How Much Data are Required to Develop and Validate a Reliable Risk Model?

Khadijeh Taiyari

Department of Statistical Science

University College London

Doctor of Philosophy

June 2017

I, Khadijeh Taiyari, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

It has been suggested that when developing risk prediction models using regression, the number of events in the dataset should be at least 10 times the number of parameters being estimated by the model. This rule was originally proposed to ensure the unbiased estimation of regression coefficients with confidence intervals that have correct coverage. However, only limited research has been conducted to assess the adequacy of this rule with regards to predictive performance. Furthermore, there is only limited guidance regarding the number of events required to develop risk prediction models using hierarchical data, for example when one has observations from several hospitals. One of the aims of this dissertation is to determine the number of events required to obtain reliable predictions from standard or hierarchical models for binary outcomes. This will be achieved by conducting several simulation studies based on real clinical data.

It has also been suggested that when validating risk prediction models, there should be at least 100 events in the validation dataset. However, few studies have examined the adequacy of this recommendation. Furthermore, there are no guidelines regarding the sample size requirements when validating a risk prediction model based on hierarchical data. The second main aim of this dissertation is to investigate the sample size requirements for model validation using both simulation and analytical methods. In particular we will derive the relationship between sample size and the precision of some common measures of model performance such as the C statistic, D statistic, and calibration slope.

The results from this dissertation will enable researchers to better assess their sample size requirements when developing and validating prediction models using both standard (independent) and clustered data.

Contents

List of Figures

LIST OF FIGURES

List of Tables

Acknowledgements

I would like to thank the following people without whom this work would either not have happened or would have been a lot more stressful and difficult.

My supervisors: first and foremost Gareth Ambler and Rumana Omar for their support, ideas, and comments all the way through my research.

My colleagues and friends at UCL for making it such a nice place to work, in particular Matina Rassias, Oya Kalaycioglu, and Chienju Lin without whom I could not go through any of those difficult times during my PhD.

My family: Mum and Dad, who supported and helped me all my life and during my PhD (specially and more importantly the last year of PhD) ; my sister, whom I can always rely on to make my life easier and more enjoyable. Last but certainly not least, my husband Ali, whose unwavering support and enthusiasm for everything I do is amazing; my little son Sam, whom I love the most, I owe them my last and most heartfelt thank you for being in my life, all their love, and encouragement.

Chapter 1

Introduction

This chapter gives an overall overview of the study covering the context of the study, the rationale for and significance of the study, and the organisation of the dissertation. The first part of this chapter, the context of the study, covers the background and general schema of the dissertation. It also contextualises the guiding research questions and the main objectives of the study. The second part mainly discuses the motivations to carry out the studies of the dissertation. Finally, the organisation of this dissertation is presented in part three.

1.1 The context of the study

Risk modelling is a method by which factors related to a desired health outcome can be modelled, provided that information for those factors are collected in the studied data. Risk models are ever increasingly being used in medical research and commonly utilised to predict a patient's outcome in order to provide additional information to both patient and clinician. That is, they are employed to provide quantitative knowledge about the probability of outcomes in a defined patient population for patients with different characteristics (Moons et al., 2009).

For example, in the field of neurosurgery when the aim is to estimate the probability of death within 14 days in patients with traumatic brain injury or is to estimate the probability of severe disability at six months in those patients, a risk model can be used to serve those aims based on the information collected from all the relevant factors.

A risk model is usually developed by fitting a multivariable regression model to a dataset known as the development dataset (Harrell (2001); Royston et al. (2009)). The choice of regression varies depending on the outcome. For example, logistic regression is commonly exploited when the outcome of interest is binary.

Risk models play a vital role in medical research and should always be developed with extra care so that they can safely be used in practice. One of the key requirements to obtain a valid risk model is to use sufficient data when developing it (Harrell (2001); Steyerberg (2009)), where a valid model can produce correct predictions for future patients. The size of a dataset should be large enough to reflect the population of study, but not to waste resources.

A common problem in situations in which sample size is small is over-fitting, where the model appears to perform well in data used to fit the model but not with new data. An over-fitted logistic prediction model, for example, overpredicts the log-odds of an event and thus the range of predictions produced by the model is too wide; that is, the predictions are too extreme.

To avoid the problem of over-fitting, several researchers have made recommendations regarding the maximum number of predictors to be included in a risk model (Peduzzi et al. (1995); Steyerberg et al. (2000); Harrell (2001), Vittinghoff and McCulloch (2007); Steyerberg (2009); Courvoisier et al. (2011a)). A common recommendation in a binary outcome setting is that the ratio of events to the candidate predictors should be at least ten where the number of events is defined as the number of deaths in the binary outcome. This ratio is called events per variable (EPV). For example, there should be at least 50 events to develop a model with 5 predictors which are planned to be in the model. There are; however, other guidelines in the literature, such as to use EPV of at least 20 (Jinks, 2012) or 5 (Vittinghoff and McCulloch, 2007) when developing a risk model.

However, 'EPV rule of ten' was originally developed to ensure the accurate and precise estimation of regression coefficients and may not be suitable when the aim of developing the risk model is prediction rather than estimation. That is because there is less interest in individual covariate effects when developing a risk prediction model. Rather, the aim is likely to measure the ability of the model to predict outcomes for future patients (Copas (1983)). One of the aims of this dissertation is to determine the number of events required to obtain reliable predictions from standard models for

binary outcomes. We will achieve this by conducting several simulation studies based on real clinical data.

Moreover, in practice, adhering to *EPV rule of at least ten* may not be possible especially if one studies rare diseases, or if the event of interest is rare. In such circumstances, it is advised to apply a single shrinkage factor to the estimated regression coefficients after model fitting. However, this may not always result in a reliable risk prediction model as the performance of the risk model is also sample size dependant. Hence, we also aim to determine the required number of events when there is a possibility of using linear shrinkage methods.

It is important that one validate the developed risk prediction model in new datasets. Care needs to be taken when calculating the required sample size for model validation, because the estimated predictive performance measures of the risk model may be unbiased, yet not precise when using a small validation dataset (Harrell, 2001, Steyerberg, 2009).

To overcome such a problem, it has been suggested that the validation data must have at least 100 events when the outcome of interest is binary (Harrell, 2001, Peek et al., 2007, Collins et al., 2015, Vergouwe et al., 2005). However, few studies have examined the adequacy of this recommendation, and so the rule is not commonly used in practice. In a review paper, Collins et al. (2014) found that only 53% of studies out of reviewed papers had 100 events or more in their validation data. The second main aim of this dissertation is to investigate the sample size requirements for model validation using both simulation and analytical methods. In particular, we will derive the relationship between sample size and the precision of some common measures of model performance such as the C statistic, D statistic and calibration slope.

Clustered data is common in medical research (Kreft and Leeuw (1998); Guo and Zhao (2000); Goldstein (2003); Twisk (2006)). For example, patients may be clustered within hospitals, clinics, or general practices. In such circumstance, patients from the same cluster have common characteristics when compared with patients from other clusters. That is, observations within clusters are dependent on each other and are independent from patients who are in other clusters.

Ignoring the dependence structure of observations in the clustered data and employing standard regression models to analysis this data may lead to unbiased but imprecise estimation for regression coefficients (Beitler and Landis (1985); Kreft and

Leeuw (1998); Localio et al. (2001); Goldstein (2003); Maas and Hox (2004); Rabe-Hesketh and Skrondal (2005); Twisk (2006); Robertson et al. (2013)).

Three common methods to model clustered binary outcomes are marginal, fixed- and random-effect regression methods (Goldstein (2003); Rabe-Hesketh and Skrondal (2005); Twisk (2006)). While marginal models can be obtained by fitting a standard logistic regression models on patients from different clusters without taking clustering into account, one can fit a standard logistic regression with fixed-effect for clusters or fit a random-effect regression model by including a random term for clusters (with specific distribution, say, Normal distribution,) into the model to account for clustering. The simplest version of a random-effect logistic model is a random-intercept logistic model, which takes clustering effect into account by allowing the intercepts to vary across clusters.

The three relevant types of predictions in the context of clustered data are marginal, median, and cluster-specific predictions. Marginal predictions reflects the average probability of the outcome for patients with the same observed values of predictors in the population, ignoring the clusters the patients came from, and can be produced by marginal, random-effect, and fixed-effect models. In contrast, while cluster-specific prediction reflects the probability of outcome for patients with regard to the clusters the patients came from, median prediction reflects the median probability of outcome for patients with the same observed values of predictors across clusters. Both median and cluster-specific predictions can be obtained using fixed- and random-effect logistic models.

Wynants et al. (2015) have suggested using an EPV of at least ten when developing a risk model using the clustered binary outcome. However, they only studied median predictions from a random-intercept logistic model. The third aim of this dissertation is to determine the sample size requirement for developing a risk prediction logistic models using the clustered binary outcome. Simulation is used. Three types of marginal, fixed-effect and random-intercept modelling techniques were investigated.

Finally, there are no guidelines on sample size requirements when validating a risk prediction models using clustered binary data. The last aim of the dissertation is to ascertain the required sample size to validate a random-intercept risk prediction logistic model using simulation methods.

1.2 Rationale for and significance of the study

The sample size is one of the most important elements of designing a study in the medical and clinical setting. In a broad sense, research design is the process of making decisions before the situation arises in which the decision has to be carried out. One of the most important decisions in designing research is to devise strategies to address the research questions with the smallest possible number of subjects. A very small sample will lead to a loss of precision and power of the study and real medical improvements are unlikely to be distinguished from chance variation. On the contrary, a very large dataset will result in a waste of resources, that is, patients and investigator's time or funding. Balance is required to ensure that a study collects sufficient data to produce statistically valid and clinically useful results, while making cost-effective use of resources and meeting deadlines.

Methodologically, before starting a probable clinical research, a sample size should always be determined. For example, in randomised controlled trials, for funding application and also for subsequent publications, a researcher ought to determine the sample size. Sample size calculations for prediction studies, which are usually retrospective, are not routinely available, meaning these studies may often be underpowered. Available formulae can be used in some particular situations, but for most analyses of prognostic data, particularly binary data, little guidance is available to researchers.

There is little guideline for addressing how many events per variable are needed to develop a risk prediction model. Thus, this study will produce a clear route for this area.

