. https://www.nice.org.uk/guidance/ng59.
11. Jonckheer P, Desomer A, Depreitere B, et al. *Low Back Pain and Radicular Pain: Development of a Clinical Pathway*. Health Services Research (HSR). Brussels: Belgian Health Care Knowledge Centre (KCE); 2017.
12. Bailly F, Trouvin AP, Bercier S, et al. Clinical guidelines and care pathway for management of low back pain with or without radicular pain. *Joint Bone Spine*. 2021;88:105227. https://doi.org/10.1016/j.jbspin.2021.105227.
13. da Silva T, Macaskill P, Kongsted A, Mills K, Maher C, Hancock M. Predicting pain recovery in patients with acute low back pain: updating and validation of a clinical prediction model. *Eur J Pain*. 2019;23:341–353. https://doi.org/10.1002/ejp.1308.
14. da Silva T, Macaskill P, Mills K, et al. Predicting recovery in patients with acute low back pain: a clinical prediction model. *Eur J Pain*. 2017;21:716–726. https://doi.org/10.1002/ejp.976.
15. Back-UP. *Personalised Prognostic Models to Improve Well-being and Return to Work After Neck and Low Back Pain*. 2020. Accessed September 14, 2021. http://backup-project.eu/.
16. Dunn K, Campbell P, Lewis M, et al. Refinement and validation of a tool for stratifying patients with musculoskeletal pain. *Eur J Pain*. 2021;25:2081–2093. https://doi.org/10.1002/ejp.1821.
17. Hill J, Garvin S, Chen Y, et al. Computer-based stratified primary Care for Musculoskeletal Consultations Compared with usual care: study protocol for the STarT MSK cluster randomized controlled trial. *JMIR Res Protoc*. 2020;9:e17939. https://doi.org/10.2196/17939.
18. Hill JC, Garvin S, Bromley K, et al. Risk-based stratified primary care for common musculoskeletal pain presentations (STarT MSK): a cluster-randomised, controlled trial. *Lancet Rheumatol*. 2022;4:E591–E602. https://doi.org/10.2139/ssrn.3925482.
19. Aun C, Lam Y, Collect B. Evaluation of the use of visual analogue scale in Chinese patients. *Pain*. 1986;25:215–221. https://doi.org/10.1016/0304-3959(86)90095-3.
20. *Visual Analogue Scale*. Physiopedia; 2021. Accessed September 14, 2021. https://www.physio-pedia.com/Visual_Analogue_Scale.
21. Von Korff M, Ormel J, Keefe F, Dworkin S. Grading the severity of chronic pain. *Pain*. 1992;50:133–149. https://doi.org/10.1016/0304-3959(92)90154-4.
22. Riley RD, Snell KIE, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: part I—continuous outcomes. *Stat Med*. 2019;38:1262–1275. https://doi.org/10.1002/sim.7993.
23. Riley RD, Snell KI, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: PART II—binary and time-to-event outcomes. *Stat Med*. 2019;38:1276–1296. https://doi.org/10.1002/sim.7992.
24. Archer L, Snell KI, Ensor J, Hudda MT, Collins GS, Riley RD. Minimum sample size for external validation of a clinical prediction model with a continuous outcome. *Stat Med*. 2020;40:133–146. https://doi.org/10.1002/sim.8766.
25. Snell KI, Archer L, Ensor J, et al. External validation of clinical prediction models: simulation-based sample size calculations were more reliable than rules-of-thumb. *J Clin Epidemiol*. 2021;135:79–89. https://doi.org/10.1016/j.jclinepi.2021.02.011.
26. Riley RD, Debray TPA, Collins GS, et al. Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Stat Med*. 2021;40:4230–4251. https://doi.org/10.1002/sim.9025.
27. Vergouwe Y, Royston P, Moons KGM, Altman DG. Development and validation of a prediction model with missing predictor data: a practical approach. *J Clin Epidemiol*. 2010;63:205–214. https://doi.org/10.1016/j.jclinepi.2009.03.017.
28. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med*. 2011;30:377–399. https://doi.org/10.1002/sim.4067.
29. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. New York, NY: Springer; 2009.
30. Debray TPA, Moons KGM, Ahmed I, Koffijberg H, Riley RD. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. *Stat Med*. 2013;32:3158–3180. https://doi.org/10.1002/sim.5732.
31. Debray TP, Damen JA, Riley RD, et al. A framework for meta-analysis of prediction model studies with binary and time-to-event

outcomes. *Stat Methods Med Res*. 2018;28:2768–2786. https://doi.org/10.1177/0962280218785504.
32. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med*. 2006;25:127–141. https://doi.org/10.1002/sim.2331.
33. Royston P, Moons KGM, Altman DG, Vergouwe Y. Prognosis and prognostic research: developing a prognostic model. *BMJ*. 2009;338:b604. https://doi.org/10.1136/bmj.b604.
34. Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. *BMJ*. 2009;338:b605. https://doi.org/10.1136/bmj.b605.
35. Moons KGM, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med*. 2015;162:W1–W73. https://doi.org/10.7326/M14-0698.
36. Van Houwelingen J, Le Cessie S. Predictive value of statistical models. *Stat Med*. 1990;9:1303–1325. https://doi.org/10.1002/sim.4780091109.
37. Collins GS, Reitsma JB, Altman DG, Moons KG. For the members of the Tg. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Eur Urol*. 2014;13:1. https://doi.org/10.1186/s12916-014-0241-z.
38. Burgess R, Mansell G, Bishop A, Lewis M, Hill J. Predictors of functional outcome in musculoskeletal healthcare: an umbrella review. *Eur J Pain*. 2019;24:51–70. https://doi.org/10.1002/ejp.1483.
39. Schellingerhout J, Heymans M, Verhagen A, Lewis M, de Vet H, Koes B. Prognosis of patients with nonspecific neck pain: development and external validation of a prediction rule for persistence of complaints. *Spine*. 2010;35:E827–E835. https://doi.org/10.1097/BRS.0b013e3181d85ad5.
40. McGirt MJ, Sivaganesan A, Asher AL, Devin CJ. Prediction model for outcome after low-back surgery: individualized likelihood of complication, hospital readmission, return to work, and 12-month improvement in functional disability. *Neurosurg Focus*. 2015;39:E13. https://doi.org/10.3171/2015.8.FOCUS15338.
41. Myhrvold BL, Kongsted A, Irgens P, Robinson HS, Thoresen M, Vøllestad NK. Broad external validation and update of a prediction model for persistent neck pain after 12 weeks. *Spine*. 2019;44:E1298–E1310. https://doi.org/10.1097/BRS.0000000000003144.
42. Campbell P, Foster NE, Thomas E, Dunn KM. Prognostic indicators of low back pain in primary care: five-year prospective study. *J Pain*. 2013;14:873–883. https://doi.org/10.1016/j.jpain.2013.03.013.
43. Costa LCM, Maher CG, McAuley JH, et al. Prognosis for patients with chronic low back pain: inception cohort study. *BMJ*. 2009;339:b3829. https://doi.org/10.1136/bmj.b3829.
44. Henschke N, Maher CG, Refshauge KM, et al. Prognosis in patients with recent onset low back pain in Australian primary care: inception cohort study. *BMJ*. 2008;337:a171. https://doi.org/10.1136/bmj.a171.
45. Mansell G, Jordan KP, Peat GM, et al. Brief pain re-assessment provided more accurate prognosis than baseline information for low-back or shoulder pain. *BMC Musculoskelet Disord*. 2017;18:139. https://doi.org/10.1186/s12891-017-1502-8.
46. Whittle R, Royle KL, Jordan KP, Riley RD, Mallen CD, Peat G. Prognosis research ideally should measure time-varying predictors at their intended moment of use. *Diagn Progn Res*. 2017;1:1. https://doi.org/10.1186/s41512-016-0006-6.
47. de Schepper EI, Koes BW, Oei EH, Bierma-Zeinstra SM, Luijsterburg PA. The added prognostic value of MRI findings for recovery in patients with low back pain in primary care: a 1-year follow-up cohort study. *Eur Spine J*. 2016;25:1234–1241. https://doi.org/10.1007/s00586-016-4423-6.
48. Jarvik JG, Hollingworth W, Heagerty PJ, Haynor DR, Boyko EJ, Deyo RA. Three-year incidence of low back pain in an initially asymptomatic cohort: clinical and imaging risk factors. *Spine (Phila Pa 1976)*. 2005;30:1541–1548. https://doi.org/10.1097/01.brs.0000167536.60002.87.

Appendix IIb - Considering the distribution of the error term, e_i, in the generation of predicted probabilities from a linear regression model

In generating the predicted probability from a linear regression model, the distribution of the error term from the linear regression model could be considered. The observed outcome value Y_i for individual i is unlikely to exactly match the predicted value in practice, and is instead equal to Y_{PRED_i} plus some residual error term e_i. This e_i is itself a random variable that, by the underlying assumptions of linear regression, follows a normal distribution with $\mu = 0$ and σ^2, defined by the residual variance of the linear regression model, ${\sigma^2}_{model}$:

$$Y_i = Y_{PREDi} + e_i$$

$$e_i \sim \mathcal{N}(0, {\sigma_{model}}^2)$$

and

$$\hat{\sigma}^2_{model} = \frac{1}{n-p-1}\sum_{i=1}^{n} {e_i}^2$$

Where n is the sample size for model development, p is the number of predictor parameters, and e_i is the residual error for individual i (thus the summation gives the total squared error across all individuals in the model development data).

For some relevant cut-off, C, the underlying distribution of Y_i is unlikely to be known at the point of prediction, thus the value of $P(Y_i < C)$ cannot be gained directly. Instead, the above information from the linear prediction model can be used as follows:

$$p_i = P(Y_i < C)$$

$$= P(Y_{PREDi} + e_i < C)$$

$$= P(e_i < C - Y_{PREDi})$$

Given the distribution of e_i is known to be normal, with distribution parameters described as above, this probability can be gained from the standardised normal distribution tables, with

$$P\left(Z < \frac{X-\mu}{\sigma} = \frac{(C - Y_{PREDi}) - 0}{\sigma_{model}}\right)$$

Or simply,

$$P(Y_i < C) = P\left(Z < \frac{C - Y_{PREDi}}{\sigma_{model}}\right)$$

This formulation exactly matches that from the alternative method described above, in the text of Chapter 3.

Appendix III - Chapter 4, associated publication

Received: 9 March 2020 | Revised: 6 August 2020 | Accepted: 11 September 2020

DOI: 10.1002/sim.8766

RESEARCH ARTICLE

Statistics in Medicine WILEY

Minimum sample size for external validation of a clinical prediction model with a continuous outcome

Lucinda Archer[1] | **Kym I. E. Snell**[1] | **Joie Ensor**[1] | **Mohammed T. Hudda**[2] | **Gary S. Collins**[3] | **Richard D. Riley**[1]

[1]Centre for Prognosis Research, School of Medicine, Keele University, Keele, UK

[2]Population Health Research Institute, St George's, University of London, London, UK

[3]Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK

Correspondence
Lucinda Archer Centre for Prognosis Research, School of Medicine, Keele University, Staffordshire, Keele ST5 5BG, UK.
Email: l.archer@keele.ac.uk

Funding information
British Heart Foundation, Grant/Award Number: FS/17/76/33286; Cancer Research UK, Grant/Award Number: C49297/A27294; European Horizon 2020 research and innovation programme, Grant/Award Number: 777090; Medical Research Council; NIHR Biomedical Research Centre, Oxford; NIHR Clinical Trials Unit Support Funding; NIHR SPCR; NIHR SPCR Evidence Synthesis Working Group, Grant/Award Number: 390; Wellcome, Grant/Award Number: 102215/2/13/2

Clinical prediction models provide individualized outcome predictions to inform patient counseling and clinical decision making. External validation is the process of examining a prediction model's performance in data independent to that used for model development. Current external validation studies often suffer from small sample sizes, and subsequently imprecise estimates of a model's predictive performance. To address this, we propose how to determine the minimum sample size needed for external validation of a clinical prediction model with a continuous outcome. Four criteria are proposed, that target precise estimates of (i) R^2 (the proportion of variance explained), (ii) calibration-in-the-large (agreement between predicted and observed outcome values on average), (iii) calibration slope (agreement between predicted and observed values across the range of predicted values), and (iv) the variance of observed outcome values. Closed-form sample size solutions are derived for each criterion, which require the user to specify anticipated values of the model's performance (in particular R^2) and the outcome variance in the external validation dataset. A sensible starting point is to base values on those for the model development study, as obtained from the publication or study authors. The largest sample size required to meet all four criteria is the recommended minimum sample size needed in the external validation dataset. The calculations can also be applied to estimate expected precision when an existing dataset with a fixed sample size is available, to help gauge if it is adequate. We illustrate the proposed methods on a case-study predicting fat-free mass in children.

KEYWORDS
calibration, continuous outcomes, external validation, prediction model, sample size, R-squared

1 | INTRODUCTION

Clinical prediction models provide individualized outcome predictions to inform patient counseling and clinical decision making, such as treatment and monitoring strategies.[1-3] Depending on the context, they may also be referred to as clinical prediction tools, diagnostic or prognostic models, risk scores, and prognostic indices, among other names. They are typically developed using a regression framework, which provides an equation to predict the outcome conditional on the values of multiple predictors (variables, covariates). In this article, we focus on prediction of continuous outcomes (such as birth weight, depression score, blood pressure or fat mass), for which the model equation is typically a linear regression. Such models can be used to predict an individual's expected outcome value, conditional on the individual's predictor values. The outcome may relate to something current (eg, fat mass level at present) or in the future (eg, pain score at 1 month after a back injury).

Recently we proposed how to calculate the minimum sample size needed to develop a prediction model with a continuous outcome.[4,5] Once a model has been developed, it is important to evaluate its predictive performance in new data, independent to that used to develop the model. This process is known as external validation, and is usually crucial regardless of how a model was developed. In particular, external validation indicates how the model performs in new data that is representative of the target population to which the model will be used in practice.[6-13] However, despite being widely encouraged and having its importance clearly demonstrated,[13-19] external validation of published prediction models is rare in practice, with researchers predominately focusing on the development of new models.[19] Even when external validation is performed, the sample size is often too small to provide reliable conclusions about a model's predictive performance and key measures are often neglected; in particular, calibration of predicted and observed outcome values is rarely examined.[16]

In this article, we propose criteria to determine the minimum sample size needed for external validation of a clinical prediction model with a continuous outcome. We suggest the minimum sample size needs to be large enough to precisely estimate three key measures of predictive performance: calibration slope (agreement between predicted and observed values across the range of predicted values), calibration-in-the-large (CITL, agreement between predicted and observed outcome values on average), and R^2 (the proportion of variance explained). Section 2 introduces these performance measures, while in Section 3, we derive three closed-form solutions for the sample size required to estimate each of them precisely. As these solutions depend on the variance of observed outcome values, we also present a fourth criterion that aims to ensure this variance is estimated precisely. Hence, our sample size calculation comprises checking four criteria, and we suggest the largest sample size calculated from the four approaches is used as the minimum required for the external validation. Section 4 applies our proposal to an applied example, and Section 5 concludes with discussion.

2 | KEY MEASURES OF PREDICTIVE PERFORMANCE FOR A CLINICAL PREDICTION MODEL WITH A CONTINUOUS OUTCOME

Assume that we wish to externally validate an existing prediction model for a continuous outcome, and have obtained a suitable external validation dataset containing a sample of individuals from the target population of interest. We now describe how to quantify the prediction model's performance in this dataset.

First, the researcher needs to calculate the existing model's predicted (expected) outcome value ($Y_{\text{PRED}i}$) for each individual (i). As the outcome is continuous, the existing prediction model equation will usually be in the form of a linear regression and so contain an intercept (α), and predictor effects (β_1, β_2, β_3, etc) corresponding to predictor variables (X_{1i}, X_{2i}, X_{3i}, etc). For example, with three predictors a simple example of an existing prediction model equation is:

$$Y_{\text{PRED}i} = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i}. \quad (1)$$

However, in practice the right hand side of the model equation (also known as the model's *linear predictor*) may be far more complex, for instance with more than three predictors and potential interactions and non-linear terms (eg, defined by splines or polynomials). A real example is given in Box 1.

BOX 1 Hudda et al prediction model for the natural logarithm of ln(fat-free mass) in children[20]

$$\text{Ln(fat-free mass)} = 2.8055 + (0.3073 \times \text{height}^2) - (10.0155 \times \text{weight}^{-1}) + (0.004571 \times \text{weight}) + (0.01408 \times \text{BA}) - (0.06509 \times \text{SA}) - (0.02624 \times \text{AO}) - (0.01745 \times \text{other}) - (0.9180 \times \ln(\text{age})) + (0.6488 \times \text{age}^{0.5}) + (0.04723 \times \text{male})$$

- Predictor variables of black (BA), south Asian (SA), other Asian (AO), or other (other) ethnic origins are all binary, with value of 1 if individual has the particular origin and 0 otherwise
- Height is measured in meters, weight in kilograms, age in years, and fat-free mass in kilograms

Clearly, the external validation dataset must contain values for all the predictors (X_{1i}, X_{2i}, X_{3i}, …) included in the prediction model equation, so that Y_{PREDi} can be calculated by applying the model's equation to each individual. The dataset must also contain the observed outcome value (Y_i) for each individual, so that the prediction model's predictive performance can then be quantified by comparing the Y_{PREDi} values to the Y_i values.

We now introduce three key statistics to quantify a model's predictive performance upon external validation, which focus on overall model fit and calibration.

2.1 | R-squared

R^2 is a well-known measure of overall model fit and quantifies the proportion of outcome variation explained.

Let $\text{var}(Y_i)$ denote the variance of Y_i values in the external validation population, and $\text{var}(Y_i - Y_{\text{PREDi}})$ denote the variance of $(Y_i - Y_{\text{PREDi}})$ values (ie, the prediction errors in the external validation population). Then the true proportion of outcome variation explained by the predicted values from the prediction model, R^2_{val}, is:

$$R^2_{\text{val}} = 1 - \left(\frac{\text{var}(Y_i - Y_{\text{PREDi}})}{\text{var}(Y_i)} \right). \tag{2}$$

Values of R^2_{val} closer to 1 indicate better fit of the Y_{PREDi} from the prediction model.

2.2 | Calibration slope and calibration-in-the-large

Calibration measures the agreement between predicted (Y_{PREDi}) and observed (Y_i) outcome values in the external validation dataset.[21] It is best shown graphically on a calibration plot, with Y_{PREDi} on the horizontal axis plotted against Y_i on the vertical axis, with every individual providing a single data point. A LOESS smoothed calibration curve should also be fitted through the points and presented on the plot.[2,11,22] Ideally, the predicted outcome values are not systematically under- or over-estimated across the entire range of predicted values. That is, the points are scattered randomly around the 45° line of perfect agreement (corresponding to a slope of 1), with little variation around the line (ie, $\widehat{R}^2_{\text{val}}$ is large), and with close agreement between predicted and observed values across the entire horizontal axis range.

To formally quantify calibration performance in an external validation dataset, a calibration model can be fitted of the form,

$$Y_i = \alpha_{\text{cal}} + \lambda_{\text{cal}}(Y_{\text{PREDi}}) + e_{\text{cali}}$$

$$e_{\text{cali}} \sim N(0, \sigma^2_{\text{cal}}), \tag{3}$$

where "cal" is used to emphasize that parameters are from the calibration model. This model can be fitted using standard estimation methods for a linear regression, such as using restricted maximum likelihood estimation. The parameter λ_{cal} represents the *calibration slope*, which measures agreement between predicted and observed outcomes across the whole range of predicted values.[2,3] As mentioned, the ideal λ_{cal} value is 1. A $\lambda_{\text{cal}} < 1$ indicates that some predictions are too

extreme (eg, predictions above the mean are too high, and/or predictions below the mean are too low) and a slope > 1 indicates that the range of predictions is too narrow. A calibration slope < 1 is often observed in external validation studies, as clinical prediction models are often developed in small datasets without adjustment for overfitting, which leads to extreme predictions (miscalibration) in new individuals external to those used for model development.[23-26] The term σ^2_{cal} measures the residual variance in the calibration model.

Note that the true calibration slope in the external validation population can also be expressed as,[27]

$$\lambda_{cal} = \sqrt{\frac{R^2_{cal}\text{var}(Y_i)}{\text{var}(Y_{PREDi})}}, \tag{4}$$

where R^2_{cal} is the proportion of variance of Y_i values explained when the calibration model (3) is fitted to the external validation population.

Systematic over- or under-prediction is still possible even when the calibration slope is 1, and thus it should always be considered alongside calibration plots and CITL. The latter measures the agreement between mean predicted ($\overline{Y}_{PRED}$) and mean observed ($\overline{Y}$) outcome values, which can be estimated in the external validation dataset using:

$$\widehat{\text{CITL}}_{val} = \overline{Y} - \overline{Y}_{PRED}. \tag{5}$$

Estimating CITL_{val} by applying Equation (5) in an external validation dataset is equivalent to estimating α_{cal} by fitting model (3) with the constraint that λ_{cal} equals 1 (see Section 3.2).

3 | SAMPLE SIZE REQUIRED TO TARGET PRECISE ESTIMATES OF PREDICTIVE PERFORMANCE

In this section, we propose four criteria for researchers to use as a basis for determining the minimum sample size required for an external validation study. The first three criteria aim to ensure the sample size is large enough to estimate R^2_{val}, CITL_{val}, and λ_{cal} precisely (ie, with a small margin of error). Closed-form solutions are derived for this purpose. As these expressions depend on the estimates of (residual) variances, a fourth criterion aims to precisely estimate these also.

3.1 | Criterion (i): Precise estimate of R^2_{val}

Our first criterion targets a precise estimate for R^2_{val} from the external validation dataset, such that the confidence interval for R^2_{val} will be narrow. There are many suggestions for deriving confidence intervals for R^2.[28] Here, we focus on the approach suggested by Wishart,[29] which uses the following approximate standard error (SE) of $\hat{R}^2_{val}$:

$$\text{SE}_{\hat{R}^2_{val}} = \sqrt{\frac{4R^2_{val}(1-R^2_{val})^2}{n}}. \tag{6}$$

Tan suggests this approximation works well when the sample size (n) is reasonably large (say >50),[28] which is likely to be the situation when externally validating a clinical prediction model (see criterion (iv)). Rearranging Equation (6) gives a closed-form sample size calculation of:

$$n = \frac{4R^2_{val}(1-R^2_{val})^2}{\text{SE}^2_{\hat{R}^2_{val}}}. \tag{7}$$

Equation (7) can now be used to calculate the sample size (n) required to meet criterion (i), by specifying a desired value for $\text{SE}_{\hat{R}^2_{val}}$ and by setting R^2_{val} at the anticipated true value for the external validation population.

For example, consider an existing prediction model with an adjusted R^2 of 0.5 in the development dataset, with this adjusted (rather than apparent) R^2 giving an unbiased estimate of expected performance in new data. Then, if we assume

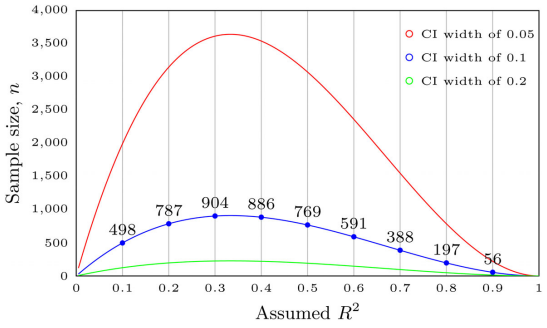


FIGURE 1 Sample size (number of participants, n) needed in an external validation dataset to target a confidence interval for R^2_{val} of a particular width (either 0.05, 0.1, or 0.2) for different assumed R^2_{val} values between 0.1 and 0.9. Sample size calculated using Equation (7) [Colour figure can be viewed at wileyonlinelibrary.com]

the validation sample is from a similar target population to the development sample, a simple starting point is to anticipate R^2_{val} upon external validation is similar to the adjusted $\widehat{R}^2$ reported in the model development study. To target a 95% confidence interval for R^2_{val} that has a narrow width of about 0.1, we need a small $\text{SE}_{\widehat{R}^2_{\text{val}}}$ of 0.0255. This stems from assuming a 95% confidence interval for R^2_{val} can be derived approximately by $\widehat{R}^2_{\text{val}} \pm (1.96 \times \text{SE}_{\widehat{R}^2_{\text{val}}})$. We can now apply Equation (7) to give,

$$n = \frac{4R^2_{\text{val}}(1-R^2_{\text{val}})^2}{\text{SE}^2_{\widehat{R}^2_{\text{val}}}} = \frac{4 \times 0.5 \times (1-0.5)^2}{0.0255^2} = 768.9$$

and so 769 participants are required to meet criterion (i). To achieve the same margin of error, 905 participants are required when assuming R^2_{val} is 0.3, and 197 participants are required when assuming R^2_{val} is 0.8. These values are reasonably close to those using more exact (but not closed-form) approaches to confidence interval derivation, such as that based on the scaled non-central F approximation proposed by Lee.[30] The *ss.aipe.R2* function within Kelley's MBESS package for the R software identifies the sample size required to ensure Lee's confidence interval for R^2_{val} is sufficiently narrow,[31-33] and so is an alternative to using Equation (7).

Figure 1 shows how the required sample size changes from R^2_{val} values between 0.1 and 0.9 based on Equation (7) and assuming $\text{SE}_{\widehat{R}^2_{\text{val}}}$ is 0.0255 to target a confidence interval width of 0.1. The required sample size will be lower when allowing for wider target confidence intervals, and higher when aiming for narrower target confidence intervals (Figure 1). However, we suggest $\text{SE}_{\widehat{R}^2_{\text{val}}} \leq 0.0255$ is a sensible compromise, as it targets a precise estimate (margin of error of 0.05 or less compared to the true value) and still gives a required sample size that will be realistic to obtain in practice.

Note that upon external validation the true R^2_{val} may be lower or higher than the adjusted $\widehat{R}^2$ reported for model development. Therefore, although the adjusted $\widehat{R}^2$ from the development study is a useful starting point, we also recommend calculating the sample size required when assuming slightly different values for the true R^2_{val}. For example, researchers might apply Equation (7) assuming R^2_{val} values $\pm$ 0.1 of the adjusted $\widehat{R}^2$ reported from the development study, and note the largest sample size across this range.

3.2 | Criterion (ii): Precise estimate of CITL

Our second criterion targets a precise estimate of CITL_{val} from the external validation dataset. We estimate CITL_{val} by using $\overline{Y} - \overline{Y}_{\text{PRED}}$ (from Equation (5)), which is equivalent to estimating the intercept when fitting (in the external validation dataset) model (3) with the predicted values as an offset term:

$$Y_i = \text{CITL}_{\text{val}} + 1(Y_{\text{PRED}i}) + e_{\text{CITL}i}$$

$$e_{\text{CITL}i} \sim N(0, \sigma^2_{\text{CITL}}). \quad (8)$$

Therefore the SE of $\widehat{\text{CITL}}$ is:

$$\text{SE}^2_{\widehat{\text{CITL}}} = \text{var}(\overline{Y} - \overline{Y}_{\text{PRED}}) = \text{var}\left(\frac{\sum_{i=1}^{n}(Y_i - Y_{\text{PRED}i})}{n}\right) = \frac{\sigma^2_{\text{CITL}}}{n} = \frac{\text{var}(Y_i)(1 - R^2_{\text{CITL}})}{n}. \quad (9)$$

We can rearrange Equation (9) to obtain an expression for the required sample size:

$$n = \frac{\text{var}(Y_i)(1 - R^2_{\text{CITL}})}{\text{SE}^2_{\widehat{\text{CITL}}}}. \quad (10)$$

Hence, the sample size required to meet criterion (ii) can be derived using Equation (10), for which the researcher must pre-specify R^2_{CITL} (the anticipated proportion of variance explained by the predictions in the external validation population), along with $\text{var}(Y_i)$ (the anticipated variance of Y_i in the target population), and the desired $\text{SE}_{\widehat{\text{CITL}}}$.

A sensible starting point is to assume CITL is zero, as then $R^2_{\text{CITL}} = R^2_{\text{val}}$ (the anticipated proportion of variance explained by the predictions upon validation), and so

$$n = \frac{\text{var}(Y_i)(1 - R^2_{\text{val}})}{\text{SE}^2_{\widehat{\text{CITL}}}}, \quad (11)$$

with R^2_{val} assumed to be the same as the adjusted $\widehat{R}^2$ reported from the development study.

If CITL is not zero then R^2_{CITL} will not equal R^2_{val}. Hence, it is also sensible to consider a range of values for R^2_{CITL} when applying Equation (10), such as $\pm$ 0.1 of the adjusted $\widehat{R}^2$ reported from the development study, and to note the largest sample size across this range.

The value that defines a precise $\text{SE}_{\widehat{\text{CITL}}}$ is context specific, as it depends on the scale of the outcome values. For example, for systolic blood pressure an SE of about 2.5 mmHg may suffice, but for BMI a smaller SE may be required as the scale is much narrower.

For instance, consider external validation of a prediction model for systolic blood pressure with a reported adjusted R^2 of 0.5 in the development study, and that the variance of the observed Y_i values is anticipated to be 400 in the target population for the validation study. Let us target an $\text{SE}_{\widehat{\text{CITL}}}$ of 2.55, as this gives a 95% confidence interval for CITL_{val} with a narrow width of about 10 mmHg, assuming a 95% confidence interval for CITL_{val} can be derived approximately by $\widehat{\text{CITL}} \pm (1.96 \times \text{SE}_{\widehat{R}^2_{\text{val}}})$. Assuming $R^2_{\text{CITL}} = R^2_{\text{val}} = 0.5$, then applying Equation (10) gives,

$$n = \frac{\text{var}(Y_i)(1 - R^2_{\text{val}})}{\text{SE}^2_{\widehat{\text{CITL}}}} = \frac{400 \times (1 - 0.5)}{2.55^2} = 30.76$$

and thus at least 31 participants are required to achieve criterion (ii).

More cautiously assuming that $R^2_{\text{CITL}} = 0.4$, the required sample size is

$$n = \frac{\text{var}(Y_i)(1 - R^2_{\text{CITL}})}{\text{SE}^2_{\widehat{\text{CITL}}}} = \frac{400 \times (1 - 0.4)}{2.55^2} = 36.91$$

and thus 37 participants are required.

It is likely that the sample size to precisely estimate CITL is smaller than that required to precisely estimate the measures outlined in criteria (i), (iii), and (iv).

3.3 | Criterion (iii): Precise estimate of calibration slope

The third criterion targets a precise estimate of λ_{cal}, which represents the calibration slope obtained from fitting calibration model (3) in the external validation dataset. As $\widehat{\lambda}_{\text{cal}}$ is the slope from a simple linear regression model, the SE of $\widehat{\lambda}_{\text{cal}}$ can be estimated by,[34]

$$SE^2_{\hat{\lambda}_{cal}} = \frac{\sigma^2_{cal}}{\sum_{i=1}^{n}(Y_{PREDi} - \overline{Y}_{PRED})^2},$$

where σ^2_{cal} is the residual variance from model (3).

By utilizing Equation (4), and also recognizing that $\sigma^2_{cal} = var(Y_i)(1 - R^2_{cal})$ and that $\sum_{i=1}^{n}(Y_{PREDi} - \overline{Y}_{PRED})^2 = (n-1)\,var(Y_{PREDi})$, we can write $SE^2_{\hat{\lambda}_{cal}}$ in terms of λ^2_{cal} and R^2_{cal} values,[27] as follows:

$$\begin{aligned}
SE^2_{\hat{\lambda}_{cal}} &= \frac{\sigma^2_{cal}}{\sum_{i=1}^{n}(Y_{PREDi} - \overline{Y}_{PRED})^2} \\
&= \frac{var(Y_i)(1 - R^2_{cal})}{(n-1)\,var(Y_{PREDi})} \\
&= \frac{var(Y_i)}{(n-1)\,var(Y_{PREDi})} - \frac{var(Y_i)R^2_{cal}}{(n-1)\,var(Y_{PREDi})} \\
&= \frac{var(Y_i)}{(n-1)\,var(Y_{PREDi})} - \frac{\lambda^2_{cal}}{(n-1)} \\
&= \frac{\lambda^2_{cal}}{(n-1)\,R^2_{cal}} - \frac{\lambda^2_{cal}R^2_{cal}}{(n-1)R^2_{cal}} \\
&= \frac{\lambda^2_{cal}(1 - R^2_{cal})}{(n-1)R^2_{cal}}.
\end{aligned} \tag{12}$$

Rearranging gives:

$$n = \frac{\lambda^2_{cal}(1 - R^2_{cal})}{SE^2_{\hat{\lambda}_{cal}} R^2_{cal}} + 1. \tag{13}$$

Equation (13) allows calculation of the required sample size for a desired $SE_{\hat{\lambda}_{cal}}$, conditional on specifying λ_{cal} (the anticipated (mis)calibration across the range of predicted values) and R^2_{cal} (the anticipated proportion of variance in observed Y_i values explained by the calibration model).

In terms of choosing $SE^2_{\hat{\lambda}_{cal}}$, a value ≤0.051 is recommended, to target a 95% confidence interval for λ_{cal} that has a narrow width ≤ 0.2 (eg, if the calibration slope was 1, the confidence interval would be 0.9 to 1.1 assuming confidence intervals derived by $\hat{\lambda}_{cal} \pm 1.96SE_{\hat{\lambda}_{cal}}$; note that replacing 1.96 by critical values of the t-distribution is unnecessary, as the sample size will not be small).

In terms of choosing λ_{cal}, a simple starting point is to assume good calibration, such that $\lambda_{cal} = 1$ and $\alpha_{cal} = 0$ in model (3). Then, $R^2_{cal} = R^2_{val}$ from criterion (i), and so R^2_{cal} might be assumed to be the same as the adjusted R^2 estimated in the model development study. For example, for external validation of a prediction model that had an estimated adjusted R^2 of 0.5 in the development dataset, a simple starting point is to anticipate the same value for R^2_{val}. Then, assuming the model's predictions will be well calibrated in the external validation dataset (ie, on average, fitting model (3) would give $\hat{\alpha}_{cal}$ of 0 and a $\hat{\lambda}_{cal}$ of 1), using Equation (13) gives,

$$n = \frac{\lambda^2_{cal}(1 - R^2_{cal})}{SE^2_{\hat{\lambda}_{cal}} R^2_{cal}} + 1 = \frac{1 \times (1 - 0.5)}{0.051 \times 0.051 \times 0.5} + 1 = 385.47$$

and thus 386 participants are required to target a confidence interval width of 0.1 for the calibration slope, under the assumptions of good calibration.

The sample size should also be large enough to precisely estimate some miscalibration. Often when a prediction model is externally validated the calibration slope is less than 1, due to overfitting during model development that was unaccounted for in the final prediction model equation (ie, penalization or shrinkage estimation methods were not used). In such situations R^2_{cal} can still be assumed to be the same as the adjusted R^2 presented for model development, as this

value specifically adjusts for optimism due to overfitting. When applying Equation (13) for fixed R^2_{cal} and $SE^2_{\hat{\lambda}_{cal}}$ values, lowering the assumed λ_{cal} below 1 will produce lower sample sizes than when assuming the prediction model is well calibrated. Hence, assuming λ_{cal} is 1 is more conservative for the sample size calculation.

Further sensitivity analyses could be undertaken if desired. For example, we could change both λ_{cal} and R^2_{cal} values. However this is complex, as Equation (4) reveals that the value of λ_{cal} depends on R^2_{cal} (and also var(Y_i) and var(Y_{PRED_i})). Therefore, changing the assumed value of λ_{cal} has implications for what the assumed value of R^2_{cal} should be. This may be too intricate for the sample size calculation. Similarly, although situations of under-prediction (where λ_{cal} is >1) may lead to larger required sample sizes, this may not be practical to consider as over-prediction situations are more common. Thus, we generally suggest to apply Equation (13) assuming good calibration (λ_{cal} =1) and set R^2_{cal} equal to the adjusted R^2 estimated for model development.

3.4 | Criterion (iv): Precise estimates of residual variances

Our final criterion targets precise estimates of $\hat{\sigma}^2_{CITL}$ and $\hat{\sigma}^2_{cal}$. This is essential because, although these residual variances are not direct measures of predictive performance themselves, their estimated values are used toward parameter estimates and, crucially, $SE_{\widehat{CITL}_{val}}$ and $SE_{\hat{\lambda}_{cal}}$.

For $\hat{\sigma}^2_{CITL}$, we can equivalently consider the sample size needed to precisely estimate a residual variance in a linear regression model with only an intercept (see model (8)). In such situations, Harrell suggests calculating the sample size to ensure the lower and upper bounds of a 95% confidence interval for the residual variance has a small multiplicative margin of error (MMOE) around the true value,[2] using

$$\text{MMOE} = \sqrt{\max\left(\frac{\chi^2_{1-\frac{\alpha}{2},n-1}}{n-1}, \frac{n-1}{\chi^2_{\frac{\alpha}{2},n-1}}\right)}, \tag{14}$$

where $\chi^2_{1-\frac{\alpha}{2},n-1}$ and $\chi^2_{\frac{\alpha}{2},n-1}$ are the critical values of a χ^2 distribution with $n-1$ degrees of freedom for which there is, respectively, a probability of $1-\frac{\alpha}{2}$ and $\frac{\alpha}{2}$ of being less than the critical value. The second term within the bracket of Equation (14) will typically give the largest MMOE.

We recommend a margin of error of within 10% of the true value ($1.0 \leq \text{MMOE} \leq 1.1$), for which Equation (14) reveals that a sample size of at least 234 participants is needed to ensure an MMOE ≤1.1 for $\hat{\sigma}^2_{CITL}$.

For precise estimation of $\hat{\sigma}^2_{cal}$, we need to adjust the sample size required for a slope parameter being estimated (see model (3)). As outlined by Riley et al,[4] the solution is simply 234 + 1, and thus 235 participants are required to ensure an MMOE of ≤1.1 for $\hat{\sigma}^2_{cal}$. Hence, in summary, at least 235 participants are needed to meet criterion (iv), and thus 235 is the minimum sample size required for any external validation of a prediction model for a continuous outcome, regardless of context and before consideration of criteria (i), (ii), or (iii).

3.5 | Summary of the criteria

Our sample size criteria aim to ensure the external validation dataset will precisely estimate R^2_{val}, CITL, calibration slope, and residual variances. The approach requires a separate sample size calculation for each criterion, and the largest sample size calculated provides the minimum needed for the external validation study. A step-by-step guide to our proposal is provided in Figure 2.

4 | APPLIED EXAMPLE

We now illustrate our sample size proposal using an applied example. Hudda et al developed a prediction model for the natural logarithm of fat-free mass in children and adolescents aged 4 to 15 years, including 10 predictor parameters based on height, weight, age, sex, and ethnicity (see Box 1 for model equation).[20] The model is required to provide an estimate of an individual's current fat mass (weight - predicted fat-free mass). The apparent calibration of the model in the development dataset is shown in Figure 3A. In the development dataset, the estimated adjusted R^2 was 0.95. An initial

STEP 1: Calculate the sample size needed to precisely estimate R^2_{val} (criterion (i))

Apply equation 7,

$$n = \frac{4R^2_{val}(1-R^2_{val})^2}{SE^2_{\hat{R}^2_{val}}}$$

after specifying suitable values for $SE_{\hat{R}^2_{val}}$ and R^2_{val}. We recommend using $SE_{\hat{R}^2_{val}} \leq 0.0255$ (to target a 0.1 width), and initially choosing R^2_{val} to equal the adjusted $\hat{R}^2$ reported for the model development study. A few other values for R^2_{val} might also be considered (e.g. values ± 0.1 the adjusted $\hat{R}^2$ reported from the development study).

STEP 2: Calculate the sample size needed to precisely estimate calibration-in-the-large (criterion (ii))

Apply equation 10,

$$n = \frac{var(Y_i)(1-R^2_{CITL})}{SE^2_{C\hat{I}TL}}$$

after specifying suitable values for R^2_{CITL} (akin to those used for R^2_{val} in step 1), $SE_{C\hat{I}TL}$ and $var(Y_i)$. The latter represents the variance of outcome values in the population of interest, and should be based on other existing knowledge (e.g. previous studies). The value of $SE_{C\hat{I}TL}$ should aim to ensure that $(\bar{Y} - \bar{Y}_{PRED}) \pm (1.96 * SE_{C\hat{I}TL})$ is narrow, and so needs to be chosen in context of what constitutes a precise estimate of the anticipated mean prediction error in the clinical setting of interest.

STEP 3: Calculate the sample size needed to precisely estimate calibration slope (criterion (iii))

Apply equation 13,

$$n = \frac{\lambda^2_{cal}(1-R^2_{cal})}{SE^2_{\hat{\lambda}_{cal}} R^2_{cal}} + 1$$

after specifying suitable values for λ_{cal}, R^2_{cal} and $SE_{\hat{\lambda}_{cal}}$. We recommend $SE_{\hat{\lambda}_{cal}} \leq 0.051$ (to target a confidence interval width ≤ 0.2), choosing R^2_{cal} to be that chosen for R^2_{val} (i.e. the adjusted R^2 reported from the model development study; see step 1), and $\lambda_{cal} = 1$ (good calibration).

STEP 4: Calculate the sample size for precisely estimating residual variances (criterion (iv))

To target residual variance estimates in the calibration models that have a margin of error of $\leq 10\%$, at least 235 participants are required based on equation 14.

STEP 5: Calculate the final sample size

The required minimum sample size is the maximum value from steps 1 to 4, to ensure that each of criteria (i) to (iv) are met.

FIGURE 2 Summary of the steps involved in our sample size calculation for external validation of a clinical prediction model for a continuous outcome

external validation was undertaken in 176 children aged 11-12 years from the UK Avon Longitudinal Study of Parents and Children (ALSPAC) study,[35,36] where the model had an estimated R^2_{val} of 0.90 Figure 3B. However, as acknowledged by Hudda et al, further external validation is warranted in a broader age range, for which a sample size calculation can be undertaken using our proposal. We assume that the validation population is similar to the development population, and work through the calculations for criteria (i) to (iv).

STEP 1: Calculate the sample size needed to precisely estimate R^2_{val} (criterion (i))

This requires us to apply Equation (7). Based on assuming an $R^2_{\text{val}} = 0.90$, as in the published external validation of the model, and a $\text{SE}_{\hat{R}^2_{\text{val}}}$ of 0.0255 to target a confidence interval width of 0.1, a sample size of 56 children is required, as:

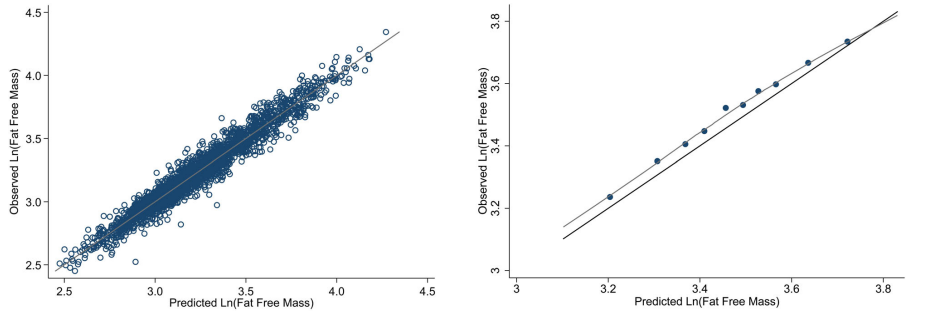


FIGURE 3 Calibration performance: A, in the development dataset; and, B, on external validation of the prediction model for ln(fat-free mass) in children, as proposed by Hudda et al.[20] The 45° line shows perfect calibration on both plots. * in B, individual level data points cannot be shown for confidentiality reasons. Data points shown are mean predicted against mean observed ln(fat-free mass) within tenths of predicted ln(fat-free mass), with a local regression smoother through the individual level data points shown in gray [Colour figure can be viewed at wileyonlinelibrary.com]

$$n = \frac{4R^2_{\text{val}}(1-R^2_{\text{val}})^2}{\text{SE}^2_{\hat{R}^2_{\text{val}}}} = \frac{4 \times 0.90 \times (1-0.90)^2}{0.0255^2} = 55.4.$$

It is sensible to also consider that the model may perform worse upon external validation, say with a 0.1 reduction in R^2_{val} to 0.80. Then, the required sample size to meet criterion (i) is 197 children. These sample size values are also identified within Figure 1.