Practically, it may not be possible to adhere to the EPV recommendations for studies based on rare diseases or events. One common approach to overcome this problem is to apply a single shrinkage factor to the estimated regression coefficients after model fitting. However, this may not always result in a reliable risk prediction model. Therefore, this dissertation will also make recommendations on the number of required EPV for such circumstances in which some sort of shrinkage factor is applied.

While there is a suggestion for the number of events to validate a risk prediction model in the literature, few studies have evaluated the adequacy of the recommendation. This study will also provides guidelines for such circumstances.

Apart from Wynants et al. (2015), in the case of developing a risk prediction model using clustered binary data, there is no recommendation about the required number of events per variable when developing a reliable risk prediction model. Furthermore, there are no guidelines on how many events are required to validate a reliable clustered risk prediction model. This dissertation will also give practical guidelines or recommendations for those situations. These guidelines will be for cases in which the risk model is developed using fixed-, random- effect and marginal logistic regression.

1.3 Organisation and overview of the dissertation

In drawing an accurate picture of the sample size requirements for developing and validating a reliable risk model in medical research, the dissertation is divided into seven chapters:

In chapter two, Key Concepts of Risk Modelling we give an overview of the central notions and terms of risk modelling in the medical setting. This chapter is composed of five parts. Those are as follows: Risk Models in Medicine, Developing a Risk Model, Logistic Regression, Validating a Risk Model including various measures of validation for both independent and clustered binary data, and Summary and Discussion.

Chapter three of the dissertation extensively investigates the Sample Size Requirements for Developing a Risk Model in the Context of Independent Binary Data. This chapter covers the following main parts: Sample Size (EPV) in Developing a Risk Model: a Review, Data, Case Study, Simulation Study, and Summary and Discussion.

In Chapter four, we investigate the Sample Size Requirements for Validating a Risk Model Using Independent Binary Data. Similar to Chapter three, this chapter is made up of four parts: Sample Size in Validating a Risk Model: a Review, Precision of Performance Measure and Sample Size, Case Study, Simulation Study, and Summary and Discussion.

The Sample Size Requirements for Developing a Risk Model with the Binary Clustered Data is presented in Chapter five. The chapter consists of six parts: Literature Review, Data, Case Study, Simulation Study, Comparing the Recommended Sample Size with the One Used in Common Practice, and Summary and Discussion.

Chapter six examines the Sample Size Requirements for Validating a Risk Model Using Binary Clustered Data. This chapter consists of four parts: Sample Size (EPV)

in Validating a Risk Model: a Review, Case Study, Simulation Study, and Summary and Discussion.

Finally, Chapter 7, Summary and Discussion, is devoted to closing remarks of the study including summary, main conclusions and contributions of this dissertation. Additionally, this chapter presents major implications of this study for further research in medical statistics including practical recommendations. It follows with Conclusion.

Chapter 2

Key concepts of risk modelling

In medicine, prognosis usually concerns the risk of a person developing a specific state of health over a particular time, based on his or her clinical and non-clinical profile. This chapter is organised into four main sections. Section 2.1 provides an overview on risk models in medicine. It briefly discusses the significance and application of risk models in medical research. In section 2.2, the steps should be taken to develop a risk model are briefly described. Then, the risk models used in this dissertation, logistic regression models, for both independent and clustered data are mathematically discussed in section 2.3. This section, also addresses methods to produce predictions of the models. The validation of a risk model including types of validation and issues related to it are elucidated in section 2.4. It concludes with explaining validation measures for both independent and clustered binary settings in sections 2.5 and 2.6.

2.1 Risk models in medicine

Risk models are ever-increasingly being employed in medical research for prognostic purposes. They enable care providers to estimate the probability that an individual develops an outcome (Harrell (2001); Steyerberg (2009)). These models are useful in various settings for different reasons: to inform patients about their forthcoming health-related problems, to guide practitioners about the possible future treatments; and to facilitate fairer, risk-adjusted comparisons between healthcare providers. Moreover, risk prediction models are useful in studies for several purposes. For instance, they assist in deciding the inclusion criteria or covariate adjustment in a randomised controlled

trial, or to adjust the co-founder or case-mix in comparing an outcome between centres in observational studies (Steyerberg et al., 2010). Such models are also referred to as prediction rules, prediction models, prognostic models, and risk scores (Moons et al., 2009). There are some risk prediction models which are commonly used in practice, such as the Framingham Risk Score (Wilson et al., 1998), QRISK2 (Hippisley-Cox et al., 2008), and Gail risk model 2 (Costantino et al., 1999). Some details of these risk models are briefly described here.

The famous risk model which is usually used as a reference in cardiovascular disease is Framingham Risk Score (Wilson et al., 1998). This sex-specific model was originally developed using data from the Framingham Heart Study and validated in the US (D'Agostino et al., 2001). The predictors included in this model were age, sex, low density lipoproteins (LDL) cholesterol, high density lipoprotein (HDL) cholesterol, blood pressure, whether the patient is treated or not for his/her hypertension, diabetes, and smoking. This risk score was updated (D'Agostino et al., 2008) to include dyslipidemia, age range, hypertension treatment, smoking, and total cholesterol. In this new version, diabetes was excluded from the model. This model is employed to estimate the 10-year cardiovascular risk of a subject.

The QRISK2 risk prediction model was developed and validated using a large general practice database in England and Wales (Hippisley-Cox et al., 2008). This sex-specific risk score is the updated version of QRISK1 (Hippisley-Cox et al., 2007). The predictors in QRISK2 are age, sex, cholesterol/HDL ratio, systolic blood pressure, diabetes, smoking status, self-assigned ethnicity, family history of coronary heart disease in a first degree relative under the age of 60, deprivation, treated hypertension, body mass index, rheumatoid arthritis, chronic kidney disease, and atrial fibrillation. The QRISK2 model is used to predict the ten year risk of developing cardiovascular disease in patients from different ethnic groups.

Gail's risk model 2 (Costantino et al. (1999); Gail et al. (1989)) is one of the most commonly employed risk prediction models for breast cancer in the clinics. The data was from a randomised placebo-controlled study conducted over six years to study the effects of tamoxifen in a population of women at high risk of breast cancer. The predictors included in this model were current age, age at menarche, age at first birth or nulliparity, number of previous breast biopsies, number of first degree relatives with breast cancer, and the presence of an atypical hyperplasia on biopsy. This model is

employed to estimate the chance of breast cancer development in currently healthy women within the next five years and during their lifetime.

2.2 Developing a risk model

The process of constructing a function of candidate factors or predictors related to the future outcome of the patients is recognised as developing a risk model. This process may not be as simple as fitting a model on a dataset. There are a number of decisions that model makers must make before developing such a model which will affect the model and the results of the study (Royston et al. (2009); Harrell (2001); Justice et al. (1999); Moons et al. (2009)). These decisions are as follows.

First, all candidate variables for likely inclusion in the model must be chosen. Candidate predictors can be obtained from patient demographics, clinical history, physical examination, disease characteristics, test results, and previous treatment.

Second, the quality of the data should be appraised and all steps to handle missing values should be determined.

Third, the strategy to select the important variables in the final model and to model continuous variables should be chosen. There are options such as stepwise variable selection, incomplete principal component, and clustering variables using clinical knowledge to select important variables (Harrell, 2001).

Finally, the choice of validation data should be decided. One may need to decide on whether data should be set aside to validate the risk model (see section 2.4).

Royston et al. (2009) noted that there are other important areas which should be considered when developing a risk model, including investigation of the robustness of the final risk model to outliers and influential observations, research for possible interactions between predictors, and determining whether and how to adjust the model for overfitting (see section 2.4.1).

Subsequent to these, one may find the suitable modelling strategy to describe a relationship between predictors and the outcome of interest. The choice of strategies for modelling depends on the type of outcome variables.

The observations of the outcome variable might be of an independent or of a clustered structure. For example, patients might be clustered within hospitals, cities, or countries.

2.3 Logistic regression

In this dissertation logistic regression, for both independent and clustered data, was used to develop risk models for two main reasons. First, the real data used in this dissertation has a binary outcome (0/1) (see section 3.2 and 5.2). This type of outcome is amongst the most common outcomes in medical research. Second, the logistic regression modelling technique is commonly used in practice to develop a risk model when the outcome of interest is binary.

2.3.1 Standard logistic regression

Let $Y_i \quad (i = 1, \ldots, N)$ be a binary outcome for the ith patient which follows the Bernoulli distribution with the probability $p_i = Pr(Y_i = 1)$, the probability that the individual experiences the event of interest, X_k the kth predictor $(k = 1, \ldots, K)$. The logistic regression model expresses p_i as a linear combination of predictors X_{ki}, using the logit as a link function:

$$\text{logit}(\text{p}_\text{i}) = \frac{\text{p}_\text{i}}{1 - \text{p}_\text{i}} = \alpha + \sum_{k=1}^{K} \beta_k X_{ki}, \tag{2.3.1}$$

The intercept α and regression coefficients β_k are estimated using maximum likelihood. The predicted probability (call it standard predictions) of an event is computed by taking the inverse logit of the linear predictor of the estimated model.

$$\eta_i = \hat{\alpha} + \sum_{k=1}^{K} \hat{\beta}_k X_{ki}, \tag{2.3.2}$$

$$\hat{p}_i = \frac{1}{1 + exp(-\eta_i)}, \tag{2.3.3}$$

In this study, the maximum likelihood logistic regression model was fitted onto the data using command *logit* from STATA 14.2 MP.

2.3.2 Logistic regression for clustered binary outcomes

Clustered data are common in medical research; for example patients may be clustered within hospitals or general practices. Several multilevel or hierarchical binary regression models have been developed to model the clustered binary outcome (Zeger et al. (1988);

Neuhaus et al. (1991); Gelman and Hill (2007)). Three very common modelling methods for clustered data are marginal, fixed-effect, and the random-effect modelling approach (Skrondal and Rabe-Hesketh (1984); Goldstein (2003); Twisk (2006); Steele (2017); Kahan (2014)).

Let Y_{ij} be the outcome for the ith patient ($i = 1, \ldots, n_j$) from the jth cluster ($j = 1, \ldots, J$) of size n_j, which takes a value of one for an event (e.g. if the patient has died) and a value of 0 for a nonevent (e.g. if the patient is alive), X_{kij} the kth predictor ($k = 1, \ldots, K$) and $p_{ij} = P(Y_{ij} = 1)$.