STEP 2: Calculate the sample size needed to precisely estimate calibration-in-the-large (criterion (ii))

This requires us to apply Equation (10), which itself requires us to specify $\text{var}(Y_i)$, the anticipated variance of outcome values in the target population for external validation. Let us illustrate how to derive this from published information. In their paper, Hudda et al reported the lower quartile (LQ) as 20.8 and the upper quartile (UQ) as 30.6 kg of fat-free mass in their development dataset. By transforming this to the ln(kg) scale, and assuming ln(fat-free mass) values are approximately normally distributed, we can derive an estimate of the SD of the ln(fat-free mass) in the development population using[37]:

$$\frac{\ln UQ - \ln LQ}{1.35} = \frac{\ln 30.6 - \ln 20.8}{1.35} = 0.286.$$

Therefore, based on the published information $\widehat{\text{var}}(Y_i) \approx 0.286^2 = 0.082$. Interestingly, when contacting the original study authors directly for this information, they calculated it to be a similar value of $\widehat{\text{var}}(Y_i) = 0.089$. We will use this value from the study authors going forward.

We must also specify the expected value for R^2_{CITL}. We begin by assuming $R^2_{\text{CITL}} = R^2_{\text{val}}$ and that this is 0.90, as in Hudda's initial external validation of the model.

The precision required to estimate CITL needs to be placed in context of the mean outcome value in the population. Hudda et al reported a median baseline fat-free mass of 24.8 kg. If we assume that the mean value is similar, then we have:

$$\overline{Y} \approx \ln 24.8 = 3.21.$$

Considering the original untransformed scale, an accuracy of approximately ±1 kg around $\overline{Y}$ seems reasonably precise. A confidence interval of about 23.8 to 25.8 on the kg scale would correspond to a 95% CI of about 3.17 to 3.25 around $\overline{Y}$, implying a target $\text{SE}^2_{\widehat{\text{CITL}}}$ of about 0.02.

TABLE 1 Summary of the sample size calculation for external validation of the prediction model of Hudda et al

Criterion	Target precision	Assumptions	Minimum sample size required
(i) Precise estimate of R^2_{val}	$SE_{\widehat{R^2_{val}}} = 0.0255$	$R^2_{val} = 0.8$	197
		$R^2_{val} = 0.9$	56
(ii) Precise estimate of **CITL**	$SE_{\widehat{CITL}} = 0.02$	$R^2_{CITL} = R^2_{val} = 0.8$, $var(Y_i) = 0.089$	45
		$R^2_{CITL} = R^2_{val} = 0.9$, $var(Y_i) = 0.089$	23
(iii) Precise estimate of λ_{cal}	$SE_{\widehat{\lambda}_{cal}} = 0.051$	$R^2_{cal} = R^2_{val} = 0.9$ $\widehat{\lambda}_{cal} = 1$	44
(iv) Precise $\widehat{\sigma}^2_{CITL}$ and $\widehat{\sigma}^2_{cal}$	$1.0 \leq MMOE \leq 1.1$	-	235

Therefore, we can now apply Equation (10) to obtain a sample size of,

$$n = \frac{var(Y_i)\,(1 - R^2_{CITL})}{SE^2_{\widehat{CITL}}} = \frac{0.089 \times (1 - 0.9)}{0.02^2} = 22.3,$$

and thus 23 participants are required to meet criterion (ii). To be conservative, let us assume a 0.1 lower value for R^2_{CITL} to 0.80. Then, the required sample size to meet criterion (ii) would increase to 45 children.

STEP 3: Calculate the sample size needed to precisely estimate calibration slope (criterion (iii))

This requires us to apply Equation (13) after choosing values for $SE_{\widehat{\lambda}_{cal}}$, R^2_{cal}, and λ^2_{cal}. Let us choose an $SE_{\widehat{\lambda}_{cal}}$ of 0.051 to target a confidence interval width of 0.2. Further, we assume $R^2_{cal} = R^2_{val}$ and take the value of 0.90 as reported by the initial validation study of Hudda et al; and assume good calibration such that λ^2_{cal} is 1. We can now apply Equation (13) to give,

$$n = \frac{\lambda^2_{cal}(1 - R^2_{cal})}{SE^2_{\widehat{\lambda}_{cal}}\, R^2_{cal}} + 1 = \frac{1 \times (1 - 0.9)}{0.051^2 \times 0.9} + 1 = 43.72,$$

and thus 44 participants are required.

STEP 4: Calculate the sample size for precisely estimating residual variances (criterion (iv))

To ensure a 10% margin of error in residual variance estimates from the calibration models, at least 235 participants are required (see Section 3.4).

STEP 5: Calculate the final sample size

Assuming we aim to validate the model of Hudda et al in a population similar to the development data, steps 1 to 4 have provided four sample sizes to ensure criteria (i) to (iv) are met. These are summarized in Table 1. Based on the largest of these sample sizes, the final minimum sample size required to meet all criteria is 235 participants. This is driven by criterion (iv), to target sufficient precision around $\widehat{\sigma}^2_{CITL}$ and $\widehat{\sigma}^2_{cal}$.

5 | WHAT IF SAMPLE SIZE FOR EXTERNAL VALIDATION IS FIXED?

Sometimes there are no resources for prospective recruitment of participants to a new study for external validation of a prediction model. Then, researchers might seek an existing (already collected) dataset from the target population of interest. However, the sample size of an existing dataset is fixed, and so the researcher (and other stakeholders such as funders and collaborators) needs to know if it is large enough for reliable external validation. In this situation, our calculations in steps 1 to 4 can be re-expressed to calculate the expected $SE_{\widehat{R^2_{val}}}$, $SE_{\widehat{CITL}}$, $SE_{\widehat{\lambda}_{cal}}$, and MMOE conditional on the known sample size and assumed values of R^2_{val}, $var(Y_i)$, R^2_{CITL}, R^2_{cal}, and λ_{cal} as before.

For example, in the initial external validation of Hudda et al, an existing dataset, from the ALSPAC study, of 176 children was used. Based on the calculation shown in Table 1, this sample size is likely to give very precise estimates of R^2_{val}, CITL, and λ_{cal} when assuming $R^2_{val} = R^2_{CITL} = R^2_{cal}$ is 0.9. However, the sample size is lower than the 235 recommended for precise estimation of $\widehat{\sigma}^2_{CITL}$ and $\widehat{\sigma}^2_{cal}$, and so the MMOE for these estimates is expected to be >10%. Nevertheless, when applying Equation (14) assuming 176 participants, the MMOE is 1.12, and thus the error is expected to be 12%, only just over the 10% recommendation. Hence, this existing dataset appears to have a reasonable sample size for external validation, which would have been useful for Hudda et al to know at the time.

6 | DISCUSSION

We have proposed closed-form sample size calculations for studies externally validating a prediction model for a continuous outcome. These aim to ensure the sample size is large enough to precisely estimate key measures of predictive performance (R^2, CITL, and calibration slope) and the residual variances in calibration models. This led to four criteria, and the largest sample size required satisfying all four criteria is the recommended minimum sample size needed in the external validation dataset. Our work builds on minimum sample size calculations for model development.[4,38]

As with any sample size calculation, assumptions are required to implement our proposed approach. In particular, researchers must specify the model's anticipated R^2_{val}, $\widehat{var}(Y_i)$, and $\hat{\lambda}_{cal}$ in the validation dataset. As discussed, a simple starting point is to assume these will be the same as those reported for the original model development study, especially if the target population (for validation) is similar to that in the model development study. Then the researcher might consider sample sizes based on slight adjustments; in particular, assuming the model may perform slightly worse than in the development dataset. Our example illustrated this for a prediction model of fat-free mass in children, where we assumed an R^2_{val} of 0.8 rather than the 0.90 or 0.95 values reported in the original model development study. Lower values may be even more important to consider in situations where the development dataset was small (such that reported performance statistics were estimated with large uncertainty); the developed prediction model did not adjust for overfitting using, for example, penalization and shrinkage techniques (such that reported performance statistics are likely to be optimistic); and in situations where the intention is to validate the model in a different population or setting from that used in the development study. Larger sample sizes may be needed if missing data are expected, and if a model's predictive performance in key subgroups (eg, males, females) is of interest.

Section 5 discussed how to use our calculations when an existing dataset (of a fixed sample size) is already available, in order to gauge the expected precision of estimates conditional on the sample size available. Ideally the dataset will be large enough to ensure precise estimates, as then more robust conclusions about predictive performance will be possible. However, we recognize that even when datasets are small, obtaining estimates of predictive performance is still useful; in particular, these could ultimately be combined in a meta-analysis.[39] It is important that datasets for external validation are high quality and applicable to the target population, setting, and timing of implementing the prediction model in practice. Adequate sample size does not overcome issues in quality and applicability.[39-41]

We chose to focus on R^2, CITL, and calibration slope as these are key performance measures; ensuring precise estimation of residual variances is also important, as they are used to calculate the aforementioned predictive performance measures and also mean-squared error. We anticipate that the largest sample size will usually be driven by criterion (i), (iii), or (iv). Further work might consider precise estimation of calibration curves,[11,22,42] and extension to non-continuous outcomes is needed, building on work of others.[11,17,43] Closed-form sample size solutions are transparent and quick to implement, but more difficult to derive for binary and time-to-event outcomes. Jinks et al do suggest closed-form sample size calculations for precisely estimating the D statistic for time-to-event prediction models.[44] Also, we only focused on statistical measures of predictive performance, and not on clinical utility or impact of using the model to inform healthcare decisions (eg, initiation of treatment).

Finally, sometimes the sample size for an external validation dataset must also be large enough for model updating, for example, when the researcher aims to recalibrate one or a few of the model parameters to the target population of interest. Then, the required sample size needs to meet the criteria described in this article (for external validation), and also those criteria proposed for model development (as model updating is akin to model development[5]). The exact sample size needed for model updating depends on how the model is to be updated (eg, which parameters, and indeed how many parameters, are to be revised) and whether additional predictors are to be included. Riley et al provide advice for this and other model development situations.[5]

ACKNOWLEDGEMENTS

We thank two anonymous reviewers for their constructive feedback that helped improve our article upon revision. We are grateful to all the families who took part in the ALSPAC study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists, and nurses. Lucinda Archer is supported by funding from the European Horizon 2020 Research and Innovation Programme under grant agreement No 777090. Kym Snell is funded by the National Institute for Health Research School for Primary Care Research (NIHR SPCR Launching Fellowship). Joie Ensor is funded by NIHR Clinical Trials Unit Support Funding, Supporting Efficient/Innovative Delivery of NIHR Research. Mohammed Hudda is supported by a British Heart Foundation PhD Studentship (FS/17/76/33286). Gary Collins is

supported by the NIHR Biomedical Research Centre, Oxford and Cancer Research UK programme grant (C49297/A27294). Lucinda Archer, Richard Riley and Kym Snell are supported by funding from the Evidence Synthesis Working Group, which is funded by the National Institute for Health Research School for Primary Care Research (NIHR SPCR) [Project Number 390]. The UK Medical Research Council and Wellcome (Grant ref: 102215/2/13/2) and the University of Bristol provide core support for the ALSPAC study.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analysed in this study.

ETHICS STATEMENTS

Ethical approval for the ALSPAC study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees. The views expressed are those of the authors and not necessarily those of the BHF, Cancer Research UK, the NHS, the NIHR, the Department of Health or the EU.

ORCID

Lucinda Archer https://orcid.org/0000-0003-2504-2613
Kym I. E. Snell https://orcid.org/0000-0001-9373-6591
Joie Ensor https://orcid.org/0000-0001-7481-0282
Mohammed T. Hudda https://orcid.org/0000-0001-7894-1159
Gary S. Collins https://orcid.org/0000-0002-2772-2316
Richard D. Riley https://orcid.org/0000-0001-8699-0735

REFERENCES

1. Riley RD, van der Windt D, Croft P, Moons KG, eds. *Prognosis Research in Healthcare: Concepts, Methods and Impact*. Oxford, UK: Oxford University Press; 2019.
2. Harrell FE Jr. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Second ed. New York: Springer; 2015.
3. Steyerberg EW. *Clinical Prediction Models: a Practical Approach to Development, Validation, and Updating*. New York: Springer; 2009.
4. Riley RD, Snell KIE, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: part I - continuous outcomes. *Stat Med*. 2019;38(7):1262-1275. https://doi.org/10.1002/sim.7993.
5. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ*. 2020;368:m441.
6. Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. *BMJ*. 2009;338:b605.
7. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med*. 1999;130(6):515-524. https://doi.org/10.7326/0003-4819-130-6-199903160-00016.
8. Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. *Ann Intern Med*. 2006;144(3):201-209.
9. Toll DB, Janssen KJ, Vergouwe Y, Moons KG. Validation, updating and impact of clinical prediction rules: a review. *J Clin Epidemiol*. 2008;61(11):1085-1094. https://doi.org/10.1016/j.jclinepi.2008.04.008.
10. Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol*. 2015;68(3):279-289. https://doi.org/10.1016/j.jclinepi.2014.06.018.
11. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol*. 2016;74:167-176. https://doi.org/10.1016/j.jclinepi.2015.12.005.
12. Royston P, Altman DG. External validation of a Cox prognostic model: principles and methods. *BMC Med Res Methodol*. 2013;13:33. https://doi.org/10.1186/1471-2288-13-33.
13. Bleeker SE, Moll HA, Steyerberg EW, et al. External validation is necessary in, prediction research: a clinical example. *J Clin Epidemiol*. 2003;56(9):826-832. https://doi.org/10.1016/S0895-4356(03)00207-5.
14. Steyerberg EW, Harrell FE Jr. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol*. 2016;69:245-247. https://doi.org/10.1016/j.jclinepi.2015.04.005.
15. Collins GS, Altman DG. An independent and external validation of QRISK2 cardiovascular disease risk score: a prospective open cohort study. *BMJ*. 2010;340:c2442. https://doi.org/10.1136/bmj.c2442.
16. Collins GS, de Groot JA, Dutton S, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol*. 2014;14:40. https://doi.org/10.1186/1471-2288-14-40.
17. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. *Stat Med*. 2016;35(2):214-226. https://doi.org/10.1002/sim.6787.
18. Riley RD, Ensor J, Snell KI, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *BMJ*. 2016;353:i3140. https://doi.org/10.1136/bmj.i3140.
19. Steyerberg EW, Moons KG, van der Windt DA, et al. Prognosis research strategy (PROGRESS) 3: prognostic model research. *PLoS Med*. 2013;10(2):e1001381. https://doi.org/10.1371/journal.pmed.1001381.

20. Hudda MT, Fewtrell MS, Haroun D, et al. Development and validation of a prediction model for fat mass in children and adolescents: meta-analysis using individual participant data. *BMJ*. 2019;366:l4293. https://doi.org/10.1136/bmj.l4293.
21. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17(1):230. https://doi.org/10.1186/s12916-019-1466-7.
22. Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. *Stat Med*. 2014;33(3):517-535. https://doi.org/10.1002/sim.5941.
23. Copas JB. Regression, prediction and shrinkage. *J R Stat Soc B Methodol*. 1983;45(3):311-354.
24. Copas JB. Using regression models for prediction: shrinkage and regression to the mean. *Stat Methods Med Res*. 1997;6(2):167-183. https://doi.org/10.1177/096228029700600206.
25. Stein C. Inadmissibility of the usual estimator of the mean of a multivariate normal distribution. *Proc Third Berkeley Symp Math Stat Prob*. 1956;1:197-206.
26. Van Houwelingen JC. Shrinkage and penalized likelihood as methods to improve predictive accuracy. *Statistica Neerlandica*. 2001;55:17-34.
27. Kirchner J. Data Analysis Toolkit #10: Simple linear regression. http://seismo.berkeley.edu/~kirchner/eps_120/Toolkits/Toolkit_10.pdf. 1996
28. Tan L. Confidence Intervals for Comparison of the Squared Multiple Correlation Coefficients of Non-nested Models. Electronic Thesis and Dissertation Repository (Paper 384). 2012
29. Wishart J. The mean and second moment coefficient of the multiple correlation coefficient in samples from a normal population. *Biometrika*. 1931;22:353-361.
30. Lee YS. Tables of the upper percentage points of the multiple correlation. *Biometrika*. 1971;59:175-189.
31. Kelley K. Confidence intervals for standardized effect sizes: theory, application, and implementation. *J Stat Softw*. 2007;20(8):24. https://doi.org/10.18637/jss.v020.i08.
32. Kelley K. Methods for the behavioral, educational, and social sciences: an R package. *Behav Res Methods*. 2007;39(4):979-984. https://doi.org/10.3758/bf03192993.
33. Kelley K. MBESS (Version 4.0.0 and higher) [computer software and manual]. https://CRAN.R-project.org/package=MBESS. 2017.
34. Montgomery DC, Peck EA, Vining GG. *Introduction to Linear Regression Analysis*. Third ed. New York: Wiley; 2001.
35. Boyd A, Golding J, Macleod J, et al. Cohort profile: the 'children of the 90s'—the index offspring of the Avon longitudinal study of parents and children. *Int J Epidemiol*. 2013;42(1):111-127. https://doi.org/10.1093/ije/dys064.
36. Fraser A, Macdonald-Wallis C, Tilling K, et al. Cohort profile: the Avon longitudinal study of parents and children: ALSPAC mothers cohort. *Int J Epidemiol*. 2013;42(1):97-110. https://doi.org/10.1093/ije/dys066.
37. Wan X, Wang W, Liu J, Tong T. Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. *BMC Med Res Methodol*. 2014;14(1):135. https://doi.org/10.1186/1471-2288-14-135.
38. Riley RD, Snell KI, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: part II - binary and time-to-event outcomes. *Stat Med*. 2019;38(7):1276-1296. https://doi.org/10.1002/sim.7992.
39. Debray TP, Damen JA, Snell KI, et al. A guide to systematic review and meta-analysis of prediction model performance. *BMJ*. 2017;356:i6460. https://doi.org/10.1136/bmj.i6460.
40. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. *Ann Intern Med*. 2019;170(1):W1-W33. https://doi.org/10.7326/M18-1377.
41. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. 2019;170(1):51-58. https://doi.org/10.7326/M18-1376.
42. Austin PC, Steyerberg EW. Bootstrap confidence intervals for loess-based calibration curves. *Stat Med*. 2014;33(15):2699-2700. https://doi.org/10.1002/sim.6167.
43. Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KG. Internal and external validation of predictive models: a simulation study of bias and precision in small samples. *J Clin Epidemiol*. 2003;56(5):441-447.
44. Jinks RC, Royston P, Parmar MK. Discrimination-based sample size calculations for multivariable prognostic models for time-to-event data. *BMC Med Res Methodol*. 2015;15:82. https://doi.org/10.1186/s12874-015-0078-y.

How to cite this article: Archer L, Snell KIE, Ensor J, Hudda MT, Collins GS, Riley RD. Minimum sample size for external validation of a clinical prediction model with a continuous outcome. *Statistics in Medicine*. 2021;40:133–146. https://doi.org/10.1002/sim.8766

Appendix IV - Chapter 5

Appendix IVa - External validation performance on the Poon 2011 model on the $\log_{10}$ grams scale

Included items:

- Figure 8.1: Distributions of Expected and Observed birthweights ($\log_{10} grams$), by external validation cohort.
- Figure 8.2: Calibration plots for the Poon 2011 model when assessed on the $\log_{10} grams$ scale, by external validation cohort.
- Figure 8.3: Forest plot for the calibration slope of the Poon 2011 model, when assessed on the $\log_{10} grams$ scale.
- Figure 8.4: Forest plot for the CITL of the Poon 2011 model, when assessed on the $\log_{10} grams$ scale.