One can obtain three types of predictions for the ith patient from the jth cluster based on whether or not there was information from jth cluster at the time of developing the model; cluster-specific (p_{iju}), median (p_{ij0}), and marginal (p_{ijm}) predictions. Rabe-Hesketh and Skrondal (2005) suggested that one should use p_{iju} for the ith patient from the known jth cluster. We will discuss how one can obtain these predictions from each type of model after describing it.

Marginal logistic regression

The standard logistic regression model is a marginal model, and is fitted on patients from different clusters without taking clustering into account when the outcome of the interest is clustered binary. The marginal logistic regression model expresses p_{ij} as a linear combination of predictors X_{kij}, using the logit as a link function:

$$\text{logit}(\text{p}_{\text{ij}}) = \alpha + \sum_{k=1}^{K} \beta_k X_{kij}, \tag{2.3.4}$$

In this model, regression coefficients β_k represent the average effects in the population and the predicted probability for a patient indicates the average probability of patients with the same observed values of predictors, ignoring the clusters to which those patients belong. The predicted probability of an event is computed by taking the inverse logit of the linear predictor of the estimated model.

$$\eta_{ij\,m} = \hat{\alpha}_m + \sum_{k=1}^{K} \hat{\beta}_{k\,m} X_{kij}, \tag{2.3.5}$$

$$\hat{p}_{ij\,m} = \frac{1}{1 + exp(-\eta_{ij\,m})}, \tag{2.3.6}$$

This type of prediction is also called population-averaged predictions.

Furthermore, marginal predictions can also be obtained using random-effect logistic regression models (further details for this type of modelling are discussed shortly) by integrating cluster-specific predictions over prior random-effect distribution (Lee and Nelder (2004); Skrondal and Rabe-Hesketh (1984)) or using a generalised estimation equation (GEE) (Goldstein, 2003). Wynants et al. (2016) has reported that the marginal predictions obtained using the standard logistic regression model were similar to those obtained using the random-intercept logistic model. Thus, we use standard logistic regression to obtain marginal predictions.

Fixed-effect logistic model

In clustered binary data, one can also use a fixed-effect logistic regression model. A fixed-effect logistic regression model is the standard logistic regression model which is fitted on patients from different clusters, taking clustering into account by including dummy variables for all clusters but one into the model (Kahan, 2014).

The fixed-effect logistic regression model expresses p_{ij} as a linear combination of predictors X_{kij} and dummy variables $I_l \quad (l = 1, ..., J-1)$, using the logit as a link function:

$$\text{logit}(\text{p}_{\text{ij}}) = \alpha + \sum_{k=1}^{K} \beta_k X_{kij} + \sum_{l=1}^{J-1} \gamma_l I_l \tag{2.3.7}$$

where I_j is one if patient i belongs to cluster j and 0 otherwise, and γ_l presents the effect of cluster l. In this study, the maximum likelihood fixed-effect logistic regression was fitted on data using command *logit* from STATA 14.2 MP.

Like the random-intercept model, the fixed-effect model is a cluster-specific model and the regression coefficients β_k present the predictor effects within clusters.

One can obtain a cluster-specific prediction $p_{iju(fe)}$ by taking the inverse logit of the linear predictor of the estimated model.

$$\eta_{iju(fe)} = \hat{\alpha}_{(fe)} + \sum_{k=1}^{K} \hat{\beta}_{k\,(fe)} X_{kij} + \hat{\gamma}_j \tag{2.3.8}$$

$$\hat{p}_{iju(fe)} = \frac{1}{1 + exp(-\eta_{iju(fe)})}. \tag{2.3.9}$$

One can also obtain a median prediction from the fixed-effect logistic regression model as follows:

$$\eta_{ij0(fe)} = \hat{\alpha}_{(fe)} + \sum_{k=1}^{K} \hat{\beta}_{k\,(fe)} X_{kij} - \frac{1}{J-1} \sum_{l=1}^{J-1} \gamma_l, \tag{2.3.10}$$

$$\hat{p}_{ij0(fe)} = \frac{1}{1 + exp(-\eta_{ij0(fe)})}, \tag{2.3.11}$$

Random-intercept logistic regression

In clustered data, a random-effect model logistic regression model can be used for model development (Twisk, 2006). The simplest version of a random-effect logistic model is the random-intercept logistic regression model. Alongside fixed-effect predictors (X_{kij}), one adds a random variable (random-intercept; u_j) to the model to take clustering into account. The random-intercept is usually assumed to be normally distributed with a mean of zero, and allows the intercepts to vary across clusters.

The random-intercept logistic model expresses p_{ij} as a linear combination of the predictors and random-intercept u_j:

$$\text{logit}(\text{p}_{\text{ij}}) = \alpha + \sum_{k=1}^{K} \beta_k X_{kij} + u_j \tag{2.3.12}$$

where $u_j \sim N(0, \sigma_u^2)$. The random-effect model is a cluster-specific model and regression coefficients β_k represent the effects of predictors within clusters. The cluster-specific predicted probability of an event given the random-intercept for the jth cluster is computed by taking the inverse logit of the linear predictor of the estimated model.

$$\eta_{iju(re)} = \hat{\alpha}_{(re)} + \sum_{k=1}^{K} \hat{\beta}_{k(re)} X_{kij} + \hat{u}_j \tag{2.3.13}$$

$$p_{iju(re)} = \frac{1}{1 + exp(-\eta_{iju(re)})}. \tag{2.3.14}$$

One can also obtain a median prediction for a patient from a cluster which was not known when developing the risk model by replacing the random-intercept with

the average random-intercept ($\hat{u}_j = 0$), and then using the inverse logit of the linear predictor of the estimated model.

$$\eta_{ij0(re)} = \hat{\alpha}_{(re)} + \sum_{k=1}^{K} \hat{\beta}_{k\,(re)} X_{kij} + 0, \tag{2.3.15}$$

$$\hat{p}_{ij0(re)} = \frac{1}{1 + exp(-\eta_{ij0(re)})}, \tag{2.3.16}$$

It is worth mentioning that the logit function is nonlinear, and thus the inverse logit of the average effect is not equal to the average, but to the median. That is the reason why this prediction refers to the median prediction.

In this study, the maximum likelihood adaptive Gaussian quadrature (AGQ) random-intercept logistic regression was fitted on data using the command *gllamm* (Skrondal and Rabe-Hesketh (1984); Rabe-Hesketh and Skrondal (2005); Rabe-Hesketh et al. (2002)) and *meqrlogit* from STATA 14.2 MP. However, the command *gllamm* was very slow for the conduction of simulation study. Therefore, we only used this command for the case study, and command *meqrlogit* for the simulation study.

2.4 Validating a risk model

Apart from developing the risk prediction model, an important part of constructing a risk prediction model is validation (Royston et al., 2009). This section of the chapter discusses the main concepts related to validating a risk model by answering the following questions:

What Does validation of a risk model mean?

The idea of validating a risk model does generally mean establishing that the model performs well in a new dataset (Harrell et al. (1996); Andreas et al. (1997); Altman and Royston (2000); Bleekera et al. (2003); Moons et al. (2009); Royston et al. (2009); Harrell (2001)). In other words, assessing the quality of predictions or performance of the model in new patients is known as validating. Thus, if the model passes this test it is said that model is validated (Altman and Royston, 2000).

Why is validation of a risk model needed?

A risk model might not successfully pass a validation test. There are three main reasons that may cause a risk model to fail to be validated (Harrell, 2001):

The standard modelling methods might not be efficient. Most of standard statistical methods, employed to fit risk models, are data-dependent. Hence, they might give an exaggerated judgement of predictive performance. For example, there is always a large number of potential variables in the development stage of the risk model. Therefore, the important variables should be selected to appear in the model. The data-dependent aspect of most models is based on the variable selection methods, where stepwise selection is commonly used by researchers. However, the desirable approach is based on applying clinical knowledge along with statistical methods to reduce the number of candidate variables and therefore the risk of an overfitted model. The overfitted model cannot produce as good predictions in the validation data as it does in the development data.

The design of prognostic studies might not be efficient. Most observational studies aim to accomplish the same quality results as experimental studies. The existence of implicit inclusion and the exclusion criteria, exclusion of many patients due to missing data, which may be missing not at random, and insufficient sample size may result in misleading findings, creating overfitting and/or bias. The definition of the characteristics of the sample is of clear importance to the clinician who wishes to know whether a model is relevant to a particular patient.

Models may not be transportable. Even with flawless methodology of a study, a model may not be generalisable for a different case-mix population (Steyerberg (2009); Altman and Royston (2000)). This will usually take place when one or more important variables is not present in the model. The problem is that one can never be sure that all important variables are in the model.

For these reasons, it is strongly recommended to examine the performance of a risk model on a new series of patients, ideally in a different location (Harrell (2001); Vergouwe et al. (2005); Royston et al. (2009)).

As mentioned, overfitting is a common problem in the development stage. Hence, we discuss it here.

2.4.1 Overfitting

A model may be overfitted if the estimated performance of it is overly optimistic. Overfitting may cause an unimportant predictor to appear predictively important (Peduzzi et al., 1995). In other words, large effect size are estimated very large and small effect size are estimated very small in a overfitted model (Steyerberg et al., 2001). Overfitting might occur because the same data are used for model development and validation, or because the development data is really small compared to complexity of the model (Steyerberg et al. (2001); Harrell (2001)). To overcome such a problem, and to improve the quality of the predictions, the application of shrinkage factor has been suggested (Harrell et al. (1996);Steyerberg et al. (2001); Ambler et al. (2011)).

The most straightforward process is to obtain shrunk regression estimates by multiplying the ML estimates by a single linear shrinkage factor. Such a process shrinks the estimated coefficient equally. It is given by

$$\hat{\eta}_{new} = \overline{\eta} + c(\hat{\eta} - \overline{\eta})$$

where c is the shrinkage factor, and $\hat{\eta}$ and $\bar{\eta}$ are estimated and mean linear predictor. In nontechnical terms, it means that the estimated prognostic index is drawn towards the average. This method is also called the postestimation shrinkage technique.

Two common linear shrinkage factors are heuristic and bootstrap. These are discussed here.