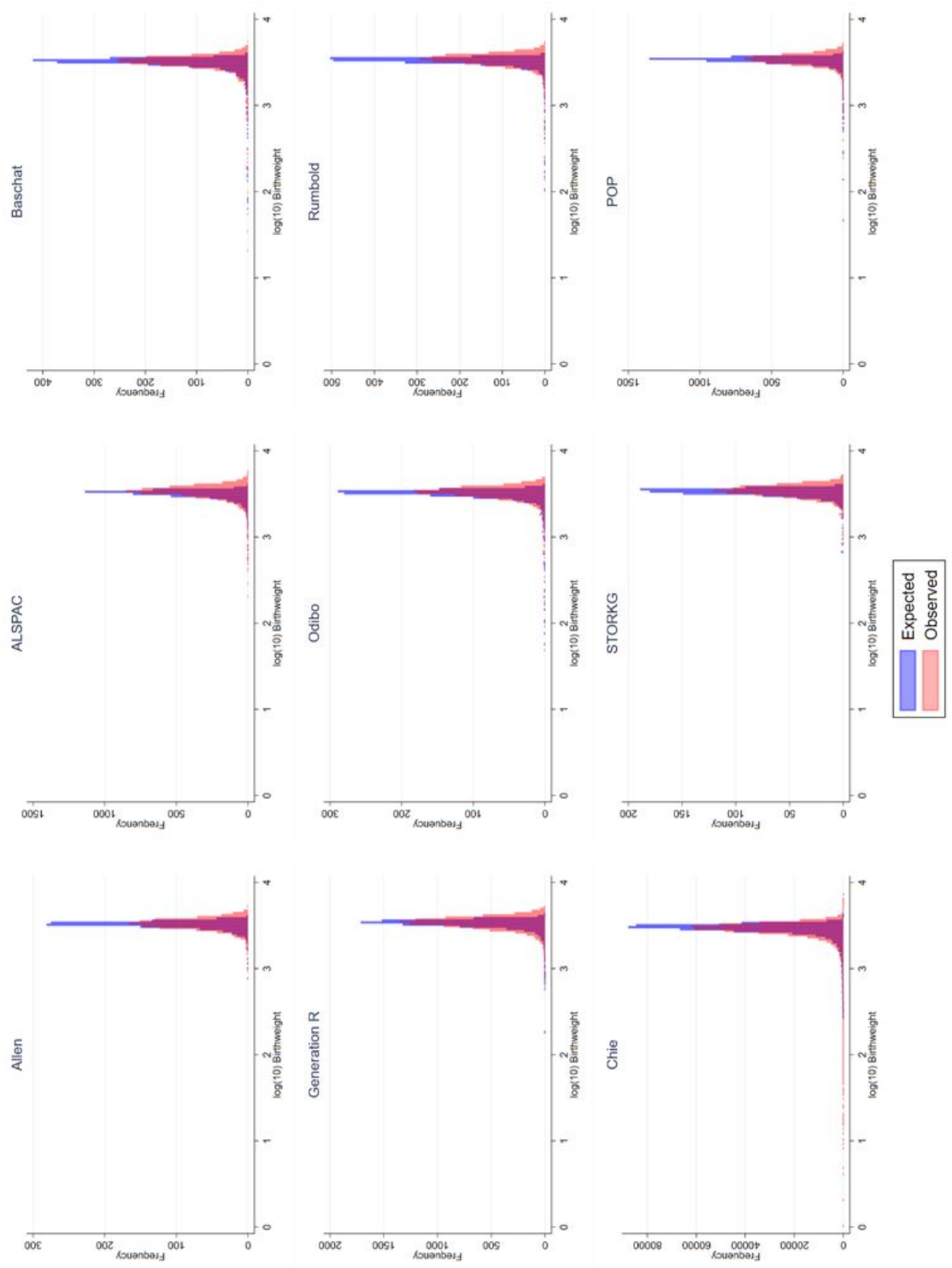
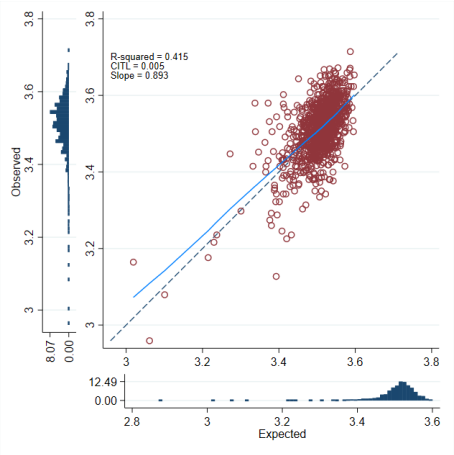
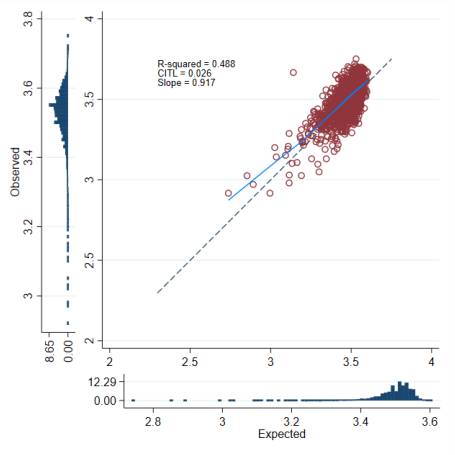


Figure 8.1: Distributions of Expected and Observed birthweights ($\log_{10} grams$), by external validation cohort.

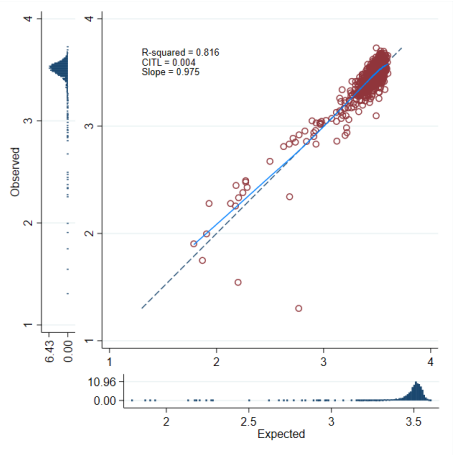
Figure 8.2: Calibration plots for the Poon 2011 model when assessed on the $\log_{10} grams$ scale, by external validation cohort.



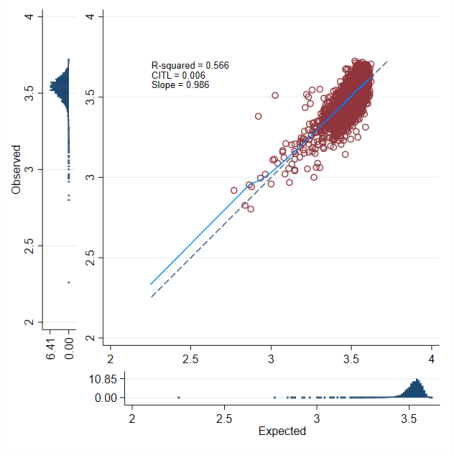
(a) Allen



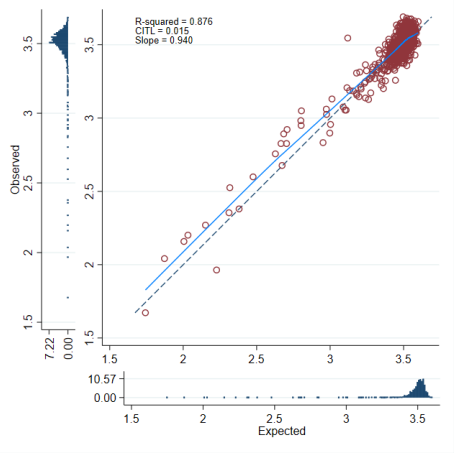
(b) ALSPAC



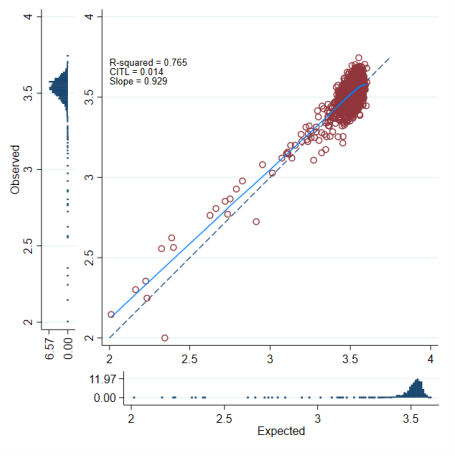
(c) Baschat



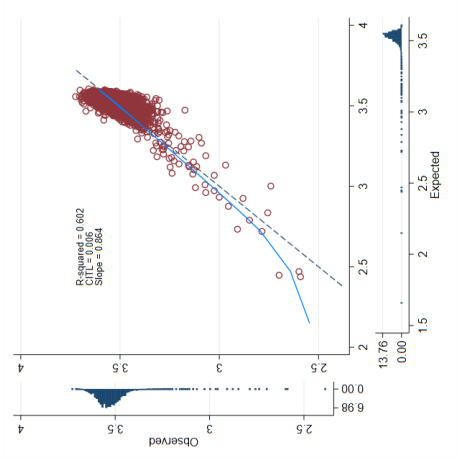
(d) Generation R



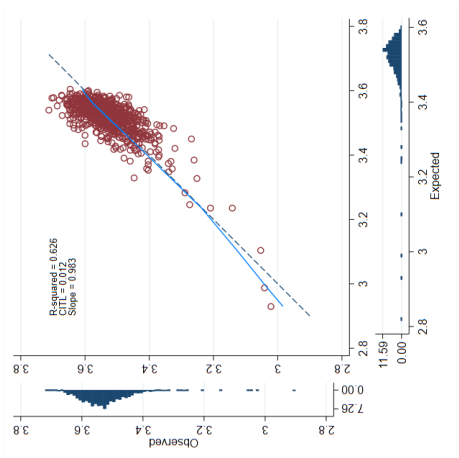
(e) Odibo



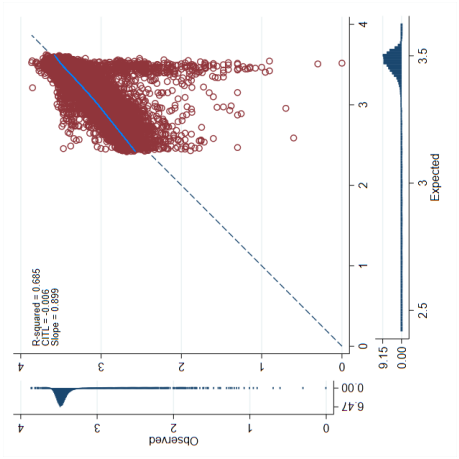
(f) Rumbold



(i) POP



(h) STORKG



(g) Chie

Figure 8.3: Forest plot for the calibration slope of the Poon 2011 model, when assessed on the $\log_{10} grams$ scale.

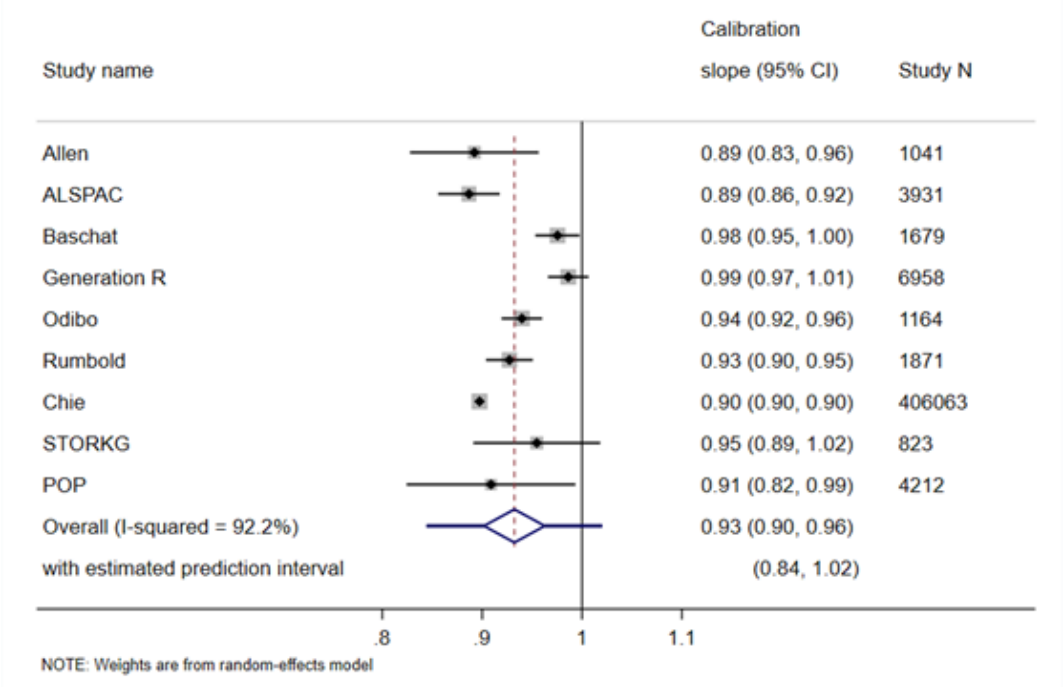
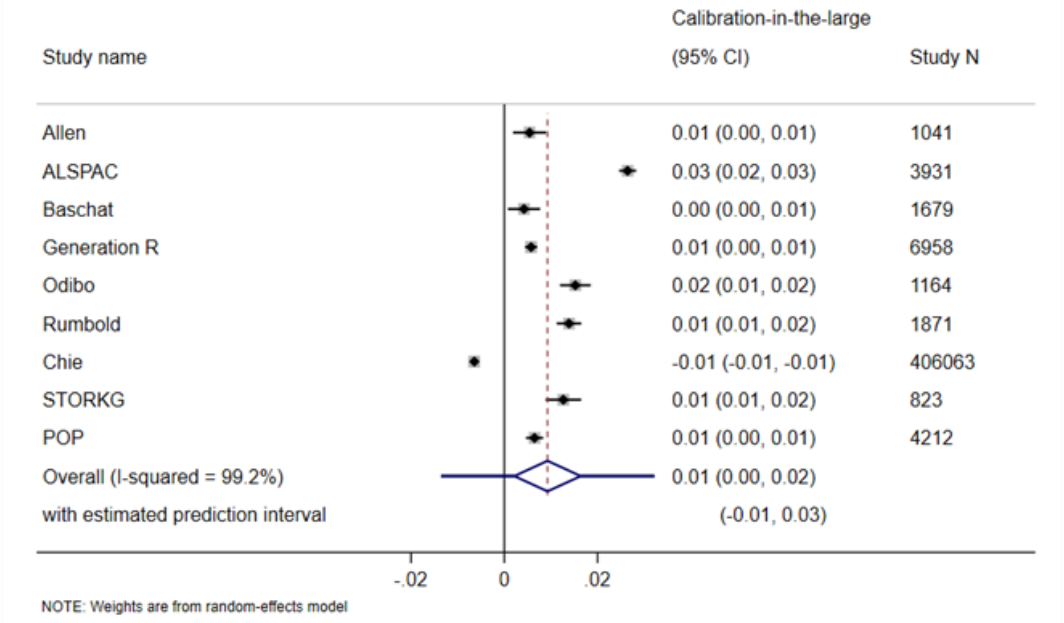


Figure 8.4: Forest plot for the CITL of the Poon 2011 model, when assessed on the $\log_{10} grams$ scale.



Appendix IVb - External validation performance on the Poon 2011 model: complete case analysis

Calibration slope

The pooled calibration slope across all cohorts of 0.97 (95% CI: 0.93 to 1.02) is consistent with that seen in the imputed analysis, vary only in the slightly wider confidence intervals. Complete case analysis continues to imply near ideal calibration of the Poon 2011 model across studies. The 95% prediction interval implies that the calibration slope expected in a new cohort has a 95% chance of falling between 0.86 and 1.08. This is also consistent with the results found in the imputation analysis. The only cohort for which there was a substantial difference in the calibration slope was Rumbold, where confidence intervals were very wider, due to low numbers available in the complete case analysis.

Figure 8.5: Forest plot for the calibration slope of the Poon 2011 model, when assessed only in the complete case data

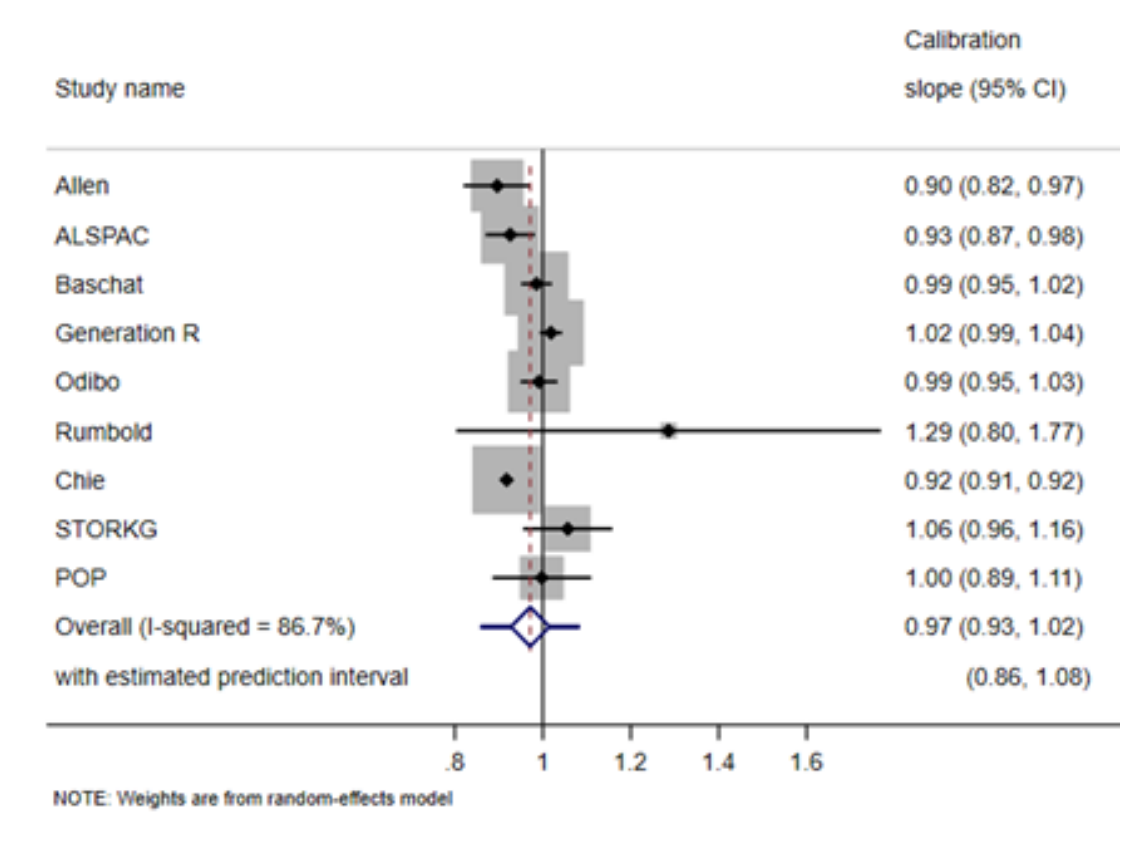
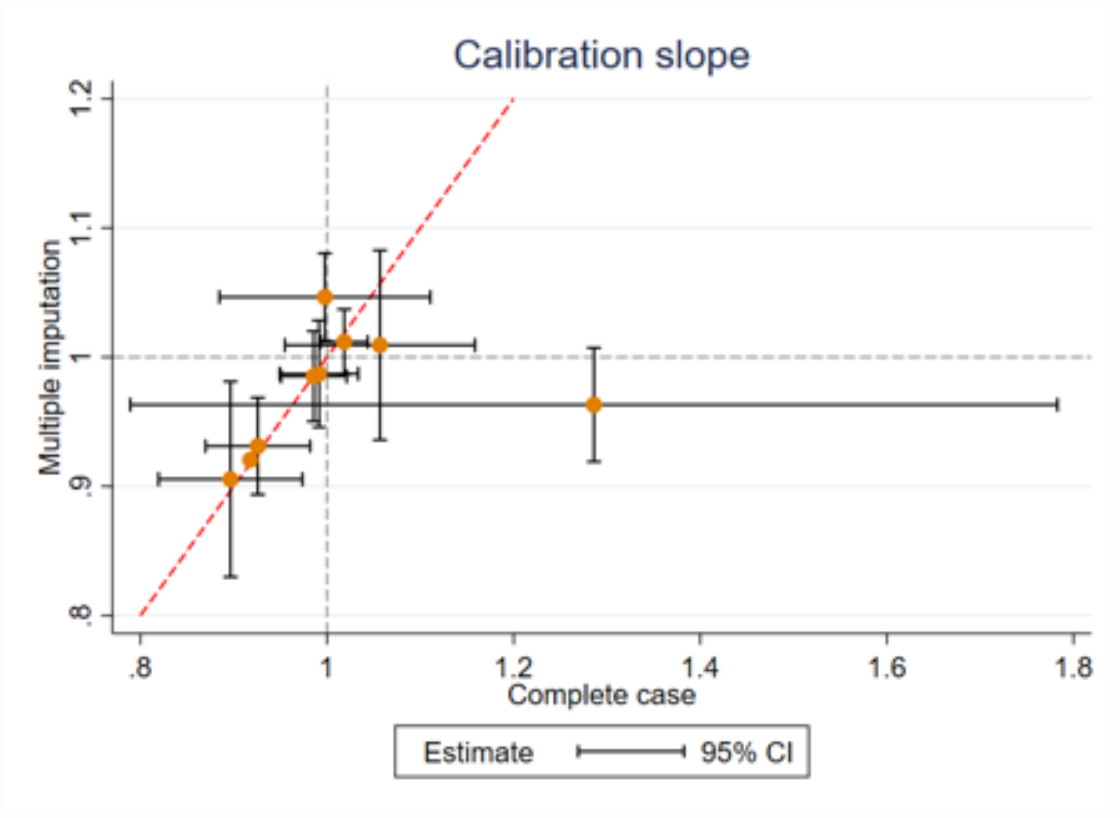


Figure 8.6: Comparison of the calibration slope in complete case and multiply imputed data. The grey lines indicate ideal value and the red line shows perfect agreement.



Calibration-in-the-large

On average, calibration-in-the-large for the complete case analysis was 89.00g (35.12g to 142.89g), which is consistent with the 90.39g calibration in the large seen in the analysis of the imputed data. Consistency in performance across analysis types was more consistent for calibration-in-the-large than the calibration slope, with most studies lying close the line of agreement.

Figure 8.7: Forest plot for the CITL of the Poon 2011 model, when assessed only in the complete case data

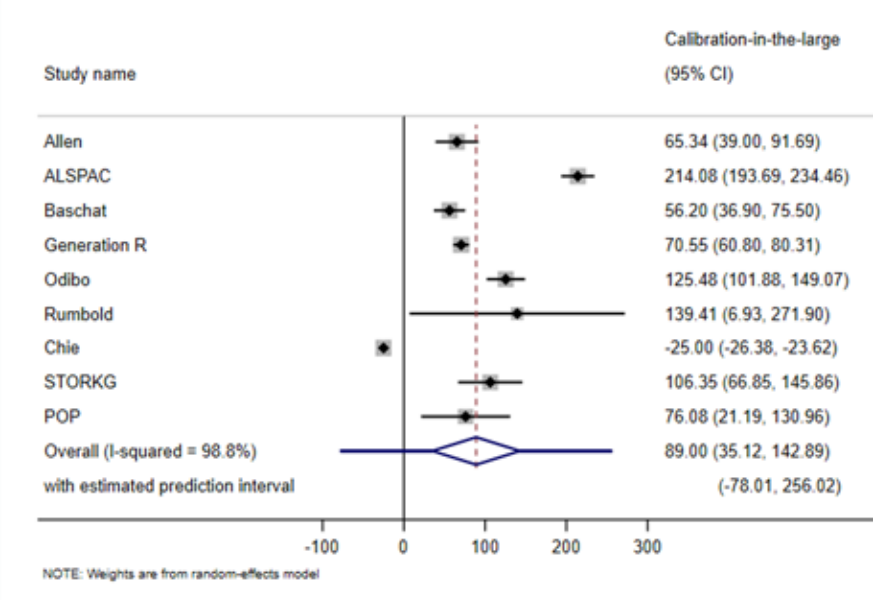
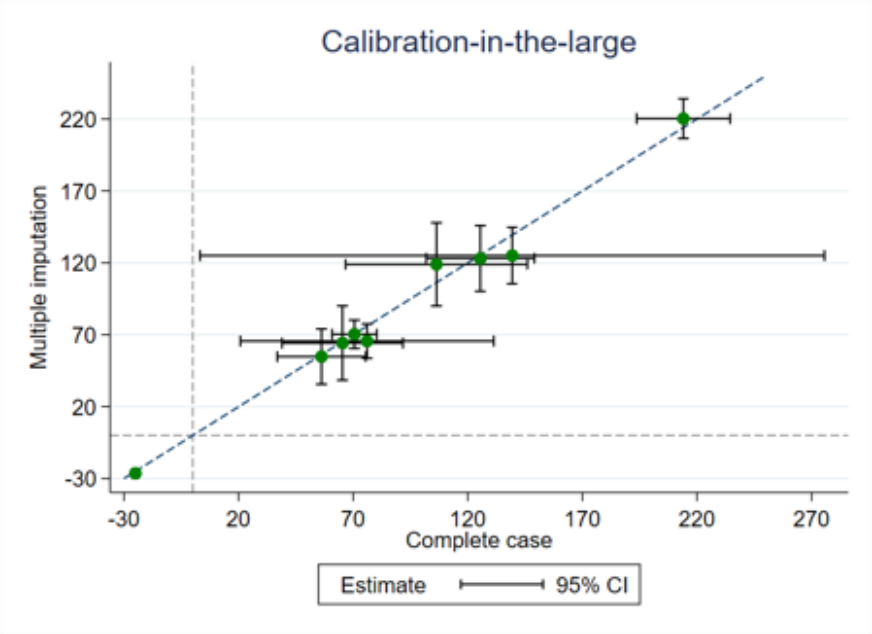


Figure 8.8: Comparison of CITL in complete case and multiply imputed data. The grey lines indicate ideal value and the red line shows perfect agreement.



Appendix IVc - Prediction models for FGR and birthweight: subsequent research

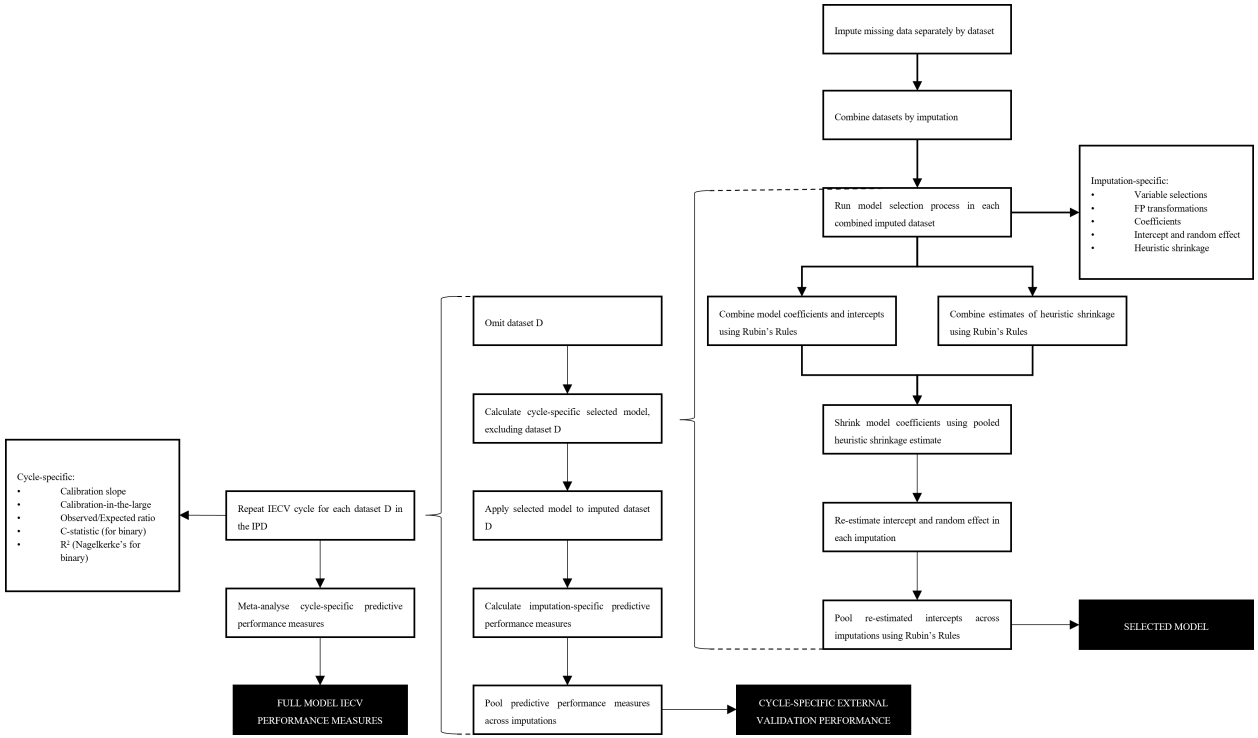
Following the literature review and external validation of existing models discussed in this chapter, it was concluded that the current literature was lacking in an acceptable model for the prediction of FGR, either directly or through use of birthweight as a proxy. Therefore, the next steps of this research project were to develop and validate a new prediction model for identifying the risk of delivering a growth restricted baby (as defined above), using data from the IPPIC collaboration data collection. Though the Poon 2011 model showed reasonable calibration performance on average across individuals, model calibration was highly variable on the individual level, thus a second model aiming to more consistently predict birthweight was also developed. Due to its satisfactory performance on average, the Poon 2011 model was used as a starting point for the development of both new models, with all predictor variables from this model being included as candidate predictors, in addition to predictors identified through consultation with clinical experts.

Model development combined IPD meta-analysis methods with multiple imputation, variable selection, and assessment of non-linear predictor-outcome relationships. An internal-external cross-validation (IECV) approach was used for model validation, as model development involved the combined IPD from multiple cohorts [268, 269]. Figure 8.9 summarises the main steps in this process, which involved the development and external validation of multiple example models across subgroups of the available datasets. For each cycle of the approach, an example model was derived using the same development processes as for the full model, using the data from all but one cohort. The reserved cohort was then used for external validation of this cycle-specific prediction model, giving cycle-specific estimates of predictive performance.

Following all IECV cycles, there were multiple values for each predictive performance measure (one

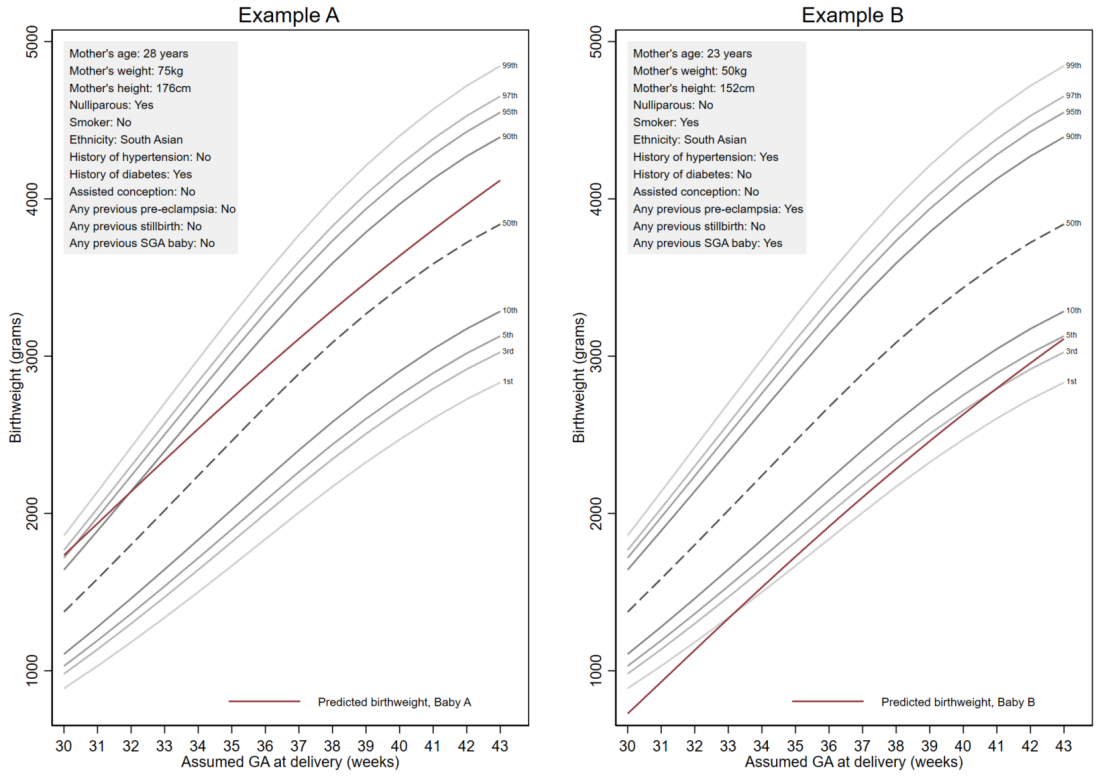
for each cohort from the full model development data, when that cohort was reserved for external validation). These estimates were summarised across cycles using random-effects meta-analysis to give pooled estimates for the full model's predictive performance, and to assess the consistency of model performance across populations and the anticipated generalisability of the estimated predictive ability of the full model to new settings. As with the external validation seen in this chapter, pooling of estimates across IECV cycles resulted in wider confidence intervals and thus less certainty in performance than in any individual cycle, reflecting the varying populations and heterogeneity in model performance across cohorts.

Figure 8.9: Flow diagram showing processes involved in development and validation of the IPPIC-FGR prediction models



An aim of this new model development was to allow predictions conditional on some assumed gestational age at delivery, allowing assessment of anticipated FGR risk or continuous birthweight if a baby were to be born at various different stages of gestation. Thus, gestational age at delivery was included as a continuous predictor in both models, as in the Poon 2011 model. Although the observed gestational age at delivery would not be available at the moment of prediction, producing the models in this way allowed for a range of potential gestational ages at delivery to be assessed for each pregnancy, with plots of predictions against gestational age ("Georgie plots") to give a more complete assessment of risk over time, as shown in Figure 8.10.

Figure 8.10: Predicted birthweight curves for two example pregnancies, compared to population percentile curves, for assumed gestational ages at delivery between 30 and 43 weeks



Development and external validation of a risk prediction model for falls in patients with an indication for antihypertensive treatment: retrospective cohort study

Lucinda Archer,[1] Constantinos Koshiaris,[2] Sarah Lay-Flurrie,[2] Kym I E Snell,[1] Richard D Riley,[1] Richard Stevens,[2] Amitava Banerjee,[3] Juliet A Usher-Smith,[4] Andrew Clegg,[5] Rupert A Payne,[6] F D Richard Hobbs,[2] Richard J McManus,[2] James P Sheppard,[2] on behalf of the STRAtifying Treatments In the multi-morbid Frail elderlY (STRATIFY) investigators

[1]Centre for Prognosis Research, School of Medicine, Keele University, Keele, UK
[2]Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, OX2 6GG, UK
[3]Institute of Health Informatics, University College London, London, UK
[4]Primary Care Unit, Department of Public Health and Primary Care, University of Cambridge, UK
[5]Academic Unit for Ageing and Stroke Research, Bradford Institute for Health Research, University of Leeds, UK
[6]Centre for Academic Primary Care, Population Health Sciences, University of Bristol, Bristol, UK

Correspondence to: J P Sheppard
james.sheppard@phc.ox.ac.uk
(ORCID 0000-0002-4461-8756)

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2022;379:e070918
http://dx.doi.org/10.1136/bmj-2022-070918

Accepted: 21 September 2022

ABSTRACT

OBJECTIVE

To develop and externally validate the STRAtifying Treatments In the multi-morbid Frail elderlY (STRATIFY)-Falls clinical prediction model to identify the risk of hospital admission or death from a fall in patients with an indication for antihypertensive treatment.

DESIGN

Retrospective cohort study.

SETTING

Primary care data from electronic health records contained within the UK Clinical Practice Research Datalink (CPRD).

PARTICIPANTS

Patients aged 40 years or older with at least one blood pressure measurement between 130 mm Hg and 179 mm Hg.

MAIN OUTCOME MEASURE

First serious fall, defined as hospital admission or death with a primary diagnosis of a fall within 10 years of the index date (12 months after cohort entry). Model development was conducted using a Fine-Gray approach in data from CPRD GOLD, accounting for the competing risk of death from other causes, with subsequent recalibration at one, five, and 10 years using pseudo values. External validation was conducted using data from CPRD Aurum, with performance assessed through calibration curves and the observed to expected ratio, C statistic, and D statistic, pooled across general practices, and clinical utility using decision curve analysis at thresholds around 10%.

RESULTS

Analysis included 1 772 600 patients (experiencing 62 691 serious falls) from CPRD GOLD used in model development, and 3 805 366 (experiencing 206 956 serious falls) from CPRD Aurum in the external validation. The final model consisted of 24 predictors, including age, sex, ethnicity, alcohol consumption, living in an area of high social deprivation, a history of falls, multiple sclerosis, and prescriptions of antihypertensives, antidepressants, hypnotics, and anxiolytics. Upon external validation, the recalibrated model showed good discrimination, with pooled C statistics of 0.833 (95% confidence interval 0.831 to 0.835) and 0.843 (0.841 to 0.844) at five and 10 years, respectively. Original model calibration was poor on visual inspection and although this was improved with recalibration, under-prediction of risk remained (observed to expected ratio at 10 years 1.839, 95% confidence interval 1.811 to 1.865). Nevertheless, decision curve analysis suggests potential clinical utility, with net benefit larger than other strategies.

CONCLUSIONS

This prediction model uses commonly recorded clinical characteristics and distinguishes well between patients at high and low risk of falls in the next 1-10 years. Although miscalibration was evident on external validation, the model still had potential clinical utility around risk thresholds of 10% and so could be useful in routine clinical practice to help identify those at high risk of falls who might benefit from closer monitoring or early intervention to prevent future falls. Further studies are needed to explore the appropriate thresholds that maximise the model's clinical utility and cost effectiveness.

WHAT IS ALREADY KNOWN ON THIS TOPIC

Serious falls are a possible side effect of antihypertensive treatment, which can adversely affect patients' quality of life and increase the risk of hospital admission, especially in older people with frailty

Existing tools that estimate an individual's risk of falls have been shown to be at high risk of bias, with only moderate discriminative ability

WHAT THIS STUDY ADDS

In the present study, a clinical prediction model for the risk of falls for up to 10 years was developed and externally validated, incorporating commonly recorded patient characteristics, comorbidities, and drugs, in patients with an indication for antihypertensive treatment

Upon external validation, the model discriminated well between patients who went on to have a serious fall and those who did not, but calibration indicated under-prediction of risk

Nevertheless, a decision curve analysis suggests the model has clinical utility and so may be useful to identify patients with a high fall risk, who may require closer monitoring or early intervention to prevent future falls

Introduction

The proportion of older adults in the population is rising,[1] and with age the risk of falls increases,[2 3] which can result in serious injury and long term disability.[4] In England, falls are associated with about 235 000 emergency hospital admissions in the over 65s and cost the National Health Service more than £2.3bn ($2.6bn; €2.6bn) every year.[5-7]

Many risk factors for falls exist, primarily related to comorbidities and frailty.[2 3 8-10] A key modifiable risk factor is prescribed drugs, including those that lower blood pressure.[11-13] Although antihypertensives are effective at reducing the risk of cardiovascular disease, typically many patients require treatment over several years to prevent a small number of events.[14] Data from randomised controlled trials show that antihypertensives are associated with an increased risk of hypotension and syncope, which may lead to falls.[15] Observational studies examining patients with frailty and multimorbidity suggest a direct association between antihypertensive treatment and falls.[11 16 17]

In patients who are prescribed antihypertensives or other drugs that substantially increase their risk of falls, doctors might want to consider altering or withdrawing treatment (ie, deprescribing),[18] along with other interventions to reduce the risk of falls (eg, advice on lower alcohol consumption, falls prevention clinics, exercises).[7] Identifying people at high risk of falls is, however, challenging. A 2021 systematic review of falls prediction models for use in the community identified a total of 72 models.[10] Most of these studies were deemed at high risk of bias, and only three of the models were externally validated. These three validated models showed moderate discriminative ability, with an area under the curve of between 0.62 and 0.69. Calibration based on internal validation was only reported in seven of the studies, and it was typically moderate to poor.[10] A further primary analysis aiming to predict falls in a general practice population showed good apparent discrimination for the model used (with an area under the curve of 0.87), but calibration performance was not assessed and no external validation was performed.[19]

To inform clinical decision making in primary care, both patients and doctors require better prediction models to accurately identify those at high risk of serious falls (defined as any fall resulting in hospital admission or death), from the population of older adults who might be considered for antihypertensive treatment. This population includes patients with a recent high blood pressure reading, including those with a new diagnosis of hypertension, as well as those in whom intensification of treatment is being considered. We used routinely collected data from electronic health records to develop and externally validate a clinical prediction model to estimate such individuals' risk of experiencing a fall resulting in hospital admission or death within 10 years. This study is part of a broader research programme investigating the association between blood pressure lowering drugs and side effects: STRAtifying Treatments In the multimorbid Frail elderlY (STRATIFY): Antihypertensives.

Methods

A retrospective observational cohort study was used to develop a prediction model for serious falls (the STRATIFY-Falls model), using data from Clinical Practice Research Datalink (CPRD) GOLD, which contains information from general practices using Vision electronic health record software (Cegedim Healthcare Solutions, London, UK). The model was externally validated using a second retrospective observational cohort comprising data from CPRD Aurum, containing data from general practices using recording software from Egton Medical Information Systems (EMIS, Leeds, UK). These data were linked to Office for National Statistics mortality data, Hospital Episode Statistics, and index of multiple deprivation data. The CPRD independent scientific advisory committee approved the protocol for this study (protocol No 19_042, see Appendix 6 in the supplementary material).

Population

Patients were eligible if they were registered at a linked general practice in England, contributing to CPRD between 1 January1998 and 31 December 2018. At the time of analysis, CPRD GOLD (development cohort) contained 4.4 million active patients from 674 general practices, whereas CPRD Aurum (validation cohort) contained seven million active patients from 738 practices. Both datasets have previously been shown to be representative of the patient population in England for age, ethnicity, and deprivation status.[20 21] To avoid duplication of patients, when practices had switched from one recording system to the other during the study timeframe, we excluded practices from CPRD Aurum (validation cohort) that were also present in the CPRD GOLD (development) dataset.

Patients were considered eligible if they were aged 40 years or older (no upper age limit applied), registered to a CPRD "up-to-standard" practice (CPRD GOLD only), and had records available during the study period. Patients entered the cohorts at the time at which they became potentially eligible for antihypertensive treatment (ie, at the time of their first systolic blood pressure reading ≥130 mm Hg) after the study start date, and they were followed for up to 10 years. This blood pressure threshold was chosen to account for varying treatment initiation thresholds specified in different international hypertension guidelines.[6] Patients with any systolic blood pressure reading >180 mm Hg were excluded from the cohort, as antihypertensive treatment would be indicated for these patients regardless of the risk of adverse events, unless clearly contraindicated for other reasons. All patient characteristics and model predictors were determined at the index date, defined as 12 months after cohort entry. The same eligibility criteria and characteristic determination methods were applied to both the development cohort and the validation cohort.

Outcomes

The primary outcome was any hospital admission or death associated with a primary diagnosis of a fall within 10 years of the index date, the same time horizon as used for cardiovascular prediction models.[22] Falls were based on ICD-10 (international classification of diseases, 10th revision) codes

doi: 10.1136/bmj-2022-070918 | BMJ 2022;379:e070918 | the bmj

documented in Hospital Episodes Statistics and ONS mortality data (applicable ICD-10 codes shown in supplementary table S4.1). Prespecified secondary outcomes were falls (defined in the same way) within one and five years of the index date. This outcome definition was consistent across both the development cohort and the validation cohort.

Model predictors

We identified clinically relevant predictors of falls from the literature and through expert clinical opinion.[2 7-9 23] These included 30 predictors (44 predictor variables), covering patient demographics (age, sex, ethnicity, area based socioeconomic deprivation (index of multiple deprivation), body mass index (BMI), systolic and diastolic blood pressure), clinical characteristics (total cholesterol level, smoking status, alcohol intake), comorbidities (previous falls, memory problems, mobility issues, history of stroke, multiple sclerosis, activity limitation, syncope, cataract), and prescribed drugs (antihypertensives, opioids, hypnotics or benzodiazepines, antidepressants, anticholinergics) (see table S4.2 in the supplementary material). A recent literature review of falls clinical prediction tools by the National Institute for Health and Care Excellence identified the need for frailty to be considered as a predictor in models for use in the community.[24] We therefore also calculated a validated electronic frailty index using the 36 comorbidities and conditions specified, including this index as a single covariate.[25] Covariates were defined by any occurrence of relevant Read or SNOMED codes at any time point before the index date, with the exception of antihypertensives, which were defined as any prescription in the 12 months before the index date.