Heuristic shrinkage factor

The Heuristic shrinkage factor was first introduced by Copas (1983) and then Van Houwelingen and Le Cessie (1990) generalised this method. It is given by

$$\hat{s}_{heur} = \frac{model\chi^2 - k}{model\chi^2},$$

where k is the number of regression parameters, excluding the intercept but including all non-linear and interaction effects. The *model* χ^2 is the difference in - *2log likelihood* of the final model and the null model (Miller et al., 1991). The heuristic shrinkage uses *model* χ^2, as a quantity of its strength, to shrink the optimism of the fitted model towards zero where the optimism is the difference between the original performance of the model and its estimated performance. When *model* χ^2 is less than the number of

parameters the estimated heuristic shrinkage becomes negative. That is, the model is only fitting noise, and is often the case when the dataset is small. In other words, when the heuristic shrinkage factor is very small or negative, it means that the number of variables in the model is large compared to the strength of the model. Harrell in his book discussed that when heuristic shrinkage is less than 0.9 in a model then either the shrunken estimator or data reduction might be required (Harrell, 2001).

Bootstrap shrinkage

Bootstrap shrinkage is another method that helps to shrink the coefficients of the model in the developing stage. Based on this method, the amount of optimism is quantified and used to shrink the model (Steyerberg (2009); Harrell (2001)). The following procedure is carried out to obtain Bootstrap shrinkage. First, a model is fitted on the development sample and linear predictors ($\hat{\eta}$) for all patients are obtained. Second, a sample with a replacement is taken from development data. This sample is also referred to as a bootstrap sample, and is the same size as the development sample. Third, the same model is fitted on Bootstrap sample and used to produce linear predictors ($\hat{\eta}_{boot}$) for all patients. Afterwards, a model is fitted on the bootstrap sample using $\hat{\eta}_{boot}$ as a sole predictor and corresponding outcomes. Then, the slope of the linear predictor (β_{boot}) is estimated. The process is performed several times, say 100. The average ($\hat{\beta}_{boot}$) over the number of bootstrap samples forms a bootstrap shrinkage. This method mimics the calculation of the calibration slope (see section 2.5), and corrects the predictions accordingly.

How should a risk model be validated?

To validate a risk model, one can use development data or new data. By ascending order according to their rigorous performance, validation methods are apparent, internal, temporal and external (Harrell, 2001, Steyerberg, 2009, Vergouwe et al., 2005). A brief description of these approaches are as follows.

Apparent validation: in apparent validation, a researcher quantifies the performance of the model only in the development data which is used to develop the risk model. That is, the data is employed for both estimating parameters of the model and testing the quality of predictions produced by that model (Steyerberg, 2009). It is known that such validation tends to produce overoptimistic measures (Moons et al., 2009).

Internal validation: the internal validity is regarded as a method that examines reproducibility of the models (Konig et al., 2007). One common type is *data-splitting*. Here, the sample is split into two parts before the modelling begins, for example 50% for development and the rest for validation. The model is derived from the development data, and its predictions are evaluated on validation data. A problem is how to split the data. Furthermore, estimates of predictive accuracy from data-splitting procedures, though unbiased, tend to be imprecise (Harrell, 2001).

The second common method of internal validation is *cross-validation* (Harrell (2001); Konig et al. (2007); Steyerberg (2009)), which is an extension of split data such that the data is split up into several parts: one part is used for validation and the rest for development at the time. This is repeated several times and the average is taken as an estimate of performance. The most extreme variant of cross-validation is the *Jack Knife*, where one observation is picked out and the rest are used for model development, and the model is validated on the observation which was left out (Harrell, 2001). One benefits from cross-validation more than data-splitting for two reasons: firstly, the size of the development samples can be much larger, so less data are discarded from the estimation process; secondly, cross-validation reduces variability by not relying on a single sample split.

The other technique in internal validation is *bootstrap validation* (Harrell (2001); Konig et al. (2007); Steyerberg (2009)). The Bootstrap validation can be computationally extensive depending on the characteristics of studied data, and the type of method which is used to fit the risk model. The Bootstrap validation is conducted as follows. First, random samples with replacement are taken from the development data; they are also called bootstrap samples. These samples are the same size as the development data. Then, the risk model is fitted on each bootstrap sample and its predictions are evaluated using the development sample (Harrell, 2001). This technique is preferable and efficient especially when a limited number of observations is available. To obtain stable results, the procedure has to be repeated several times, usually at least 100 times.

The internal validation techniques differ with regard to how well they predict the performance of the model in independent patients but all still operate only in the original data. Further validation procedures may be required for a more stringent check of the generalisability of the model (Hosmer and Lemeshow (2001); Konig et al. (2007); Altman et al. (2009)).

Temporal validation: in this approach, the performance of a model is evaluated on subsequent patients within the same centre(s) (Hosmer and Lemeshow (2001); Konig et al. (2007); Altman et al. (2009)). In principle, this technique is similar to splitting a single dataset into two cohorts seen in different time periods. Structural differences between datasets arise from temporal changes, such as revised diagnostic criteria, or different referral schemes in the hospitals. Typically, it is difficult to eliminate this variability because of employing specific inclusion/exclusion criteria. Thus, the case-mix might be different (Konig et al. (2007); Altman et al. (2009)).

External validation: neither internal nor temporal evaluation addresses the wider issue of the generalisability of the model. The external approach is desirable since it evaluates a model on new data collected in a different centre. The external validity checks the generalisability of the risk model or its transportability (Justice et al. (1999); Harrell et al. (1996); Vergouwe et al. (2005); Konig et al. (2007); Moons et al. (2009)). The results from this validation may differ from those from internal validation, since many aspects may be different between settings including the selection of patients, definition of variables, and diagnostic or therapeutic procedures (Miller et al. (1991); Moons et al. (2009)).

What aspects of a risk model should be validated?

The performance of the models is commonly assessed in terms of calibration, discrimination, and overall performance (Spiegelhalter (1986); Miller et al. (1991); Vergouwe et al. (2005)).

Calibration: this refers to the agreement between observed outcome frequencies and predicted probabilities. For instance, if the model predicts that in-hospital mortality after heart valve surgery is 70%, that model is well calibrated only if the observed in-hospital mortality is 70%. Measures such as the calibration slope and Hosmer-Lemeshow test are used to investigate calibration of risk prediction models.

Discrimination: this refers to the ability of the model to distinguish patients from different risk groups. For example, a risk model with good discriminating ability can adequately separate low- and high-risk patients. The discriminatory ability of a risk prediction model is commonly measured using the area under the ROC curve or C statistic. Other methods such as the D statistic are also used in the literature.

Overall performance: this measures the quality of predictions for each subject in the dataset. Overall performance measures incorporate both calibration and discrimination aspects (Harrell, 2001). The Brier score is one of the most commonly used measures to quantify the overall performance of a risk prediction model.

2.5 Validation measures for independent binary data

This section elucidates measures used throughout this dissertation to measure the validity of risk models. These measures are evaluated in the validation data using predictions derived from the models fitted on the development data.

Calibration slope

The calibration slope measures calibration or the agreement between the estimated log-odds and actual log-odds (Miller et al., 1991). The calibration slope is given by:

$$\text{logit}(p_i) = \alpha + \beta\hat{\eta}_i$$

If the model is well calibrated, then, the estimated calibration slope $\hat{\beta}$ should be close to one. The overfitted model tends to show a slope of less than one indicating that low predictions are too low, and high predictions are too high (Miller et al., 1991). The $\hat{\alpha}$ is different from zero indicates that the predicted probabilities are systematically too high ($\hat{\alpha} < 0$), or too low ($\hat{\alpha} > 0$). If $\hat{\alpha}$ is different from zero and $\hat{\beta}$ different from one, the interpretation of the mis-calibration is hard since the values of $\hat{\alpha}$ and $\hat{\beta}$ are not independent (Miller et al., 1991). In practice, the calibration slope is less than one reflecting a need for the shrinkage of regression coefficients that were estimated in the development dataset (Copas (1983); Van Houwelingen and Le Cessie (1990); Harrell (2001); Steyerberg et al. (2003); Steyerberg (2009)).

C statistic

The concordance statistic is defined as a measure of the discrimination ability of the risk model (Harrell, 2001), and is the proportion of all pairs of patients in which the patient with the outcome of interest had a higher predicted probability. This statistic is identical to the area under the receiver operating a characteristics curve (ROC) when the outcome of interest is binary (Harrell et al., 1984). The ROC is a graph of the true

positive rate or sensitivity against false positive rates or one minus specificity of the model for all possible cut-off points (Harrell et al., 1984), where sensitivity is the ability of a test to correctly classify a patient as diseased. For example, if 100 patients known to have a disease are tested, and the test result for only 23 of them is positive, then the test has 23% sensitivity. On the contrary, the ability of a test to correctly classify a healthy individual as disease-free is called specificity. For instance, if 100 with no disease are tested, and 80 return a negative result, then the test has 80% specificity.

The C statistic of a model with no discrimination ability is expected to be around 0.5. In contrast, the model with perfect discrimination would have a C statistic of about one (Harrell, 2001). The C statistic equal one is rare in practice though. In the risk prediction modelling, it typically ranges from 0.6 to 0.85; higher values are seen in diagnostic settings (Vergouwe et al. (2005); Altman and Royston (2000); Harrell (2001)).

The C statistic can be estimated in both parametric and nonparametric approaches (Dorfman and Alf (1969); Metz (1978)). The nonparametric approach will be taken in this study as it does not require any distributional assumption for the linear predictor.