To ensure consistency with commonly used risk calculators,[26 27] our prediction models do not account for changes in prescriptions of drug type or amount over time, and as such give an estimation of falls risk assuming treatment assignment policy in any application setting is similar to that in the development data.[28]

Final model equations, five and 10 years:

$$\ln\left(\frac{p_{5y}}{1-p_{5y}}\right) = \alpha_{5y} + (\beta_{5y} \times LP) + (\gamma_{5y} \times (LP) \times \ln(LP))$$

$$\ln\left(\frac{p_{10y}}{1-p_{10y}}\right) = \alpha_{10y} + (\beta_{10y} \times LP) + (\gamma_{10y} \times (LP) \times \ln(LP))$$

Final model equation for one year:

$$1\text{ year risk} = 1 - (1 - CIF_{1y})^{\exp(LP)}$$

Where p_{5y} and p_{10y} are the predicted probabilities of a fall at five and 10 years, respectively; and LP is the linear predictor from the original model, as shown in table 2:

$$LP_i = \left(\left(\frac{Age}{100}\right)^3 - 0.242\right) \times 4.102 + \begin{cases} 0.223 \text{ if female} \\ 0 \text{ if male} \end{cases}$$

$$+ \left(\left(\frac{TC}{10}\right)^{-0.5} - 1.381\right) \times 0.393$$

$$+ \begin{cases} -0.425 \text{ if black ethnicity} \\ -0.381 \text{ if South Asian ethnicity} \\ -0.352 \text{ if other ethnicity} \end{cases} + \begin{cases} 0.038 \text{ if IMD2} \\ 0.072 \text{ if IMD3} \\ 0.169 \text{ if IMD4} \\ 0.229 \text{ if IMD5} \end{cases} + \begin{cases} 0.114 \text{ if former smoker} \\ 0.236 \text{ if current smoker} \end{cases}$$

$$+ \begin{cases} -0.105 \text{ if occasional drinker} \\ -0.065 \text{ if light drinker} \\ -0.009 \text{ if moderate drinker} \\ 0.451 \text{ if heavy drinker} \\ -0.068 \text{ if drinker (unknown quantity)} \end{cases} + \left(\left(\frac{FI}{0.1}\right) - 0.576\right) \times 0.197$$

$$+ (0.275 \text{ if previous falls}) + (0.155 \text{ if memory problems})$$
$$+ (-0.08 \text{ if mobility problems}) + (0.11 \text{ if using ACE inhibitors})$$
$$+ (0.17 \text{ if using angiotensin 2 receptor blockers})$$
$$+ (0.082 \text{ if using calcium channel blockers}) + (0.071 \text{ if using diuretics})$$
$$+ (0.068 \text{ if using } \beta \text{ blockers}) + (0.041 \text{ if using } \alpha \text{ blockers})$$
$$+ (-0.045 \text{ if using other hypertensives}) + (0.101 \text{ if using opioids})$$
$$+ (0.142 \text{ if using hypnotics/anxiolytics}) + (0.146 \text{ if using antidepressants})$$
$$+ (0.034 \text{ if using anticholinergic}) + (0.133 \text{ if history of stroke})$$
$$+ (0.537 \text{ if history of multiple sclerosis})$$

Fig 1 | Final model equations for predicting risk of falls at one, five, and 10 years in patients with an indication for hypertensive treatment. Age is measured in years. Ln=natural logarithm; IMD2-IMD5=indices of multiple deprivation; TC=total cholesterol; FI=electronic frailty index. The full algorithm code (including the α, β, γ, and CIF values) is freely available for research use and can be downloaded at https://process.innovation.ox.ac.uk/software/

Sample size

The prespecified sample size calculation for model development was 2194 participants (15 358 person years), assuming a maximum of 40 predictors would be included in the final model (see extended methods in the supplementary material).[29] For the external validation, the estimated sample size required was 12 000 patients (with at least 708 experiencing falls), sufficient to target a 95% confidence interval of width 0.2 around the estimate of the calibration slope (see extended methods in the supplementary material).[30] The actual sample sizes in both the development cohort and the validation cohort far exceeded these estimates.

Statistical analysis

We calculated descriptive statistics for baseline characteristics in the model development and external validation cohorts separately.

Missing data

Multiple imputation with chained equations was used to impute missing data in both the development cohort and the validation cohort, with 10 imputations generated for the development and validation datasets. Two separate and independent imputation procedures were used, one for model development and one for model validation. The imputation models included all model covariates within each dataset, along with the Nelson-Aalen estimator for the cumulative baseline cause specific hazards for falls and for the competing event of death, and binary event indicators for each of these possible event types.[31 32] When information was missing on the diagnosis of comorbidities or prescribed drugs, it was assumed that no diagnosis or prescription was present. Predictor variables requiring imputation were cholesterol, ethnicity, deprivation score (validation cohort only), smoking status, and alcohol consumption.

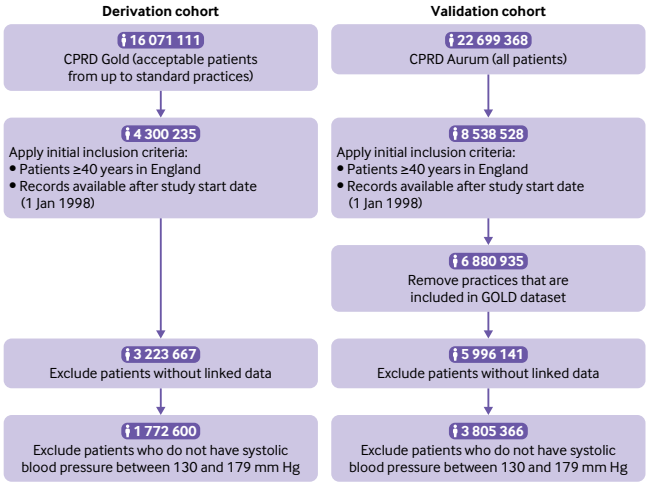


Fig 2 | Flow of participants through study. CPRD=Clinical Practice Research Datalink

Imputations were assessed for consistency by comparing density plots, histograms, and summary statistics across imputations and back to the complete values. The model coefficients and predictive performance measures were then estimated in each imputed dataset separately, before being combined across imputations using Rubin's rules.[33]

Model development

Researchers at the University of Oxford (CK, JPS) conducted the model development and apparent validation. Multivariable prediction models were fitted in each imputed dataset using a Fine-Gray subdistribution hazard model, taking into account the competing risk of death by other causes.[34] The aim of accounting for the competing risk in this way was to avoid overestimation of the predicted probabilities of falls as defined in the Fine-Gray paper.[34 35] Predictor effects in the model are reported as subdistribution hazard ratios with 95% confidence intervals, and the post-estimation baseline cumulative incidence for falls was estimated using a Breslow type estimator.[34] Analyses were undertaken using the *fastcmprsk* package in RStudio.[36] Automated variable selection methods were not used, since the variables were all predetermined based on the literature and expert opinion, and given the large sample size would result in nearly all predictors having a statistically significant association with the outcome, regardless of effect size. To ensure a parsimonious model, we excluded variables with little or no association in multivariable analysis before fitting the final model.

Fractional polynomial terms were examined to identify the best fitting functional form of all continuous variables.[37] Fractional polynomials were identified separately within each imputed dataset, and we selected the most consistent transformation across the imputations, choosing lower order fractional polynomial terms whenever possible for the sake of parsimony. We then forced the selected fractional polynomial format for each continuous variable into the model for all imputations to ensure consistency in coefficient estimation.

Interactions between age, sex, and antihypertensive treatments were considered but excluded from the model development owing to problems with stability or convergence, or for the sake of parsimony.

We examined the Schoenfeld residuals to check the proportional hazards assumption for each predictor.[38]

Apparent validation using development data

Observed outcome probabilities were defined using pseudo values: jack-knife estimators representing an individual's contribution to the cumulative incidence function for falls, accounting for competing risk, calculated by the Aalen–Johansen method. Pseudo values were generated separately in 50 groups by linear predictor value, for stability, and to account for the competing risk of death and non-informative right censoring.[39 40]

The model's apparent calibration performance was assessed using calibration plots comparing the observed to predicted risks at one, five, and 10 years. The calibration plots were produced using observed pseudo values and included a smooth (non-linear) calibration curve to show apparent calibration across the spectrum of predicted risks,[41] with 95% confidence intervals. Plots were generated in each imputed dataset separately and were checked for consistency across imputations. A single, representative example is reported.

When plots showed miscalibration, we recalibrated the original Fine-Gray model separately at each

doi: 10.1136/bmj-2022-070918 | *BMJ* 2022;379:e070918 | the bmj

Table 1 | Descriptive statistics for model development and validation cohorts, in full cohorts and stratified by outcome type at 10 years. Values are numbers (percentages) unless stated otherwise

	Development cohort			Validation cohort		
Variables	Total (n=1 772 600)	Falls (n=62 691)	Mortality (n=181 731)	Total (n=3 805 366)	Falls (n=206 956)	Mortality (n=334 552)
Mean (SD) age (years)	59.4 (13.2)	73.6 (12.7)	74.3 (12.0)	58.6 (13.3)	72.8 (12.7)	73.1 (12.3)
Women	921 853 (52)	39 955 (64)	91 676 (50)	1 959 489 (52)	134 945 (65.2)	165 689 (49.5)
Systolic blood pressure (mm Hg)	143.5 (11.9)	146.3 (12.7)	146.9 (12.8)	143.8 (12.3)	147.2 (13.2)	147.7 (13.3)
Diastolic blood pressure (mm Hg)	83.8 (9.6)	81.6 (10.0)	81.7 (10.0)	83.9 (9.8)	81.9 (10.2)	82.0 (10.3)
Cholesterol (mmol/L)	5.3 (1.1)	5.2 (1.2)	5.2 (1.2)	5.5 (1.2)	5.4 (1.3)	5.4 (1.3)
Missing	868 461 (48.9)	32 661 (52)	104 094 (59)	1 839 116 (48.3)	109 708 (53.0)	195 390 (58.4)
Ethnicity:						
White	734 149 (41)	59 608 (95)	105 077 (40.5)	2 041 505 (54)	194 311 (93.9)	206 384 (61.7)
Black	10 799 (0.6)	339 (0.54)	826 (0.45)	115 279 (3)	2239 (1.1)	4019 (1.2)
South Asian	14 799 (0.8)	505 (0.81)	991 (0.55)	94 485 (3)	2449 (1.2)	3673 (1.1)
Other	15 731 (0.9)	587 (0.94)	1229 (0.68)	832 614 (22)	3442 (1.7)	21 458 (6.4)
Missing	997 122 (56)	1652 (2.6)	73 608 (57.8)	721 483 (19)	4515 (2.2)	99 018 (29.6)
Index of multiple deprivation score:						
1	420 765 (23.7)	12 624 (20.2)	35 529 (19.6)	790 311 (20.8)	41 786 (20.2)	66 606 (19.9)
2	406 775 (22.9)	13 429 (21.4)	39 652 (21.8)	732 246 (19.2)	41 820 (20.2)	68 147 (20.4)
3	376 765 (21.3)	13 239 (21.1)	39 279 (21.6)	684 288 (18)	40 665 (19.7)	67 130 (20.1)
4	313 595 (17.7)	12 031 (19.2)	35 183 (19.4)	630 482 (16.6)	40 383 (19.5)	65 342 (19.5)
5	254 700 (14.4)	11 317 (18.1)	31 909 (17.6)	597 180 (15.7)	42 141 (20.4)	67 024 (20.0)
Missing	0 (0)	0 (0)	0 (0)	370 859 (9.7)	161 (0.1)	303 (0.1)
Smoking status:						
Non-smoker	847 205 (48)	29 500 (47.1)	74 646 (41)	1 475 708 (39)	77 990 (37.7)	109 249 (32.7)
Former smoker	471 005 (27)	17 440 (27.8)	50 884 (28)	1 236 061 (33)	39 087 (18.9)	75 081 (22.4)
Current smoker	363 440 (21)	10 720 (17.1)	38 478 (21.2)	838 404 (22)	66 836 (32.3)	105 363 (31.5)
Missing	90 950 (5)	5031 (8.0)	17 905 (9.9)	255 193 (7)	23 043 (11.1)	44 859 (13.4)
Median (IQR) frailty index score	0.03 (0-0.08)	0.08 (0.06-0.14)	0.08 (0.06-0.14)	0.06 (0.03-0.08)	0.08 (0.06-0.17)	0.08 (0.06-0.17)
Alcohol intake status:						
Non-drinker	289 472 (16)	14 172 (22.6)	37 568 (20.7)	864 865 (23)	59 364 (28.7)	89 537 (26.8)
Occasional drinker	488 289 (28)	15 195 (24.2)	42 645 (23.5)	998 948 (26)	47 088 (22.8)	71 739 (21.4)
Light drinker	239 732 (14)	6472 (10.3)	18 863 (10.4)	696 369 (18)	26 635 (12.9)	44 924 (13.4)
Moderate drinker	179 102 (10)	3891 (6.2)	12 926 (7.1)	246 468 (7)	9378 (4.5)	17 491 (5.2)
Heavy drinker	22 760 (1.3)	891 (1.4)	2336 (1.3)	74 005 (2)	5124 (2.5)	6845 (2.1)
Unknown amount	291 649 (16)	9962 (15.9)	165 132 (14.4)	237 464 (6)	9631 (4.7)	12 117 (3.6)
Missing	216 596 (15)	12 108 (19.3)	41 261 (22.7)	687 247 (18)	49 736 (24)	91 899 (27.5)
Risk factors:						
Previous falls	108 745 (6)	10 514 (16.8)	22 459 (12.4)	140 886 (3.7)	21 697 (10.5)	25 124 (7.5)
Memory problems	28 276 (1.6)	3860 (6.2)	10 556 (5.8)	99 264 (2.6)	15 996 (7.7)	28 636 (8.6)
Mobility problems	20 425 (1.2)	2462 (3.9)	7347 (4.0)	85 675 (2.3)	13 999 (6.8)	22 928 (6.9)
Stroke	44 339 (2.5)	4320 (6.9)	14 167 (7.8)	111 462 (2.9)	15 704 (7.6)	26 703 (8
Multiple sclerosis	6367 (0.4)	300 (0.5)	798 (0.4)	11 328 (0.3)	975 (0.5)	1373 (0.4)
Antihypertensive drugs:						
ACE inhibitors	219 506 (12)	12 039 (19.2)	38 096 (20.9)	478 778 (13)	38 867 (18.8)	67 787 (20.3)
Angiotensin 2 receptor blockers	59 075 (3)	3167 (5.1)	7628 (4.2)	136 926 (4)	11 018 (5.3)	14 308 (4.3)
α blockers	34 338 (2)	2088 (3.3)	6794 (3.7)	68 131 (2)	6335 (3.1)	11 388 (3.4)
β blockers	216 122 (12)	10 885 (17.4)	31 341 (17.3)	461 329 (12)	36 317 (17.6)	59 019 (17.6)
Calcium channel blockers	193 141 (11)	11 570 (18.5)	35 859 (19.7)	426 151 (11)	37 590 (18.2)	63 764 (19.1)
Diuretics	180 065 (10)	10 706 (17.1)	29 783 (16.4)	397 980 (11)	36 418 (17.6)	55 934 (16.7)
Other antihypertensives	10 784 (0.6)	400 (0.8)	1594 (0.9)	19 235 (1)	1437 (0.7)	2471 (0.7)
Other drugs:						
Opioids	553 344 (31)	26 060 (41.6)	69 496 (38.2)	1 213 876 (32)	84 108 (40.6)	121 303 (36.3)
Hypnotics and anxiolytics	376 885 (21)	17 703 (28.2)	48 636 (26.8)	750 584 (20)	52 854 (25.5)	78 627 (23.5)
Antidepressants	383 647 (21)	17 159 (27.4)	42 767 (23.5)	793 690 (21)	52 820 (25.5)	71 452 (21.4)
Anticholinergics	207 345 (11)	11 085 (17.7)	29 384 (16.2)	388 513 (10)	31 542 (15.2)	46 255 (13.8)
Median (IQR) follow-up (years)	6.2 (2.6-10)	4.3 (1.8-7.0)	3.7 (1.6-6.3)	6.7 (2.7-10)	4.3 (1.9-7.1)	3.8 (1.6-6.5)

ACE=angiotensin converting enzyme; IQR=interquartile range; SD=standard deviation.

time point by fitting a generalised linear equation with a logit link function directly to the observed pseudo values in the development dataset. The linear predictor from the original model was the only variable included in the recalibration model, which allowed for a non-linear recalibration effect using fractional polynomials.

External validation

Researchers at Keele University (LA, KIES, RDR) conducted the external validation of the prediction model, independent of the model development team. The prediction model algorithms presented in figure 1 (both the original and the final) were applied to each individual in the external validation cohort to give the

Table 2 | Prediction model for falls. Values are subdistribution hazard ratios and 95% confidence intervals

Predictors	Full case analysis (n=358 207)	Multiple imputation model (n=1 772 600)
Age	30.1 (27.7 to 32.7)	60.46 (57.87 to 63.17)
Sex (women)	1.32 (1.28 to 1.35)	1.25 (1.23 to 1.27)
Total cholesterol	1.55 (1.44 to 1.67)	1.48 (1.36 to 1.61)
Ethnicity:		
White	Reference	Reference
Black	0.68 (0.59 to 0.79)	0.65 (0.58 to 0.74)
South Asian	0.67 (0.60 to 0.75)	0.68 (0.61 to 0.77)
Other	0.66 (0.59 to 0.74)	0.70 (0.63 to 0.78)
Index of multiple deprivation score:		
1	Reference	Reference
2	1.05 (1.00 to 1.09)	1.04 (1.01 to 1.07)
3	1.06 (1.02 to 1.12)	1.07 (1.05 to 1.10)
4	1.14 (1.01 to 1.19)	1.18 (1.15 to 1.21)
5	1.23 (1.18 to 1.29)	1.35 (1.31 to 1.39)
Smoking status:		
Non-smoker	Reference	Reference
Former smoker	1.06 (1.04 to 1.09)	1.12 (1.10 to 1.4)
Current smoker	1.26 (1.22 to 1.31)	1.27 (1.24 to 1.30)
Alcohol intake status:		
Non-drinker	Reference	Reference
Occasional drinker	0.87 (0.84 to 0.90)	0.90 (0.85 to 0.95)
Light drinker	0.93 (0.89 to 0.98)	0.94 (0.88 to 1.00)
Moderate drinker	0.99 (0.94 to 1.05)	0.99 (0.93 to 1.06)
Heavy drinker	1.71 (1.55 to 1.87)	1.57 (1.28 to 1.93)
Unknown amount	0.97 (0.95 to 1.02)	0.93 (0.89 to 0.98)
Frailty index score	1.11 (1.09 to 1.14)	1.22 (1.20 to 1.23)
Risk factors:		
History of falls	1.40 1.35 to 1.46)	1.32 (1.29 to 1.35)
Memory problems	1.25 (1.17 to 1.35)	1.17 (1.12 to 1.21)
Mobility problems	0.99 (0.93 to 1.07)	0.92 (0.87 to 0.98)
Stroke	1.28 (1.22 to 1.34)	1.14 (1.11 to 1.18)
Multiple sclerosis	1.48 (1.23 to 1.78)	1.71 (1.51 to 1.94)
Antihypertensive drugs:		
ACE inhibitors	1.04 (1.01 to 1.07)	1.12 (1.10 to 1.14)
Angiotensin 2 receptor blockers	1.07 (1.02 to 1.12)	1.19 (1.15 to 1.23)
α blockers	1.00 (0.95 to 1.06)	1.04 (1.02 to 1.06)
β blockers	0.97 (0.96 to 1.00)	1.07 (1.02 to 1.12)
Calcium channel blockers	0.99 (0.97 to 1.03)	1.08 (1.06 to 1.11)
Diuretics	0.98 (0.95 to 1.01)	1.07 (1.05 to 1.10)
Other antihypertensives	1.08 (0.97 to 1.21)	0.96 (0.88 to 1.04)
Other drugs		
Opioids	1.10 (1.07 to 1.13)	1.11 (1.08 o 1.13)
Hypnotics and anxiolytics	1.04 (1.00 to 1.07)	1.15 (1.13 to 1.18)
Antidepressants	1.14 (1.10 to 1.18)	1.16 (1.13 to 1.18)
Anticholinergics	1.11 (1.06 to 1.14)	1.03 (1.02 to 1.05)

ACE=angiotensin converting enzyme.
Variable transformations: Age=((age/100)^3)−0.242; cholesterol=((cholesterol/10)^−0.5)−1.381; frailty index=(frailty index/0.1)−0.576.

predicted probabilities of experiencing a fall within one, five, and 10 years, taking account of the competing risk of death by other causes.[42] Model calibration was assessed through comparison of predicted probabilities to observed pseudo values, estimated using jack-knife estimators representing an individual's contribution to the cumulative incidence function for falls, accounting for competing risks, calculated by the Aalen–Johansen method in the external validation cohort.

Predictive performance was quantified by calculating the observed to expected ratio, Harrell's C statistic, Royston's D statistic with its associated R^2 statistic,[43] each applied to the same pseudo values as above, and by using calibration plots and curves. Calibration plots were generated separately in each imputed dataset and checked for consistency (one illustrative example is shown for each model). All measures were calculated in each imputed dataset separately and, when appropriate, combined across imputations using Rubin's rules. When Rubin's rules did not apply (eg, when the posterior distribution was not expected to be normal), performance was summarised across imputations using the median and interquartile range.[44]

Heterogeneity in model performance across different general practices was assessed using a random effects meta-analysis, using restricted maximum likelihood estimation, given that the case mix and incidence of falls were expected to vary between practices (see extended methods in the supplementary material).[45] The observed to expected ratio was pooled across practices on the natural log scale, the C statistic on the logit scale (with the standard errors of logit C calculated using the delta method), and the D statistic on its original scale.[46 47] Pooled estimates are reported with prediction intervals to give an indication of expected model performance in a new general practice.

Clinical utility was assessed by plotting the one year, five year, and 10 year risk of falls against the 10 year risk of cardiovascular disease, calculated using the Qrisk2 algorithm.[22] Clinical utility was also examined using net benefit analysis, where the harms and benefits of using a model to guide treatment decisions were offset to assess the overall consequences of using the STRATIFY-Falls prediction models for clinical decision making.[48] The original and final models were compared with one another at five and 10 years and with model blind methods of introducing falls prevention measures (which may include deprescribing) for all patients, or not introducing falls prevention measures (starting or continuing treatment) for all patients, regardless of falls risk. We assessed net benefit across the full range of possible threshold probabilities, with a falls risk above 10% at 10 years specified a priori as being a threshold of clinical interest, to align with current thresholds for an individual's risk of cardiovascular disease.[49]

The same external validation methods as described earlier were employed in subgroups by age (<65 years, ≥65 years), sex (women, men), and ethnicity (white, black, South Asian, other), to assess the models' predictive performance in these clinically relevant groups.