For a pair of patients (i, j) the nonparametric C statistic is given by

$$C = Pr(p_i > p_j | Y_i = 1 \; and \; Y_j = 0), \; \forall (i, j)$$

from one-to-one transformation between the probability of event of interest (p) and the linear predictor (η); the above expression can be written as:

$$C = Pr(\eta_i > \eta_j | Y_i = 1 \; and \, Y_j = 0), \quad \forall (i, j).$$

To compute the parametric C statistic, let N_0 and N_1 denote the total number of subjects who experienced event and nonevent, respectively. Also, let p_0 and p_1 denote the probability of nonevent and event, respectively. Therefore, the C statistic can be estimated as follows

$$C = \frac{1}{N_0 N_1} \sum_i \sum_j I(\eta_i^{(1)}, \eta_j^{(0)}),$$

where $\eta^{(1)}, \eta^{(0)}$ are linear predictors that correspond to the ith person who did and the jth person who did not experience the event of interest; $I(\eta_i^{(1)}, \eta_j^{(0)})$ is defined as

follows.

$$I(\eta_i^{(1)}, \eta_j^{(0)}) = \begin{cases} 1 & \text{if } \eta_i^{(1)} > \eta_j^{(0)} \\ 0.5 & \text{if } \eta_i^{(1)} = \eta_j^{(0)} \\ 0 & \text{if } \eta_i^{(1)} < \eta_j^{(0)} \end{cases}$$

D statistic

The D statistic estimates separation or discrimination between subjects with low- and high-risk (Royston and Sauerbrei, 2004). This statistic was initially suggested by Royston and Sauerbrei. This is calculated by first transforming each patient's estimated linear predictor $\hat{\eta}_i$ to give standard normal order rank statistics. These rank statistics are then divided by a factor of $\sqrt{\frac{8}{\pi}}$. It can be formalised as follows.

$$z_i = (\sqrt{\frac{8}{\pi}})^{-1} \Phi^{-1} \left(\frac{i - 3/8}{N + 1/4} \right)$$

where i is the rank of the estimated linear predictor $\hat{\eta}_i$ sorted within the population of the study, N is the number of observations, $\Phi^{-1}(.)$ is the inverse standard normal distribution function, and $\sqrt{\frac{8}{\pi}} \approx 1.60$. The scaled normalised estimated linear predictor z_i is distributed as $N(0, \pi/8)$. Obtaining z for all patients in the validation data, a logistic regression is then fitted to the validation data with sole predictor z. That is,

$$\text{logit}(Y_i = 1 | z_i) = \beta_z z_i + \alpha_z,$$

$\hat{\beta}_z$ is an estimate of the D statistic. In terms of the values that the D statistic can take, it ranges between zero and plus infinity ($+\infty$) for the models with no and good prognostic information. The D statistic will be close to zero, if the average of the predicted risk ordering doesn't show any relationship with true risk ordering. Consider two groups of the lower and upper half of the estimated linear predictor; in non-technical words, the D statistic is an estimate of *log-odds* of having the event of interest between low- and high-risk groups, respectively. For example, a D statistic of three in the binary outcome setting means the log-odds of event in low-risk group is three times larger than high-risk group.

Brier score

The Brier score (BS) is often used to quantify the predictive accuracy of a risk model. This measure is a quadratic scoring rule, where the mean square differences between observed response y and estimated probability $\hat{p}$ are calculated (Glenn, 1950), where n is the number of observations. That is:

$$\text{BS} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{p}_i)^2$$

The Brier score for a risk model can range from zero for a perfect model to one for a model with no prognostic information (Ash and Shwartz, 2003). A Brier score is so close to zero when the outcome prevalence is low (Harrell, 2001), which may not mean that the model has good prognostic information.

Other measures

There are other measures available in the literature to study performance of logistic risk prediction models.

One common method, usually readily available in standard software, is *Nagelkerke's* R^2 (Nagelkerke, 1991). This measure can be obtained by rescaling the fit of the model in accordance with -2 log *likelihood*. The Nagelkerke's R^2 is the logarithmic scoring rule which is calculated by $Y \times log(P) + (1 - Y) \times log(1 - P)$.

Another, yet not so common, measure is Pearson's R^2 which considers the square differences between predictions p and the outcome y. This measure is similar to the Brier score; however, it does not involve averaging over the total number of observations.

Calibration curve is a measure to quantify calibration ability of the risk prediction model, and is the visual presentation of the relationships between observed outcome frequencies and predicted probabilities.

Calibration-in-the-large is another method to measure calibration. This is the calibration intercept when the slope is set to one.

Furthermore, the Hosmer-Lemeshow test is frequently used in risk prediction modelling to assess whether or not the observed event rates match the predicted event rates in subgroups of the sample, where subgroups are typically used based on deciles of the estimated predictions. The risk prediction model is said to be well calibrated if the predicted and observed event rates in the subgroups are similar.

2.6 Validation measures for clustered binary data

In clustered data, the performance of the risk prediction models should be assessed taking clustering into account. The predictive performance measures are often obtained using two approaches; within-cluster and overall approaches.

To calculate within-cluster measures, one calculates the standard measures (see section 2.5) within each cluster, and takes average over all cluster. On the contrary, one computes the overall measures by calculating standard measures within the population.

For each measure in the clustered data, there are as many types of measures as there are types of predictions type. For example, one can compute the within-cluster and overall calibration slope for each type of prediction when validating a random-intercept risk prediction logistic model. All these versions of the calibration slope have the same interpretation to the standard calibration slope, based on the reference value of one (see Section 2.5).

We considered using cluster-specific performance measures. However, we encountered a problem in small datasets when calculating the C statistic in cases where there was no event in some clusters, and when calculating the calibration slope where there were less than two events in some clusters. Moreover, according to Wynants et al. (2015), overall and within-cluster performance measures perform similarly. Therefore, we only used overall measures in chapters five and six. The details of four common measures are discussed now.

Calibration slope

The calibration slope (CS) for clustered binary outcomes can be obtained using the same method as the standard calibration slope. However, one can conduct it by fitting either a random-intercept logistic model with the linear predictor $\hat{\eta}_{iju}$ as the only predictor and the outcome variable Y_{ij}, or a standard logistic model with those variables thereof. If one uses a random-intercept logistic regression, the results take the following form:

$$\text{logit}(p_{iju}) = \alpha + \beta_u \hat{\eta}_{iju} + u_j, \tag{2.6.1}$$

where β_u is the estimated calibration slope. On the contrary, if one uses a standard logistic regression, the results will be in the following form:

$$\text{logit}(p_{iju}) = \alpha + \beta\hat{\eta}_{iju}, \tag{2.6.2}$$

where β is the estimated calibration slope.

The estimated β_u is equal to the estimated β, because the random effects are already included in $\hat{\eta}_{iju}$.

The same approach to that discussed above can be taken to obtain the calibration slope based on median ($\hat{\eta}_{ij0}$) and marginal ($\hat{\eta}_{ijm}$) linear predictors, respectively, but by replacing $\hat{\eta}_{iju}$ with the corresponding linear predictors.

C statistic

For a pair of subjects (i, k) from clusters (j, l), respectively, where i and k correspond to subjects who had an event and those who did not, respectively, with the probability (p_{iju}, p_{klu}) or linear predictors (η_{iju}, η_{klu}), the C statistic is defined as

$$C = Pr(p_{iju} > p_{klu}) = Pr(\eta_{iju} > \eta_{klu}), \tag{2.6.3}$$

where a pair may consist of subjects from the same cluster or from different clusters. For the subjects from different clusters, the cluster-specific random effect (u) contributes in determining whether a pair is concordant, even if both subjects have the same predictor values. For the subjects from the same cluster, however, u does not contribute in determining a concordant pair, as they share the same value of u.

The definition is the same for median and marginal predictions. Note that marginal predictions are re-scaled values of median predictions; marginal predictions re-scaled by integrating out the u in clustered-specific predictions. Therefore, the rank order based on both median and marginal predictions would be identical. This results in the C statistic obtained using them being identical.

To estimate the C statistic, one uses the same approach described in section 2.5. Note that the C statistic obtained using cluster-specific predictions is greater than that obtained using marginal predictions if clustering exists in the data, and they are equal if there is no clustering.

D statistic

The same approach as described in section 2.5 is taken to obtain the D statistic when using the clustered binary outcome. That is, one transforms linear predictors η_{iju} to z_{ij}, and then fits either a random-intercept or standard logistic model to the validation sample, consisting of z_{ij} as the only predictor and outcome variable. However, as the random effects are already included in z_{ij}, both approaches would give a similar D statistic.

For median and marginal predictions, the D statistic is obtained in a similar manner to that described above. All these versions of the D statistic have the same interpretation as those for the C statistic.

Brier score

The Brier score (BS) can be obtained by averaging the squared differences between the predicted probabilities and the observed outcomes (see section 2.5). One conducts this by using either cluster-specific, median, or marginal prediction to obtain the corresponding Brier score. Unlike the C statistic, the Brier score is obtained using median predictions that are not equal to those obtained using marginal predictions, as those predictions are not equal.

2.7 Summary and discussion

In this chapter, the main components of risk modelling in medical research have been briefly reviewed.

The next chapters of the dissertation include our main contribution to the literature. While chapters three and four will examine the sample size requirements to develop and validate a risk prediction model exploiting the independent data, chapters five and six address the same problem but in the context of clustered data.

Chapter 3

Sample Size Requirements for Developing a Risk Model Using Independent Binary Data

Risk prediction models have a vital role in medicine and need to be developed carefully so that they reflect the information in the data accurately. These models should also be valid when applied to new data. However, risk models developed in small datasets may not be accurate as they may be overfitted (Harrell, 2001). Therefore, adequate sample sizes should be used when developing a risk model.

This chapter discusses the sample size required to develop a risk prediction model in the context of independent binary outcomes. The chapter consists of the following parts. A literature review is provided in section 3.1. Section 3.2 describes the dataset used in this chapter and Chapter 4. Section 3.3 is devoted to a case study conducted to give an insight into the issue of sample size and events per variable. Section 3.4 describes and presents the results of a simulation study which was conducted to explore factors which may affect the accuracy of risk models. The final section, 3.5, is the discussion.

Literature search

For the literature search in chapters 3 to 6, we first searched online materials (e.g using Google and Google scholar) for combinations of the following keywords "Events per Variable", "EPV", "number of events", "prediction modelling", "prediction risk

modelling", "risk modelling", "developing", and "validating" (for chapters 4 and 6). We also searched online sources based on citations and references of all known articles (e.g. Harrell et al. (1984, 1985), Concato et al. (1995), Peduzzi et al. (1996) for chapter 3). However, we did not find any article using either of the aforementioned approaches for chapter 6.

3.1 EPV in developing a risk model: a review

Type of sample size calculation

One critical part of study design is determining the sample size to develop a risk model. Traditionally, the sample size for a prognostic study is calculated based on either precision or hypothesis testing (Cochran (1977), Rosner (2006)). For precision-based calculations, the sample size is obtained to find a confidence interval of a sufficiently narrow width, where the aim of the study is estimating the mean of the population or the difference in means or proportions. A small study, for instance, found that the sensitivity of a new technique to detect methicillin-resistant staphylococcus aureus (MRSA) was around 90%. Therefore, a new study should recruit 385 patients to estimate this sensitivity with a 95% confidence interval of ± 3. Note that MRSA is an infection caused by a group of bacteria called staphylococcus aureus.

Sample size, for a hypothesis testing calculation, is computed when the aim of the study is detecting a difference as statistically significant with specific power. Such a study is conducted to compare the mean or proportion in two groups. As an example, consider a randomised control trial in which patients with colonic cancer received post-op fluids either in accordance with standard practice or a restricted intake. Let us assume that the aim of the study is establishing whether two approaches differ with respect to gastric emptying time. In this study a sample size of 20 patients in each group will be sufficient to detect a difference of 30 minutes on gastric emptying time using two sample t-test with a standard deviation of 29 minutes, a power of 90%, and a significance level of 5%.