Patient and public involvement

This study was developed and conducted with the help of our patient and public advisor Margaret Ogden. As a member of our study advisory group, they commented on the study protocol and have been present in all team meetings discussing results and reporting. We also held a focus group with several older adults during the study to discuss broader themes related to drugs for cardiovascular disease prevention and adverse events, which informed the interpretation of this work.

doi: 10.1136/bmj-2022-070918 | BMJ 2022;379:e070918 | the bmj

Table 3 | Predictive performance statistics of the falls prediction models on external validation in Clinical Practice Research Datalink Aurum

	1 year	5 years		10 years	
Statistics	Original model	Original model	Pseudo value recalibration	Original model	Pseudo value recalibration
Observed to expected ratio					
Pooled effect size (95% CI)	0.162 (0.158 to 0.166)	1.702 (1.674 to 1.730)	1.906 (1.874 to 1.939)	1.682 (1.657 to 1.707)	1.839 (1.811 to 1.865)
Prediction interval	0.090 to 0.289	1.116 to 2.586	1.246 to 2.915	1.139 to 2.484	1.284 to 2.638
τ^2	0.089 (0.080 to 0.099)	0.046 (0.042 to 0.052)	0.0479 (0.043 to 0.054)	0.038 (0.035 to 0.043)	0.0342 (0.031 to 0.038)
C statistic					
Pooled effect size (95% CI)	0.866 (0.862 to 0.869)	0.843 (0.841 to 0.844)	0.843 (0.841 to 0.844)	0.833 (0.832 to 0.835)	0.833 (0.831 to 0.835)
Prediction interval	0.794 to 0.915	0.789 to 0.881	0.789 to 0.881	0.789 to 0.870	0.789 to 0.870
τ^2	0.068 (0.056 to 0.083)	0.026 (0.023 to 0.030)	0.026 (0.023 to 0.030)	0.022 (0.019 to 0.025)	0.022 (0.019 to 0.025)
D statistic					
Pooled effect size (95% CI)	2.160 (1.987 to 2.333)	1.903 (1.754 to 2.051)	1.894 (1.746 to 2.042)	1.643 (1.515 to 1.771)	1.597 (1.472 to 1.721)
Prediction interval	1.99 to 2.33	1.75 to 2.05	1.75 to 2.04	1.51 to 1.77	1.47 to 1.72
τ^2	0.000 (0.000 to 0.039)	0.000 (0.000 to 0.023)	0.000 (0.000 to 0.022)	0.000 (0.0000 to 0.0168)	0.000 (0.000 to 0.016)
Royston and Sauerbrei's R^2					
Range	0 to 86.0	28.0 to 91.4	25.9 to 91.4	21.3 to 91.4	21.6 to 91.4
Median (IQR)	58.1 (52.3 to 62.2)	47.4 (43.5 to 51.8)	47.3 (43.2 to 51.7)	39.9 (36.4 to 43.8)	38.6 (35.4 to 42.4)
Mean (SD)	56.5 (0.10)	47.9 (0.07)	47.7 (0.07)	40.8 (0.07)	39.4 (0.07)

CI=confidence interval; IQR=interquartile range; SD=standard deviation.

Results

Study population characteristics

Figure 2 shows the flow of study participants for both the development cohort and the validation cohort. A total of 1 772 600 patients were included in the model development cohort (CPRD GOLD), with a mean age of 59 years (standard deviation (SD) 13 years) and a mean systolic blood pressure of 144 mm Hg (SD 12 mm Hg) at study inclusion (table 1). The 10 year prevalence of falls was 3.5% (n=62 691), with 10.3% of patients (n=181 731) experiencing death by other causes before any fall occurred, and a median follow-up of 6.2 years (interquartile range (IQR) 2.6-10 years) across the cohort.

In total, 3 805 366 patients were included in the validation cohort, with 206 956 (5.4%) experiencing fall events during 10 year follow-up. A further 334 552 (8.8%) patients died during follow-up from unrelated causes, before any fall occurred. Median follow-up time in the validation cohort was 6.7 years (IQR 2.7-10 years). Total cholesterol level was missing in 48% of participants, and ethnicity data were more complete in the validation cohort than development cohort (81% *v* 44% complete data).

Model development

The original model consisted of 24 predictors, after the exclusion of variables with little or no association in multivariable analysis (table 2). Compared with men, women were more likely to experience a fall during follow-up (subdistribution hazard ratio 1.25, 95% confidence interval 1.23 to 1.27). Increasing age, white ethnicity, and being a smoker, a heavy drinker, or more deprived were predictors associated with an increased risk of falls (table 2). Increasing frailty was one of the strongest predictors of falls, with an increased falls risk of 22% for about every four deficits accrued (1.22, 1.20 to 1.23). Of the previous medical conditions examined, the strongest predictors of falls were having a history of falls (1.32, 1.29 to 1.35) and multiple sclerosis (1.71, 1.51 to 1.94). Drugs most strongly associated with falls were angiotensin 2 receptor blockers (1.19, 1.15 to 1.23), antidepressants (1.16, 1.13 to 1.18), hypnotics and anxiolytics (1.15, 1.13 to 1.18), angiotensin converting enzyme inhibitors (1.12, 1.10 to 1.14), and opioids (1.11, 1.08 to 1.13). To ensure a parsimonious final model, systolic and diastolic blood pressure, BMI, activity limitation, syncope, and cataract were excluded from the model owing to a lack of association with falls risk. No violations of the proportional hazards assumption were detected.

Internal validation and recalibration using pseudo values

At five and 10 years, apparent calibration plots in the model development data showed significant miscalibration, with under-prediction for patients with a low predicted risk and substantial over-prediction for those with a high predicted risk (see supplementary figure S3.1). We therefore recalibrated the original model to the observed pseudo values and this improved apparent calibration (in the model development data) considerably (fig 4 and fig 5). Apparent calibration of the original model at one year was good, therefore recalibration was not required (see fig 3).

External validation

Predictive performance

Upon external validation, the original model showed excellent discrimination (table 3) but poor calibration (see supplementary figure S3.1), with considerable heterogeneity across general practices (see supplementary figure S3.2). Recalibration of the model corrected miscalibration in the model development cohort, but under-prediction of risk was still present in the validation cohort (fig 3, fig 4, and fig 5). This miscalibration was less extreme than that of the original model, in the narrower range of predicted probabilities between 0 to 0.2. On average, the recalibrated model showed a pooled observed to expected ratio at 10 years

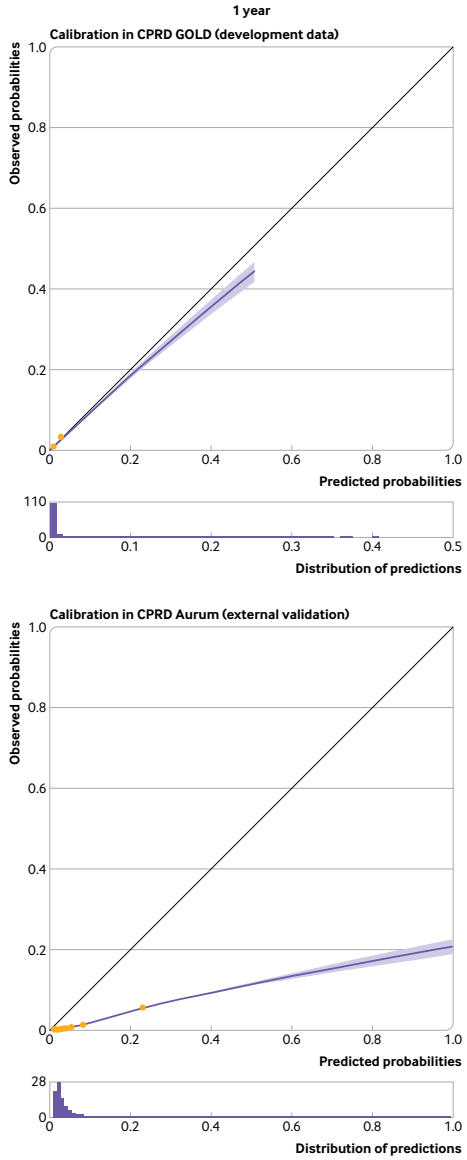


Fig 3 | Calibration curves for apparent performance of the final STRATIFY-Falls model in CPRD GOLD at one year, and calibration on external validation in CPRD Aurum at one year. Groups represent 10ths of linear predictor, as created between deciles. Histogram shows distribution of predicted probabilities. The model is not recalibrated to pseudo values in the development data. CPRD=Clinical Practice Research Datalink; STRATIFY=STRAtifying Treatments In the multi-morbid Frail elderly

of 1.839 (95% confidence interval 1.811 to 1.865, 95% prediction interval 1.284 to 2.638), suggesting that the observed incidence of falls would be around 84% (relatively) higher than expected when using the model to generate predictions. Under-prediction of 10 year falls risk was consistent across all subgroups, with the exception of the "other ethnicity" group, where both the falls incidence and the observed to expected ratio were considerably lower than in the full population (see extended results in supplementary material section 2.2).

The ordering of participants' predicted probabilities altered only slightly on recalibration; thus discriminative ability of the recalibrated models remained excellent at each of the analysis time points, with C statistics of 0.843 (95% confidence interval 0.841 to 0.844, 95% prediction interval 0.789 to 0.881) at five years, and 0.833 (0.831 to 0.835, 95% prediction interval 0.789 to 0.870) at 10 years, and D statistic values of 1.894 (1.746 to 2.042, 95% prediction interval 1.75 to 2.04) at five years, and 1.597 (1.472 to 1.721, 95% prediction interval 1.47 to 1.72) at 10 years (table 3). Model performance varied more among smaller practices, with more consistent performance seen as practice size increased (fig 6).

The model's discriminative ability at 10 years was consistent across age and sex subgroups (see supplementary tables S2.1 and S2.2). The pooled C statistic was lowest in those of white ethnicity (0.796, 95% confidence interval 0.793 to 0.798) and highest among those of other ethnicity (0.834, 0.830 to 0.839) (see supplementary table S2.3).

Clinical utility analysis

Net benefit and decision curve analysis of the original and recalibrated models indicated potential clinical utility at five and 10 years around the predefined threshold of 10% (fig 7). At 10 years, basing clinical management decisions on predicted probabilities of falls yielded a benefit over the two strategies of introducing falls prevention measures (which may include deprescribing) for all and not introducing falls prevention measures (starting or continuing treatment) for all patients, when using a treatment decision threshold of 7% or higher from the original model, or a treatment decision threshold of 6% or higher from the final recalibrated model. Thus, for either model, when using our prespecified treatment decision cut-off of 10% risk of falls at 10 years, we would expect a benefit to patients over and above model blind treatment strategies (usual care). This treatment decision threshold of 10% showed a net benefit in all subgroups except other ethnicity, where a cut-off of at most 3% was required for the model to be superior to usual care for all (see supplementary figure S2.6). In the analysis at five years, using a treatment decision threshold of 3% risk or higher gave a net benefit above starting or continuing treatment for all, for both models.

In analyses comparing the risk of falls with the risk of cardiovascular disease in CPRD GOLD, 198 654 (11%) patients had a high risk of falls (>10%) but low

doi: 10.1136/bmj-2022-070918 | *BMJ* 2022;379:e070918 | the bmj

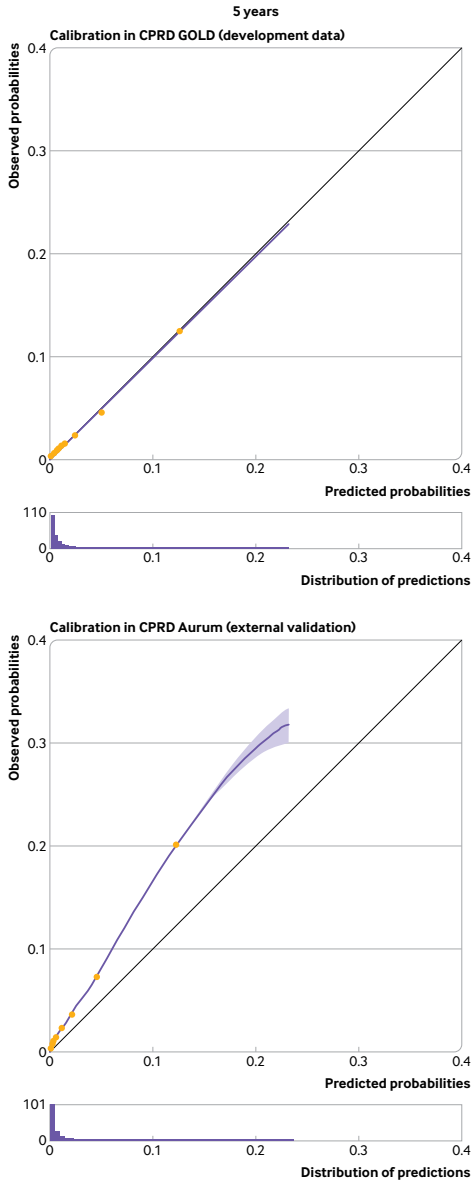


Fig 4 | Calibration curves for apparent performance of the final STRATIFY-Falls model in CPRD GOLD at five years, and calibration on external validation in CPRD Aurum at five years. Groups represent 10ths of linear predictor, as created between deciles. Histogram shows distribution of predicted probabilities. CPRD=Clinical Practice Research Datalink; STRATIFY=STRAtifying Treatments In the multi-morbid Frail elderlY

risk of cardiovascular disease (<10%) at 10 years (fig 8). A further 128 458 (7%) patients were classified as high risk of both, and 571 274 (32%) had a low falls risk but high risk of cardiovascular disease.

Discussion

Principal findings

We developed and externally validated a clinical prediction model to determine an individual's risk of experiencing a fall resulting in hospital admission or death within 10 years of being indicated for antihypertensive treatment (owing to raised blood pressure readings). The model incorporates routinely recorded information, including a history of previous falls, multiple sclerosis, heavy alcohol consumption, high deprivation score, and prescribed drugs, which were all strong predictors of subsequent falls, conditional on the other model variables.

The final recalibrated model showed good discrimination upon external validation, suggesting that it can help distinguish those at a higher risk of falling, which may improve how doctors identify patients who might benefit from targeted fall prevention strategies, including multifactorial or exercise based interventions,[50] and drug reviews including deprescribing. Calibration performance of the prediction model was inconsistent across the development and validation datasets, with miscalibration leading to under-prediction of fall risk across the full range of predicted probabilities. Nevertheless, such under-prediction of risk may be deemed acceptable if the model is intended to inform whether treatment should be stopped to avoid adverse effects—particularly if the treatment in question also carries benefits. Indeed, the clinical utility analysis showed that at risk thresholds around 10%, the net benefit of the model is higher than for other strategies currently employed in usual care.

Strengths and limitations of this study

Strengths of this work include the large, population based cohorts used, incorporating routinely collected patient data that have been shown to be representative of the patients across England, suggesting that the findings could be generalised across this (or a similar) population.[20 21] Analyses accounted for the competing risk of death in both model development and external validation, ensuring that falls risk was not over-estimated. This is particularly important in individuals with frailty and multiple long term conditions, where an over-estimation of falls risk might preclude prescription of antihypertensive drugs in those who could still derive benefit from continued treatment. This analysis method is superior to most prediction models in widespread use, which do not take into account competing risks.[22] In these models, the stated risk of an event (cardiovascular disease, for example) is by design too high, as the actual risk of an event would be diminished by death from other (eg, non-cardiovascular) causes, particularly in older people.[35]

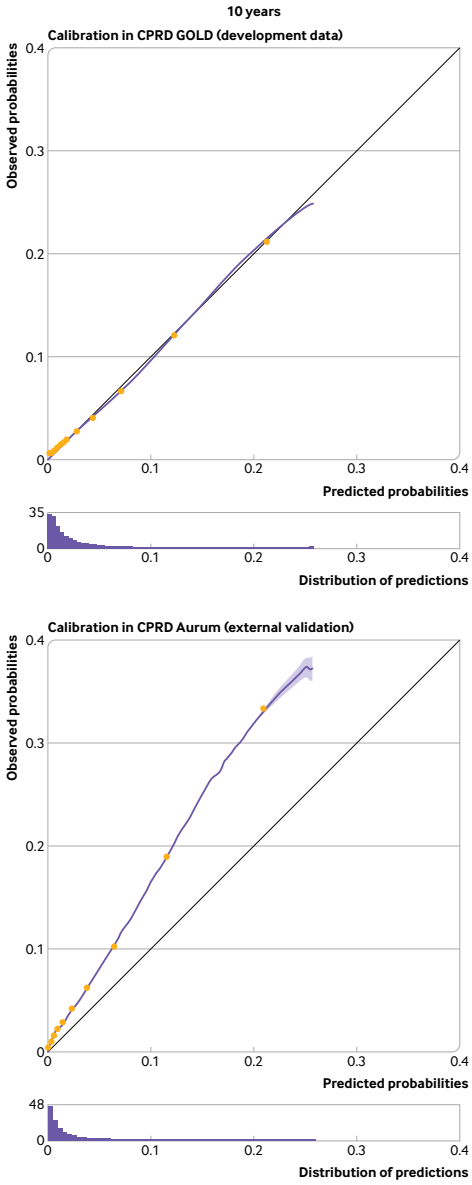


Fig 5 | Calibration curves for apparent performance of the final STRATIFY-Falls model in CPRD GOLD at 10 years, and calibration on external validation in CPRD Aurum at 10 years. Groups represent 10ths of linear predictor, as created between deciles. Histogram shows distribution of predicted probabilities. CPRD=Clinical Practice Research Datalink; STRATIFY=STRAtifying Treatments In the multi-morbid Frail elderlY

All data were derived from routine electronic health records, including the outcome definition of falls. Such a definition might not capture all events that could be included in the ProFaNE (Prevention of Falls Network Europe) consensus definition of a fall (ie, an unexpected event in which the participants come to rest on the ground, floor, or lower level),[51] and therefore the model results should be interpreted in this context. It is possible that some of these fall events were not reported or captured correctly within the electronic health record, therefore potentially underestimating the incidence of falls, which could have affected the performance of the model.

Assessments of the models' predictive performance were conducted across a range of general practices, with different case mix and outcome prevalence, giving an indication of the expected spread of performance across a range of subpopulations. Model performance varied more among smaller practices, with more consistent performance seen as practice size increased. This reflects the increased uncertainty in the estimation of the predictive performance measures in practices of low sample sizes, many of which individually would have failed to meet the required sample size for this external validation. Prediction intervals from meta-analyses across general practices give an indication of how well our falls models would be expected to perform in new practices, helping to inform decisions on implementation in practice. In the present study, the prediction intervals were relatively narrow across a range of performance statistics, suggesting that the models would perform similarly in a new practice from a similar population.

All variables included in our model were predetermined based on the literature, although we did choose to exclude some variables at the model development stage that had exhibited a negligible effect on the outcome. These variables were excluded because they did not contribute substantially to model predictions and served to unnecessarily increase the complexity of the equation. We did not use statistical selection methods such as backwards or forwards elimination, as these can lead to overfitting. Although our approach may have meant that some statistically significant (but clinically insignificant) predictors were excluded from the final model, these exclusions are unlikely to have led to overfitting given the large sample size or been the reason for miscalibration in the external validation.

For these models, we defined binary variables for antihypertensive drugs as any prescription within the year before (and including) the index date, without accounting for any changes to drugs during follow-up. Not allowing for the time varying nature of treatment could potentially affect the observed associations with falls risk, and so too the predicted risks obtained from the model. However, our model is intended to give a prediction for risk of falls over the next 1-10 years, from a particular moment in time, in the context of current care. The latter is important, because, for example, if a patient has low risk, then it means that

doi: 10.1136/bmj-2022-070918 | *BMJ* 2022;379:e070918 | the bmj

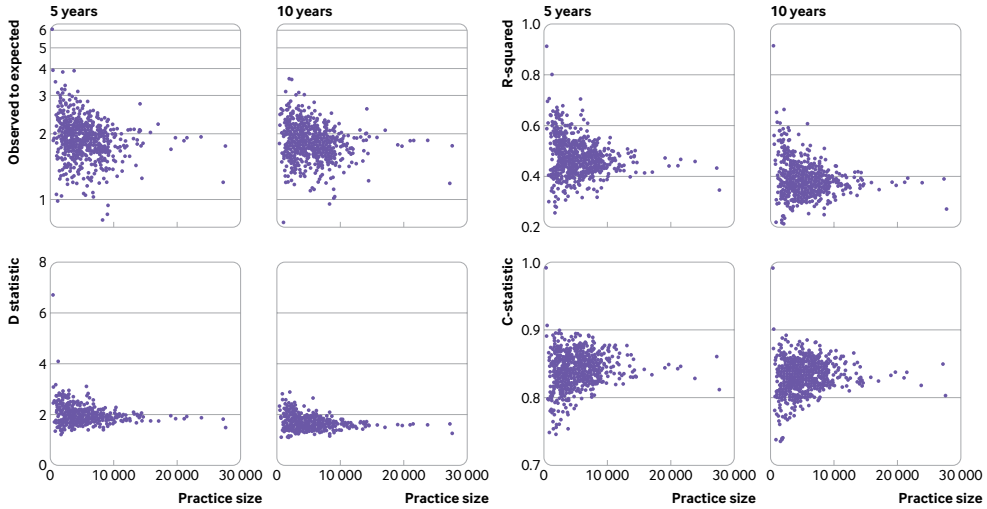


Fig 6 | Performance variability of the final STRATIFY-FALLS model on external validation across general practices, with observed to expected ratio, R^2 statistic, D statistic, and C statistic. STRATIFY=STRAtifying Treatments In the multi-morbid Frail elderlY

current care (ie, treatments and monitoring strategies over the next 1-10 years) is likely to be adequate for this individual. In contrast, if an individual's risk is high, it means that current care is likely insufficient and that additional or alternative approaches are potentially needed.

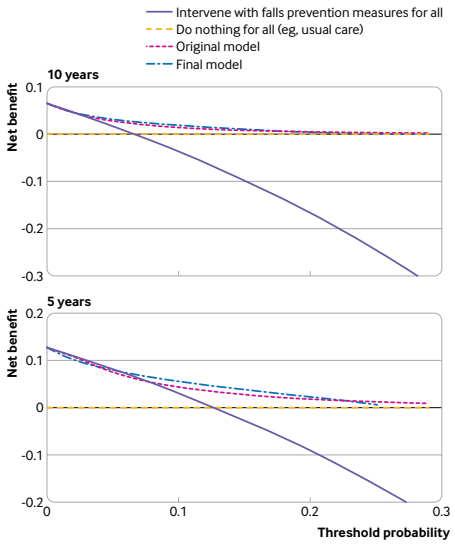


Fig 7 | Decision curve analysis showing net benefit of using prediction models across different threshold probabilities for assigning treatment

Calibration performance of the prediction model was inconsistent across the model development and validation datasets. Such miscalibration was surprising, as populations were similar across both datasets for predictor distributions and the incidence of falls and of death (with the exception of self-reported characteristics such as smoking status, alcohol consumption, and ethnicity, which may reflect differences in how these data are captured within the electronic health record systems that underlie these databases). Distributions of the linear predictor were also consistent across the development and validation datasets, suggesting miscalibration could be due to differences in the outcomes or the outcome recording or coding. This is representative of real life, where outcome definitions vary, and both models still exhibited useful discrimination and potential clinical utility across the full population for a range of treatment decision threshold probabilities, although the predicted risk for individuals may be different (miscalibrated) from their actual risk. Indeed, miscalibration was most evident in the 5-10% of patients with the highest predicted risk (those above a threshold of 10%), and in these patients, doctors may interpret the exact predicted risks with caution, even though these patients can still be considered at higher risk.