In current practice, the common sample size calculation for risk models is not based on using confidence interval (precision-based) or testing coefficients (hypothesis testing). Simple calculations are used for exploratory or prognostic studies involving risk modelling where the risk models are serving to predict the likely outcome of a disease or

ailment. These calculations obtain the size of the sample which is required to develop a risk model and differ in relation to the types of outcome. The effective sample size (Harrell et al., 1984) for a continuous outcome is equated with the total number of observations. It is, in contrast, equal to the number of events, if this is the smallest category, when analysing the binary outcome. The required effective sample size is also obtained with taking the number of potential variables into account. That is the number of events per variable (EPV) (Harrell et al., 1996). Now, the existing body of knowledge and literature on sample size using EPV is reviewed.

Using EPV

For years, many model developers such as Harbarth et al. (2000), Judith et al. (2002), Lassnigg et al. (2004), Jonas and Johnny (2005), Voerman et al. (2007), Akins et al. (2008), Stone et al. (2011) and William et al. (2013) used ten events per variable (EPV) to build a risk model and cited Peduzzi et al. (1995, 1996). However, this rule of thumb is based on the work of Harrell et al. (1984) and Harrell et al. (1985). The last four studies will be summarised now.

Harrell et al. (1984)

Harrell et al. (1984) conducted a study to quantify the performance of regression models built using development data of different sizes to predict survival in independent validation data. They used three Cox regression models; one with stepwise variable selection, one that is the incomplete principle components Cox model, and the one with clinical indexes, where the Cox regression model is a common modelling technique when the outcome of interest is time-to-event.

Harrell et al. (1984) employed data from 4226 patients with possible angina undergoing cardiac catheterisation of whom 10% died from cardiovascular disease. Their response variable was measured on a continuous scale with no censoring. They always used half of the data and three subsets of it for development and the other half of it for validation. Moreover, in their data there were 30 potential variables, both continuous and categorical.

Those three methods are described now. While in the standard stepwise variable selection model, one selects variables from candidates until no other variable remains significant at the studied significant levels; in incomplete principle components (PC)

regression, one replaces variables with subsets of PCs which explain most of the variation in the variables across patients. Each PC is the linear combination of variables. This is incomplete as in this method one does not transform the estimated coefficients of PCs back to the coefficients of the variables. In the regression model with clinical indexes, one makes clusters of variables using the clustering method, assigns weights to each variable in the clusters (if required) based on clinical intuition and uses the clusters (indexes) instead of potential variables.

In the study, Harrell et al. (1984) assessed the performance of the model using only one discrimination measure (C index) and concluded that the Cox regression model with stepwise variable selection should not be employed in studies when the EPV is smaller than ten; however, this issue was not the main focus of the study and so investigation of that aspect was limited in scope.

Harrell et al. (1985)

Harrell et al. (1985) conducted the study to compare the performance of the three regression methods that they used in previous work plus the method of recursive partitioning to predict the probability of recovery from Non-Hodgkin's Lymphoma using logistic regression. Note that the recursive partitioning methods yield a prediction rule by making subgroups of patients within which the response variable is relatively homogeneous(on the basis of their values on a set of predictor variables) and assigning each subgroup a predicted response.

Harrell et al. (1985) used ten years of data on 450 patients with Non-Hodgkin's Lymphoma, of whom 190 recovered after six doses of chemotherapy. In their data, they had 25 potential variables, both continuous and categorical. They divided the first five years into two independent development samples and always utilised the last five years for validation.

Like in the previous study, Harrell et al. (1985) used only one discrimination measure (C statistic) and drew the same conclusion as before. Harrell et al. (1985) suggested the use of data reduction method before performing regression analysis when the EPV is smaller than ten. However, again the issue of EPV was not the main focus of the publication and so was not investigated further.

Peduzzi et al. (1995), Concato et al. (1995), and Peduzzi et al. (1996)

In a similar vein, Peduzzi et al. (1995) and Concato et al. (1995) carried out a simulation study to find the effect of small EPV on accuracy, precision of estimation of the Cox regression coefficients, and the power of significance testing. Their data was from 673 patients who had coronary artery bypass surgery of which 252 died during the first ten years of follow-up. They used only categorical predictors in their model with a wide range of prevalences and hazard ratios.

Peduzzi et al. and Concato et al. assessed convergence of the models, bias in the estimation of regression coefficients, and power of the statistical test. They reported that most of the problems happened at EPV<10 and concluded that the results of studies with EPV<10 should be interpreted cautiously.

Peduzzi et al. (1996) repeated the previous study with logistic regression using the same dataset and drew the same conclusions as before.

The focus of both studies was on the accuracy and precision of the estimated regression coefficients rather than predictions. Moreover, they did not report the simulation standard error which aids judgement on stability of the results. Furthermore, in these studies, the researchers only looked into the effect of EPV, holding the sample size, the distribution, and the effects of the predictors constant at the values observed in the dataset.

Vittinghoff and McCulloch (2007)

Vittinghoff and McCulloch (2007) carried out a simulation study to find out in which situations EPV<10 was enough for acquiring correct coverage for the confidence interval of and unbiased estimation of regression coefficients. They studied both logistic and Cox regression models.

Their study was carried out from an epidemiological perspective; it focused only on a primary predictor and considered the rest of the variables as covariates. The primary predictor in their study was either discrete or continuous, but the covariates were always normally distributed. Apart from altering the EPV in their simulation process, they changed other factors; namely sample size, number of events, total number of predictors, effect size of primary predictor (varying prevalence for the categorical predictor or different coefficients for the continuous predictor). For logistic models, Vittinghoff

and McCulloch generated the required number of events and non-events separately, mimicking the case-control design. In contrast, for the Cox regression models, they first selected the required number of events and then censored the remainder at simulated time values. They concluded that the problem of a biased estimation of regression coefficients, large standard error for estimated coefficients, and non-coverage in the confidence interval appeared more at EPV less 5.

Nevertheless, by taking an epidemiological approach, the investigators did not study collinearity, or the presence of noise variables in the model, presence of various types of covariates. Similar to previous studies, Vittinghoff and McCulloch did not assess the effect of the EPV in the presence of other scenarios on the predictive performance of the model. They also did not report the simulation standard error in order for readers to judge the stability of their results.

Courvoisier et al. (2011a)

In line with previous works, Courvoisier et al. (2011a) carried out a simulation study to examine the accuracy and precision of the estimated logistic regression coefficients in relation to the EPV. They generated a set of standard normal predictors, examining the following scenarios; strength of the predictors, collinearity, the number of variables in the model (one variable, 25 variables, and seven significant variables out of 25 variables which are in the model) and the proportion of significant variables from the total number of variables in the model. They also studied different outcome prevalences using univariate models (only one predictor in the model). They simulated a population and sampled events and nonevents separately to achieve the required EPV.

They observed the estimation bias was around 20% at EPV$\leq$10 in univariable models and that it was about 15% at the same EPV in models with seven variables. In addition, the level of bias in the model with 25 variables was the same as it was in models with seven variables when EPV$\leq$ 7.

In a small part of their study Courvoisier et al. (2011a) also looked into the bias in the estimation of the C statistic in the development sample (apparent discriminating ability of the model) for all simulated models and compared the results with those in the population. They found an overestimation of the C statistic in the development sample by $>$ 10% at all EPV in univariate models with weak variables, by $>$ 10% at EPV$<$ 10 in models with seven weak variables, and by $<$ 10% at all EPV in models with

25 variables. Courvoisier et al. concluded that even if EPV exceeds 10, an unbiased and precise estimation of regression parameters may not be guaranteed.

Courvoisier et al.'s model always included continuous variables. These researchers did not study the presence or otherwise of noise variables in the model. Courvoisier et al. (2011a) only examined the apparent discrimination ability of the models. The researchers may not be able to ensure the reliability of their model unless they use data to test it that is different to the data used to build it (Harrell, 2001, Omar et al., 2004, Steyerberg, 2009). Moreover, Courvoisier et al. did not report the simulations standard error.

Steyerberg et al. (2011)

In their response to Courvoisier et al. (2011a), Steyerberg et al. (2011) indicated that Courvoisier et al.'s model with seven significant variables may not be successfully validated in future datasets because of selection bias (Steyerberg et al., 2011). Where Selection bias refers to the amount of bias in estimated regression coefficients when one selects only statistically significant predictors to be included into the model. In their response they illustrated that selection bias in estimated regression coefficients is larger than the bias that may occur in a prespecified model.

Courvoisier et al. (2011b)

Later, Courvoisier et al. (2011b) responded using results from a new simulation study. They found that bias at EPV less than 5 was higher whenever predictors were selected based on statistical significance. Thus, they confirmed that due to the impact of omitting nonsignificant variables, the estimation of significant variables effects was biased (Courvoisier et al., 2011b).

Ogundimu et al. (2016)

They conducted a resampling study to evaluate the effect of the EPV on the accuracy and precision of estimated regression coefficients and the accuracy of predictions for models with binary predictors which have low prevalence.

Ogundimu et al. (2016) used the data from The Health Improvement Network (THIN) which consisted of two million patients' information. They studied fully pre-

specified models and models derived using automated variable selection (backward elimination).

In their study, they used a case-control design for sampling (select events and non-events separately) and examined the number of variables in the model. Ogundimu et al. investigated the predictive accuracy of the models using the C index, D statistic, and measures of explained variation and calibration slope.

Ogundimu et al. (2016) concluded that a higher EPV (say, >20) is required when there are many low-prevalence binary predictors in the model; this is to ensure unbiased regression coefficients and accurate/improved predictions compared to true predictions from the true model which was developed using the entire THIN dataset.

Although Ogundimu et al. (2016) employed a large dataset for their study, they studied a limited number of scenarios (number of variables in the model). For instance, they did not examine collinearity between variables or different outcome prevalences. They also only evaluated the Cox regression model.

van Smeden et al. (2016)

They conducted Monte Carlo simulations to investigate small sample bias, coverage of 90% confidence intervals, and mean square error of regression coefficients.

van Smeden et al. (2016) show that the estimates of true associations were largely biased (about 30%) for data sets with small EPV, and bias may not disappear even for a large EPV (say, 150). They also reported the coverage above the nominal level for EPV less than 30, and that mean square error was large (greater than 0.2) for those predictors with large true effects (β equals to log(4) or log(0.25) when with an EPV ≤ 30.

van Smeden et al. (2016) found that apart from EPV, other factors, such as the total sample size, and true effect size, associated with the problems of low EPV. They concluded that the available evidence supporting EPV rules is weak. Nevertheless, they have not investigated the effect of EPV on the predictive performance of logistic regression models.