Comparison with previous literature

Several prediction models can now estimate an individual's risk of falls, including those for use in the

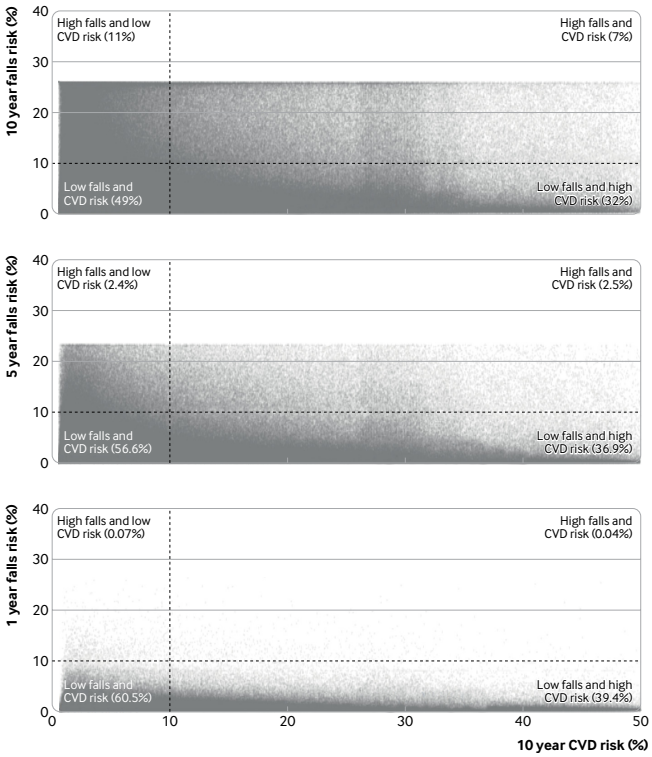


Fig 8 | Comparison of 10 year cardiovascular disease risk (Qrisk2) and fall risk in Clinical Practice Research Datalink GOLD dataset. High risk for both conditions was defined as a risk >10%. CVD=cardiovascular disease

community. A recent systematic review of development and validation studies identified a total of 72 existing models.[10] These were typically poorly reported, with only 40 studies (56%) reporting discrimination statistics and seven studies (10%) reporting calibration. Only three models were externally validated. Discrimination was reported with area under the curves of 0.49 to 0.87 for internally validated models and 0.62 to 0.69 for externally validated models. Calibration was moderately good but presented in 10ths of risk across a small range of risk thresholds (eg, 0-10% [52]) making it difficult to determine how calibration varied across the full range of predicted probabilities. All studies were deemed at high risk of bias owing to methods of analysis and outcome assessment along with restrictive eligibility criteria.

In contrast, our final model, reported in line with the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) guidelines for reporting of clinical prediction models[53] (see supplementary table S4.3), showed excellent discrimination upon external validation, with an area under the curve of 0.84. It demonstrated reasonable calibration across the low range of predicted risks typically examined by previous risk models (eg, 0-10%) and although miscalibration was present at higher predicted probabilities, there was still clinical utility based on the decision curve analysis. This suggests that the present model is the most promising clinical prediction model for falls available to date, and that it may be effective in identifying individuals at high risk of falls from those in primary care with raised blood pressure.

Implications for policy and practice

As patients age, their risk of a fall resulting in serious injury and long term disability increases.[4] Identifying those most at risk is therefore important to enable targeting of fall prevention strategies.[7] The present model provides primary care doctors with a method of estimating the risk of falls using data routinely available in electronic health records and could

doi: 10.1136/bmj-2022-070918 | *BMJ* 2022;379:e070918 | the bmj

have uses beyond predicting falls in patients being considered for antihypertensive treatment.[54]

Among patients aged 40 years and older, with an indication for antihypertensive drugs owing to raised blood pressure, the model was shown to distinguish well those at high risk of falls in the next 1-10 years. Miscalibration was noted, with an under-prediction of risk seen particularly at higher predicted probabilities. Depending on how the model might be used, such under-prediction might be less of a concern—for example, if the model was being used to inform treatment changes only above a certain threshold of predicted risk. In this context, doctors could be confident that the true risk is at least at this threshold, if not higher. Further studies are, however, needed to explore the appropriate thresholds that maximise the model's clinical utility and cost effectiveness, and to examine whether recalibration is possible in local settings.

The model may also be used to target falls prevention strategies to patients with the highest risk. These strategies might include multifactorial or exercise based interventions,[50] or review of prescribed drugs, with those drugs likely to increase the risk of falls being considered for deprescribing.[418] Such drug reviews are increasingly being encouraged in routine clinical practice, and the STRATIFY-Falls model may be useful for informing these reviews.[55] For example, in patients prescribed antihypertensive treatment, the model might be used alongside a cardiovascular risk prediction algorithm to compare the potential for benefit and harm from continued treatment prescription.[262756] For individuals with a high risk of falls but low risk of cardiovascular disease, a doctor might consider whether new or continued antihypertensive treatment is still appropriate. We examined the prevalence of this scenario in our model development population (fig 8) and identified an important number of individuals (11%) who would be classified in this way, when comparing risks at 10 years. More common, however, were individuals with a low risk of falls but high risk of cardiovascular disease (affecting one in three patients). For these patients, doctors could use the model to illustrate the minimal risk of harm for individuals, potentially improving uptake of, adherence to, and persistence with antihypertensive treatment, which is known to be poor currently.[57]

Conclusions

The STRATIFY-Falls prediction model helps to identify those at high risk of falls and could be used by doctors wanting to identify patients who might benefit from targeted fall prevention strategies, including multifactorial or exercise based interventions[50] and drug reviews. Used alongside other prediction tools such as those for cardiovascular risk, such a model could be valuable when used as part of a wider risk assessment for falls prevention.

The STRAtifying Treatments In the multi-morbid Frail elderlY (STRATIFY) investigators include the authors already listed and: John Gladman, professor of medicine of older people, School of Medicine, University of Nottingham; Simon Griffin, professor of primary care, Department of Public Health and Primary Care, Primary Care Unit, University of Cambridge; and Margaret Ogden, patient and public involvement advisor.

We thank Lucy Curtin for administrative support throughout the project and Margaret Ogden, Simon Griffin, and John Gladman for their contributions as STRATIFY Investigators to the project. The Hospital Episode Statistics data used in this analysis are reused with permission of NHS Digital, which retains the copyright for those data. We thank the Office for National Statistics for providing data on mortality. The ONS and NHS Digital bear no responsibility for the analysis or interpretation of the data. Finally, we are grateful to all those patients who permitted their anonymised routine NHS data to be used for this research.

Contributors: JPS conceived the project and wrote the protocol with FDRH, RJM, RS, and RDR. CK and SLF extracted data for analysis. CK developed the model under supervision of JS and RS. LA validated the model under supervision of RDR and KIES. LA, KIES, and RDR wrote the first draft of the manuscript. All authors revised the manuscript and approved the final version. JPS is the guarantor for this work and accepts full responsibility for the conduct of the study, had access to the data, and controlled the decision to publish. The corresponding author (JPS) attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Funding: JPS and CK were funded in whole, or in part, by the Wellcome Trust and Royal Society via a Sir Henry Dale fellowship held by JPS (ref: 211182/Z/18/Z) and the National Institute for Health and Care Research (NIHR) School for Primary Care (project 430) awarded to JPS. JPS also receives funding via an NIHR Oxford Biomedical Research Centre (BRC) senior fellowship. RJMcM is supported by an NIHR senior investigator award. FDRH acknowledges part support from the NIHR ARC Oxford Thames Valley and the NIHR Oxford University Hospitals BRC. KIES is funded by an NIHR School for Primary Care Research (SPCR) launching fellowship. SLF was part funded by the NIHR BRC and NIHR Applied Research Collaboration (ARC) Oxford and Thames Valley. AB has received research funding from AstraZeneca, NIHR, BMA Medical Research Foundation, and UK Research and Innovation. RAP receives funding from the NIHR. AC is part funded by NIHR ARC Yorkshire and Humber and Health Data Research UK, an initiative funded by UK Research and Innovation Councils, NIHR and the UK devolved administrations, and leading medical research charities. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. The sponsor and funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: authors had financial support from the Wellcome Trust, Royal Society, and National Institute for Health and Care Research for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Ethical approval: The study protocol was approved by the Clinical Practice Research Datalink (CPRD) independent scientific advisory committee in February 2019 before obtaining the data relevant to the project (protocol given in the eAppendix in the supplementary material). As all data are fully anonymised, no consent was required. A project summary is published on the CPRD website (https://www.cprd.com/isac).

Data sharing: Data were obtained via a Clinical Practice Research Datalink (CPRD) institutional licence. Requests for data sharing should be made directly to the CPRD. The algorithm is freely available for research use and can be downloaded from https://process.innovation.ox.ac.uk/software/. Code lists used to define variables included in the dataset are available at https://github.com/jamessheppard48/STRATIFY-BP/tree/STRATIFY-Falls

The manuscript's guarantor (JPS) affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as originally planned (and, if relevant, registered) have been explained.

Dissemination to participants and related patient and public communities: Findings from this study will be press released alongside publication of this manuscript. Social media (eg, Twitter)

will be used to draw attention to the work and stimulate debate about its findings. We will also publish a lay summary of our findings on our study website https://www.phc.ox.ac.uk/research/stratified-treatments/studies/stratifying-treatments-in-the-multi-morbid-frail-elderly-stratify-antihypertensives and make the underlying developed algorithms freely available for academic use here: https://process.innovation.ox.ac.uk/software/.

Provenance and peer review: Not commissioned; externally peer reviewed.

1 Office for National Statistics. Living longer: how our population is changing and why it matters 2018. www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/ageing.
2 Tinetti ME, Speechley M, Ginter SF. Risk factors for falls among elderly persons living in the community. *N Engl J Med* 1988;319:1701-7. doi:10.1056/NEJM198812293192604
3 Gale CR, Cooper C, Aihie Sayer A. Prevalence and risk factors for falls in older men and women: The English Longitudinal Study of Ageing. *Age Ageing* 2016;45:789-94. doi:10.1093/ageing/afw129
4 Gill TM, Murphy TE, Gahbauer EA, Allore HG. Association of injurious falls with disability outcomes and nursing home admissions in community-living older persons. *Am J Epidemiol* 2013;178:418-25. doi:10.1093/aje/kws554
5 Office for Health Improvement & Disparities. Public Health Outcomes Framework 2019-2020. 2021. https://fingertips.phe.org.uk/profile/public-health-outcomes-framework/data
6 Whelton PK, Carey RM, Aronow WS, et al. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Hypertension* 2018;71:e13-115.
7 National Clinical Guideline Centre. Falls in older people: Assessing risk and prevention. [NICE guideline CG161.] NICE, 2013. https://www.nice.org.uk/guidance/cg161
8 White AM, Tooth LR, Peeters GMEEG. Fall risk factors in mid-age women: the Australian Longitudinal Study on Women's Health. *Am J Prev Med* 2018;54:51-63. doi:10.1016/j.amepre.2017.10.009
9 Robbins AS, Rubenstein LZ, Josephson KR, Schulman BL, Osterweil D, Fine G. Predictors of falls among elderly people. Results of two population-based studies. *Arch Intern Med* 1989;149:1628-33. doi:10.1001/archinte.1989.00390070138022
10 Gade GV, Jørgensen MG, Ryg J, et al. Predicting falls in community-dwelling older adults: a systematic review of prognostic models. *BMJ Open* 2021;11:e044170. doi:10.1136/bmjopen-2020-044170
11 de Vries M, Seppala LJ, Daams JG, van de Glind EMM, Masud T, van der Velde N, EUGMS Task and Finish Group on Fall-Risk-Increasing Drugs. Fall-Risk-Increasing Drugs: A Systematic Review and Meta-Analysis: I. Cardiovascular Drugs. *J Am Med Dir Assoc* 2018;19:371.e1-9. doi:10.1016/j.jamda.2017.12.013
12 Seppala LJ, Wermelink AMAT, de Vries M, et al, EUGMS task and Finish group on fall-risk-increasing drugs. Fall-Risk-Increasing Drugs: A Systematic Review and Meta-Analysis: II. Psychotropics. *J Am Med Dir Assoc* 2018;19:371.e11-7. doi:10.1016/j.jamda.2017.12.098
13 Seppala LJ, van de Glind EMM, Daams JG, et al, EUGMS Task and Finish Group on Fall-Risk-Increasing Drugs. Fall-Risk-Increasing Drugs: A Systematic Review and Meta-analysis: III. Others. *J Am Med Dir Assoc* 2018;19:372.e1-8. doi:10.1016/j.jamda.2017.12.099
14 Ettehad D, Emdin CA, Kiran A, et al. Blood pressure lowering for prevention of cardiovascular disease and death: a systematic review and meta-analysis. *Lancet* 2016;387:957-67. doi:10.1016/S0140-6736(15)01225-8
15 Albasri A, Hattle M, Koshiaris C, et al, STRATIFY investigators. Association between antihypertensive treatment and adverse events: systematic review and meta-analysis. *BMJ* 2021;372:n189. doi:10.1136/bmj.n189
16 Leipzig RM, Cumming RG, Tinetti ME. Drugs and falls in older people: a systematic review and meta-analysis: II. Cardiac and analgesic drugs. *J Am Geriatr Soc* 1999;47:40-50. doi:10.1111/j.1532-5415.1999.tb01899.x
17 Tinetti ME, Han L, Lee DS, et al. Antihypertensive medications and serious fall injuries in a nationally representative sample of older adults. *JAMA Intern Med* 2014;174:588-95. doi:10.1001/jamainternmed.2013.14764
18 Reeve E, Gnjidic D, Long J, Hilmer S. A systematic review of the emerging definition of 'deprescribing' with network analysis: implications for future research and clinical practice. *Br J Clin Pharmacol* 2015;80:1254-68. doi:10.1111/bcp.12732
19 Smith MI, de Lusignan S, Mullett D, Correa A, Tickner J, Jones S. Predicting Falls and When to Intervene in Older People: A Multilevel Logistical Regression Model and Cost Analysis. *PLoS One* 2016;11:e0159365. doi:10.1371/journal.pone.0159365
20 Herrett E, Gallagher AM, Bhaskaran K, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol* 2015;44:827-36. doi:10.1093/ije/dyv098
21 Wolf A, Dedman D, Campbell J, et al. Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum. *Int J Epidemiol* 2019;48:1740-1740g.
22 Hippisley-Cox J, Coupland C, Vinogradova Y, et al. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ* 2008;336:1475-82. doi:10.1136/bmj.39609.449676.25
23 Wehner-Hewson N, Watts P, Buscombe R, Bourne N, Hewson D. Racial and Ethnic Differences in Falls Among Older Adults: a Systematic Review and Meta-analysis. *J Racial Ethn Health Disparities* 2021. doi:10.1007/s40615-021-01179-1
24 National Clinical Guideline Centre. 2019 surveillance of falls in older people: assessing risk and prevention. [NICE guideline CG161.] Appendix A: Summary of evidence review. NICE, 2019. https://www.nice.org.uk/guidance/cg161/resources/2019-surveillance-of-falls-in-older-people-assessing-risk-and-prevention-nice-guideline-cg161-pdf-8792148103909
25 Clegg A, Bates C, Young J, et al. Development and validation of an electronic frailty index using routine primary care electronic health record data. *Age Ageing* 2016;45:353-60. doi:10.1093/ageing/afw039
26 Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ* 2017;357:j2099. doi:10.1136/bmj.j2099
27 D'Agostino RBSr, Vasan RS, Pencina MJ, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation* 2008;117:743-53. doi:10.1161/CIRCULATIONAHA.107.699579
28 van Geloven N, Swanson SA, Ramspek CL, et al. Prediction meets causal inference: the role of treatment in clinical prediction models. *Eur J Epidemiol* 2020;35:619-30. doi:10.1007/s10654-020-00636-1
29 Riley RD, Snell KI, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Stat Med* 2019;38:1276-96. doi:10.1002/sim.7992
30 Riley RD, Collins GS, Ensor J, et al. Minimum sample size calculations for external validation of a clinical prediction model with a time-to-event outcome. *Stat Med* 2022;41:1280-95. doi:10.1002/sim.9275
31 White IR, Royston P. Imputing missing covariate values for the Cox model. *Stat Med* 2009;28:1982-98. doi:10.1002/sim.3618
32 Lau B, Lesko C. Missingness in the Setting of Competing Risks: from missing values to missing potential outcomes. *Curr Epidemiol Rep* 2018;5:153-9. doi:10.1007/s40471-018-0142-3
33 Rubin DB. Multiple Imputation for Nonresponse in Surveys. 1987. Wiley & Sons, New York. doi:10.1002/9780470316696
34 Fine JP, Gray RJ. A Proportional Hazards Model for the Subdistribution of a Competing Risk. *J Am Stat Assoc* 1999;94:496-509. doi:10.1080/01621459.1999.10474144.
35 Feakins BG, McFadden EC, Farmer AJ, Stevens RJ. Standard and competing risk analysis of the effect of albuminuria on cardiovascular and cancer mortality in patients with type 2 diabetes mellitus. *Diagn Progn Res* 2018;2:13. doi:10.1186/s41512-018-0035-4
36 Kawaguchi ES, Shen J, Li G, Suchard M. A Fast and Scalable Implementation Method for Competing Risks Data with the R Package fastcmprsk. arXiv: Computation. 2019.
37 Royston P, Ambler G, Sauerbrei W. The use of fractional polynomials to model continuous risk variables in epidemiology. *Int J Epidemiol* 1999;28:964-74. doi:10.1093/ije/28.5.964
38 Schoenfeld D. Partial residuals for the proportional hazards regression model. *Biometrika* 1982;69:239-41. doi:10.1093/biomet/69.1.239.
39 Graw F, Gerds TA, Schumacher M. On pseudo-values for regression analysis in competing risks models. *Lifetime Data Anal* 2009;15:241-55. doi:10.1007/s10985-008-9107-z
40 Andersen PK, Perme MP. Pseudo-observations in survival analysis. *Stat Methods Med Res* 2010;19:71-99. doi:10.1177/0962280209105020
41 Royston P. Tools for Checking Calibration of a Cox Model in External Validation: Approach Based on Individual Event Probabilities. *Stata J* 2014;14:738-55. doi:10.1177/1536867X1401400403.
42 Austin PC, Fine JP. Practical recommendations for reporting Fine-Gray model analyses for competing risk data. *Stat Med* 2017;36:4391-400. doi:10.1002/sim.7501
43 Rahman MS, Ambler G, Choodari-Oskooei B, Omar RZ. Review and evaluation of performance measures for survival prediction models in external validation settings. *BMC Med Res Methodol* 2017;17:60. doi:10.1186/s12874-017-0336-2

44 Riley RD, Ensor J, Snell KI, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *BMJ* 2016;353:i3140. doi:10.1136/bmj.i3140
45 Marshall A, Altman DG, Holder RL, Royston P. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. *BMC Med Res Methodol* 2009;9:57. doi:10.1186/1471-2288-9-57
46 Snell KI, Ensor J, Debray TP, Moons KG, Riley RD. Meta-analysis of prediction model performance across multiple studies: Which scale helps ensure between-study normality for the C-statistic and calibration measures?*Stat Methods Med Res* 2018;27:3505-22. doi:10.1177/0962280217705678.
47 Debray TP, Damen JA, Riley RD, et al. A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes. *Stat Methods Med Res* 2019;28:2768-86. doi:10.1177/0962280218785504
48 Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ* 2016;352:i6. doi:10.1136/bmj.i6
49 National Clinical Guideline Centre. Cardiovascular risk assessment and the modification of blood lipids for the primary and secondary prevention of cardiovascular disease. [NICE clinical guideline CG181.] NICE, 2014. https://www.nice.org.uk/guidance/cg181
50 Guirguis-Blake JM, Michael YL, Perdue LA, Coppola EL, Beil TL. Interventions to Prevent Falls in Older Adults: Updated Evidence Report and Systematic Review for the US Preventive Services Task Force. *JAMA* 2018;319:1705-16. doi:10.1001/jama.2017.21962
51 Lamb SE, Jørstad-Stein EC, Hauer K, Becker C, Prevention of Falls Network Europe and Outcomes Consensus Group. Development of a common outcome data set for fall injury prevention trials: the Prevention of Falls Network Europe consensus. *J Am Geriatr Soc* 2005;53:1618-22. doi:10.1111/j.1532-5415.2005.53455.x
52 Womack JA, Murphy TE, Bathulapalli H, et al. Serious Falls in Middle-Aged Veterans: Development and Validation of a Predictive Risk Model. *J Am Geriatr Soc* 2020;68:2847-54. doi:10.1111/jgs.16773
53 Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* 2015;350:g7594. doi:10.1136/bmj.g7594
54 Banerjee A, Clementy N, Haguenoer K, Fauchier L, Lip GY. Prior history of falls and risk of outcomes in atrial fibrillation: the Loire Valley Atrial Fibrillation Project. *Am J Med* 2014;127:972-8. doi:10.1016/j.amjmed.2014.05.035
55 Stewart D, Madden M, Davies P, Whittlesea C, McCambridge J. Structured medication reviews: origins, implementation, evidence, and prospects. *Br J Gen Pract* 2021;71:340-1. doi:10.3399/bjgp21X716465
56 Conroy RM, Pyörälä K, Fitzgerald AP, et al, SCORE project group. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. *Eur Heart J* 2003;24:987-1003. doi:10.1016/S0195-668X(03)00114-3
57 Schulz M, Krueger K, Schuessel K, et al. Medication adherence and persistence according to different antihypertensive drug classes: A retrospective cohort study of 255,500 patients. *Int J Cardiol* 2016;220:668-76. doi:10.1016/j.ijcard.2016.06.263

Supplementary information: additional material