Jinks (2012)

In her PhD dissertation, Jinks conducted a study to explore the problem of sample size, EPV, in risk survival modelling. She used a number of different survival datasets

from cancer with various sample size and outcome prevalence. To quantify model performance she only used the discrimination measure (D statistic). In this study Jinks developed multivariable fractional polynomial models (Sauerbrei and Royston, 1999). These models use backward elimination as a variable selection method. They used the bootstrap method to minimise optimism in the estimate of the measure due to employing the same dataset for both developing and validating. This researcher studied the presence of noise variables in the data, the outcome prevalence, and sample sizes along with the EPV.

Jinks concluded that with EPV greater than and equal to 10 in a dataset, there is no chance of producing a valid risk model where the valid model was defined as the one that has an acceptable amount of optimism (median optimism was 18% here) in its related D statistic. Ultimately, she recommended that to minimise optimism and to accurately estimate the D statistic EPV should exceed 20 or 30.

Jinks exploited a variety of datasets (studying different case mixes) but only used survival data and evaluated her models with just the D statistic. Furthermore, Jinks used the multivariable fractional polynomial (Royston and Sauerbrei, 2008) method to develop risk models. This method uses a backward elimination procedure to select variables. A closer look at the literature reveals that this variable selection method is always criticised for the bias that it introduces and may not choose clinically important variables (Harrell, 2001, Steyerberg, 2009, Ambler et al., 2011).

EPV summary

So far, the review of the available literature has revealed that, with the exception of a very small part of Courvoisier et al. (2011a) and the study by Ogundimu et al. (2016), investigators have mainly paid attention to the effect of sample size using EPV on the estimation of regression coefficients. Nevertheless, in the case of a risk prediction model, there is less interest in individual covariate effects. Rather, the main focus is likely to be measuring the ability of the model to predict outcomes for future patients, or to discriminate between groups of patients. Copas (1983) asserted that "a good model may include variables which are 'not significant', exclude others which are, and may involve coefficients which are systematically biased". Hence, basing sample size decisions on the significance or unbiased estimation of model coefficients alone may not result in the best risk prediction model.

In practice, it may not be possible to accomplish the EPV $\geq$ 10 advice when developing a risk model under some circumstances (Ambler et al., 2011); for instance, datasets might contain few events due to rare event cases. As such, the performance of the risk model is affected negatively (Harrell, 2001). However, the predictive accuracy of the risk model may be improved by using the post-estimation shrinkage factor (Harrell, 2001, Ambler et al., 2002, Steyerberg, 2009, Ambler et al., 2011). The factor can be estimated using the bootstrap or the heuristic approach.

Ambler et al. (2011)

The use of shrinkage factors when developing risk models has been suggested for some decades (Copas, 1983, Van Houwelingen and Le Cessie, 1990). However, few studies have investigated the influence of applying shrinkage on the performance of risk prediction models with regard to EPV. Ambler et al. (2011) carried out a simulation study based on two real datasets (the penile cancer data for rare events and the mechanical failure of artificial heart valves data for rare disease) to investigate differences in the performance of three risk modelling methods (ridge (Verweij and van Houwelingen, 1994), lasso (Tibshirani, 1996), and garotte (Breiman, 1995)) and the standard Cox model with post-estimation shrinkage when there is a low EPV. In their penile cancer data, there was information on 129 patients (20 events), whilst the heart valve data consisted of 3118 patients (56 events). Now those methods and some details of their study are described.

The penalised estimation methods of ridge, the least absolute shrinkage and selection operator (lasso), and garotte maximise a function which is the sum of the usual Cox partial likelihood and a penalty term (a function of only the regression coefficients and a parameter that controls the amount of shrinkage). The difference between Ridge and Lasso is in the penalty term: in the Ridge method the penalty term is proportional to the sum of the squares of the regression coefficients, but in the lasso method it is proportional to the sum of the absolute value of the regression coefficients. The non-negative garotte individually shrinks each maximum likelihood regression coefficient under a constraint on the sum of the corresponding shrinkage parameters.

Ambler et al. observed that the method of maximum likelihood with linear shrinkage factor (LSF) and ridge were often under-fitted with decreasing EPV. Ambler et al. also learnt that garotte and lasso produced the best calibration. Furthermore, in terms of

the discrimination ability of the models, they found that ridge performed best in all EPV, followed by lasso and the full model. In conclusion, Ambler et al. suggested using some type of shrinkage for EPV<30 when developing risk models (Ambler et al., 2011).

Summary

To sum up, the consensus might be that at least ten EPV is a reasonable guideline to develop a reliable risk model to achieve acceptable accuracy and precision in the estimation of the regression coefficients (Harrell, 2001, Peduzzi et al., 1995, 1996). However, a researcher can develop a model with EPV greater than and equal to five with care (Vittinghoff and McCulloch, 2007) or may need a larger EPV (say, >20) when there are many low-prevalence binary variables in the model. Although these results are reassuring, the required number of events to develop a reliable risk prediction model might be different.

There is no consensus on how many EPV are needed to develop a reliable risk prediction model using binary outcomes. Courvoisier et al. (2011a) suggested that EPV≥10 can not guarantee the accurate estimation of the C statistic. Further, Jinks (2012) justified that EPV should be at least 20 or 30 to develop a valid risk prediction model.

This study is carried out to ascertain the required number of events to develop a reliable risk prediction model. To do this, we illustrate the dependence of the risk model performance on EPV in the case study. This will be conducted using several performance measures of discrimination, calibration, and overall performance. We also performed a simulation study evaluating various scenarios including different outcome prevalence in the dataset and the strength of linear prediction.

Moreover, Ambler et al. (2011) recommended the use of shrinkage technique when developing a risk model at EPV $\leq$ 30. Therefore, the trend of improvement in performance of the risk models when applying linear shrinkage with EPV is also studied in this chapter. Despite Ambler et al.'s finding showed that the postestimation linear shrinkage performed the worst at low EPV among all methods that they used, we will employ linear postestimation shrinkage in this chapter because we believe that their finding was only from two case mixes and may not be the case in our study.

3.2 Data

The data used in this dissertation, the heart valve surgery data, was from the Society of Cardiothoracic Surgeons of Great Britain and Ireland (SCTS). It was based on patients who underwent aortic and/or mitral heart valve surgery, both repair and replacement, between April 1995 and April 2003 (Keogh and Kinsman, 2003). The clinical outcome was in-hospital mortality, recoded as either dead or alive. This data set has been used by Ambler et al. (2005) to develop and validate a risk model. They developed the risk model on the dataset from the first five years (N=16,679) and evaluated its performance on the remaining data (N=16,160). This model was to predict in-hospital mortality for aortic and/or mitral heart valve patients with or without concomitant Coronary Artery Bypass Grafting (CABG). The overall in-hospital mortality was 6.4%.

In this chapter the first five years of data is utilised, but in Chapter four (sample size requirement to validate a reliable risk model using independent binary data) the entire data (N=32,839) is employed. Therefore, the data used in these two chapters is described in this section.

A set of predictors all with prognostic information were chosen to use in this study (except the case study in Chapter four where we used most of the predictors of Ambler et al.'s model to specify a perfect risk model), having a mixture of continuous (age at surgery, and body mass index or BMI), categorical (renal problem, ejection fraction, operative priority, operation sequence, preoperative arrhythmias, year of operation and valve operation), and binary (concomitant CABG surgery, diabetes and hypertension, sex) variables. Categories were combined for two variables, renal problem (high creatinine and dialysis) and operation sequence (second and third or more), to avoid perfect prediction problems (Albert and Anderson, 1984) due to low prevalence (less than 1%). Therefore, renal problem and operation sequence changed to be binary variables.

Logistic regressions models were fitted to the entire dataset (N=32,839) using maximum likelihood (ML) to investigate the importance of each predictor. For each predictor, two models were fitted, one all predictors including and one excluding the predictor. The decrease in χ^2 ($\Delta\chi^2$) was then calculated for each predictor (Table 3.1).

Operative priority was the most important predictor followed by age and renal failure. Table 3.1 also summarises the prevalence of each category for categorical variables.

3. SAMPLE SIZE REQUIREMENTS FOR DEVELOPING A RISK MODEL USING INDEPENDENT BINARY DATA

Table 3.1: Importance of the predictors in a multivariable model (estimated using maximum likelihood) for the heart valve data (N=32,839).

Variable	Category	D.F.	M (p_M)	$\hat{\beta}$ (SE)	OR (95% CI)	$\Delta\chi^2$
Operative priority	Elective	2	23,926 (0.05)		1.00	432.246
	Urgent		7,510 (0.09)	0.576 (0.053)	1.75 (1.58 , 1.94)	
	Emergency		1,403 (0.25)	1.680 (0.078)	5.33 (4.57 , 6.21)	
Age at surgery		1		0.035 0.002)	1.04 (1.03 , 1.04)	227.727
Age at surgery	< 50	4	3,559 (0.03)		1.00	217.896
	$50 - 59$		5,225 (0.04)	0.211 (0.122)	1.28 (1.01 , 1.63)	
	$60 - 69$		10,573 (0.06)	0.535 (0.109)	1.79 (1.44 , 2.22)	
	$70 - 79$		11,340 (0.08)	0.950 (0.107)	2.76 (2.23 , 3.41)	
	> 70		2,142 (0.10)	1.347 (0.126)	4.11 (3.21 , 5.27)	
Renal failure	No	1	31,121 (0.06)		1.00	153.243
	Cr>200&dialysis		1,718 (0.21)	0.942 (0.072)	2.12 (1.89 , 2.37)	
Operation sequence	First	1	28,673 (0.06)		1.00	140.657
	Second&more		4,166 (0.12)	0.751 (0.061)	1.87 (1.70 , 2.06)	
Ejection fraction	Good(>49)	2	21,816 (0.05)		1.00	106.855
	Fair(30-49)		8,783 (0.08)	0.218 (0.054)	1.24 (1.12 , 1.38)	
	Poor(<30)		2,240 (0.16)	0.776 (0.072)	2.19 (1.90 , 2.52)	
Concomitant CABG surgery	No	1	21,865 (0.05)		1.00	74.677
	Yes		10,974 (0.09)	0.443 (0.051)	1.57 (1.42 , 1.73)	
Valve Operation	Aortic	2	21,143 (0.06)		1.00	61.764
	Mitral		9,651 (0.07)	0.254 (0.054)	1.29 (1.16 , 1.43)	
	Aortic+mitral		2,045 (0.11)	0.642 (0.084)	1.90 (1.61 , 2.24)	
Preoperative arrhythmias	No	2	23,060 (0.04)		1.00	46.938
	AF/flutter		9,160 (0.10)	0.340 (0.051)	1.40 (1.27 , 1.55)	
	VT/VF		619 (0.09)	0.384 (0.150)	1.46 (1.08 , 1.96)	
BMI	Low(<20)	2	2,788 (0.11)		1.00	38.664
	Normal(20-25)		12,076 (0.08)	-0.350 (0.076)	0.71 (0.61 , 0.82)	
	High(>25)		17,975 (0.06)	-0.486 (0.076)	0.62 (0.54 , 0.72)	
Year	1995	8	1,039 (0.09)		1.00	22.597
	1996		1,867 (0.08)	-0.142 (0.147)	0.86 (0.65 , 1.15)	
	1997		2,882 (0.08)	-0.106 (0.137)	0.90 (0.68 , 1.17)	
	1998		3,986 (0.07)	-0.288 (0.134)	0.74 (0.57 , 0.97)	
	1999		5,359 (0.06)	-0.379 (0.131)	0.67 (0.52 , 0.87)	
	2000		5,053 (0.06)	-0.311 (0.132)	0.71 (0.55 , 0.92)	
	2001		5,178 (0.06)	-0.244 (0.132)	0.76 (0.59 , 0.98)	
	2002		5,950 (0.05)	-0.432 (0.133)	0.63 (0.48 , 0.81)	
	2003		1,525 (0.05)	-0.366 (0.167)	0.67 (0.48 , 0.93)	
Diabetes	No	1	15,189 (0.07)		1.00	16.628
	Yes		1,269 (0.11)	0.303 (0.072)	1.36 (1.18 , 1.56)	
Hypertension	No	1	20,866 (0.07)		1.00	12.528
	Yes		11,973 (0.08)	0.179 (0.050)	1.19 (1.08 , 1.32)	
Full model (with age)		24				$LR\chi^2$=1883.675
Full model (with age groups)		27				$LR\chi^2$=1873.843

D.F. denotes the number of parameters included in the model for that predictor.

M and p_M refers to the number of patients and the rate of in-hospital mortality in each category, where p_M=m/M and m is the number of patients with the event in each category.

$\hat{\beta}$ (SE) refer to the estimated regression coefficient and corresponding standard error

Odds Ratio (OR) and its related 95% confidence interval (CI) was estimated with either age or agegrp (age group) in the model.

$\Delta\chi^2$ denotes the decrease in the χ^2 statistic for the model when the predictor is omitted and the model refitted.

3.3 Case study

From reviewing the literature it become clear that the available sample size advice to develop a risk prediction model is not sufficient. There is little advice regarding the prediction ability of a risk model when the outcome of interest is binary, in contrast to the advice regarding the accuracy of the effect size of predictors.

Therefore, a case study was conducted using the heart valve surgery data (described in 3.1) to understand how the accuracy of predictions produced using logistic regression models can be affected by the EPV.

3.3.1 Method

The following method was used to carry out the case study. The data was randomly split into two parts for development (80%) and validation (20%). Then, the required subset of the development sample was taken for each EPV (by separate sampling from events and nonevents). A standard logistic regression model was fitted on the development sample and its performance was quantified in the correspondent validation sample (using four measures). These measures were the C statistic, D statistic, calibration slope, and Brier score. The process was repeated several times (i.e. 200) for each EPV. The EPV values were 2.5, 5, 7.5, 10, 12.5, 15, 20, 25 and 30. These values include all recommended EPV (5, 10, 20, and 30).

In all simulation and case studies for the entire thesis we used the standard error of the estimated measures to determine the adequate number of simulations to perform. That is, we obtained the required number of simulations using $\text{SE} = \text{SD}/\sqrt{n}$, where SE and SD are standard error and standard deviation, respectively. For instance, to achieve a standard error of 0.001 for the estimated mean we require 400 simulations based on a standard deviation of 0.02.

3.3.2 Results

Table 3.2 displays the mean values of each performance measure over 200 samples for each EPV. From the table, it is clear that the performance of the risk model is dependant on the EPV.

The discrimination ability of the models deteriorated by decreasing the EPV for the C statistic and D statistic. That is, the mean value of the C statistic diminished from

Table 3.2: mean values of each performance measures over 200 samples for each EPV level.

EPV	C statistic (SE*)	D- statistic (SE)	Calibration slope (SE)	Brier score (SE)
30	0.74 (0.02)	1.47 (0.12)	0.96 (0.01)	0.06056 (0.0030)
25	0.74 (0.01)	1.48 (0.11)	0.96 (0.01)	0.06119 (0.0030)
20	0.73 (0.02)	1.45 (0.12)	0.94 (0.01)	0.06115 (0.0029)
15	0.73 (0.01)	1.45 (0.11)	0.92 (0.01)	0.06149 (0.0031)
12.5	0.73 (0.02)	1.41 (0.11)	0.89 (0.01)	0.06162 (0.0033)
10	0.72 (0.02)	1.38 (0.12)	0.88 (0.01)	0.06192 (0.0031)
7.5	0.72 (0.02)	1.39 (0.13)	0.84 (0.01)	0.06155 (0.0031)
5	0.72 (0.02)	1.32 (0.16)	0.77 (0.01)	0.06238 (0.0032)
2.5	0.69 (0.03)	1.15 (0.19)	0.60 (0.01)	0.06452 (0.0041)

* SE denotes the empirical standard error of the measure in 200 simulations.

0.74 for EPV equal to 30 to 0.69 for EPV equal to 2.5, respectively. Also, the value of the D statistic declined from 1.47 for EPV equal to 30 to 1.15 for EPV equal to 2.5, respectively. The major drop was at EPV=2.5 for the C statistic. It was at EPV ≤ 5 for the D statistic.

Furthermore, there were signs of overfitting for all EPV as shown by the calibration slope. Moving from EPV of 30 to 2.5, the calibration slope changed from 0.99 to 0.60. However, models fitted using EPV ≤ 12.5 were overfitted. This problem was severe for EPV ≤ 5

The overall performance of the models measured by the Brier score slowly decreased with regard to the decreasing EPV. There is a jump at EPV ≤ 5. Brier scores for different EPV are so close together due to the low outcome prevalence in our dataset and also our scaling.

3.3.3 Discussion

From the results of this case study, it is evident that the performance of the risk model was affected by the EPV.

Further investigation was conducted to understand the reasons for the trends seen in Table 3.2. The effect of the EPV on the separation between $\hat{\eta}^{(0)}$ and $\hat{\eta}^{(1)}$ is illustrated in Figure 3.1 where $\hat{\eta}^{(0)}$ and $\hat{\eta}^{(1)}$ are the linear predictors that correspond to the nonevent and event groups respectively. For each panel (each EPV), an overlap of all 200 $\hat{\eta}^{(0)}$

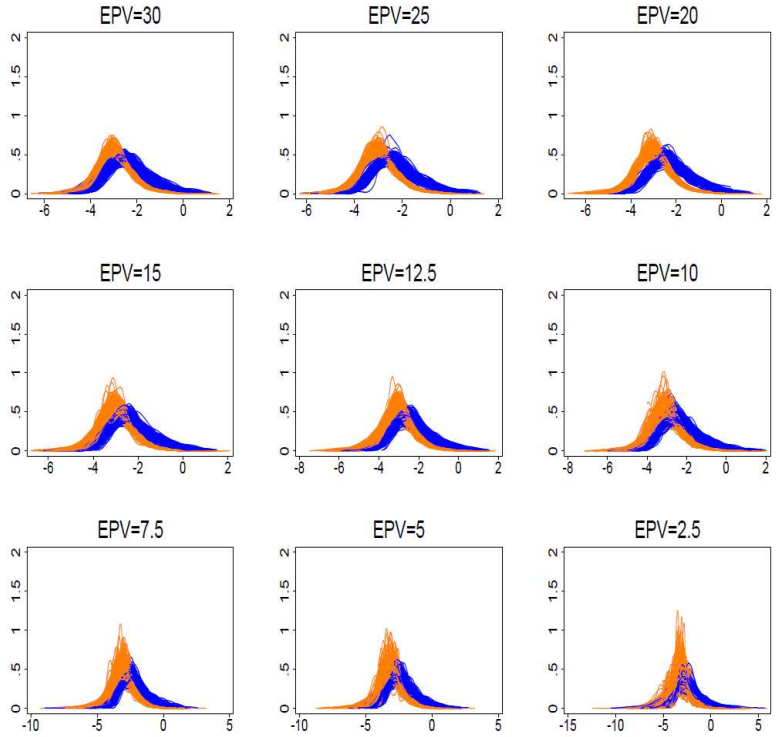


Figure 3.1: Separation between $\hat{\eta}^{(0)}$ and $\hat{\eta}^{(1)}$ by EPV. $\hat{\eta}^{(0)}$ and $\hat{\eta}^{(1)}$ are linear predictors correspond to nonevent and event groups, respectively.

and $\hat{\eta}^{(1)}$ distribution have been overlaid to judge the separation by decreasing EPV. As can be seen, the separation between $\hat{\eta}^{(0)}$ and $\hat{\eta}^{(1)}$ is better when the sample size (EPV) is large compared to when the sample size is small.

This indeed was the pattern observed in the D statistic and C statistic, both measures decreased by decreasing the EPV.

The questions that we also like to address in this chapter are whether EPV is the only factor that affects model performance or whether there are other additional factors. Therefore, we need to change some characteristics of the data and observe whether model performance changes when other factors in the dataset apart from EPV are varying. That will bring us to the next section in which we discuss the quality of

the risk models that were studied using diverse scenarios and EPV.

3.4 Simulation study

The case study illustrates that model performance is influenced by EPV in terms of calibration, discrimination, and overall performance. However, it is still unclear whether performance is affected by other factors such as outcome prevalence, strength of the model or presence of noise variables in the model.

To learn how model performance and predictive accuracy depend on EPV and other factors, a number of simulation studies were performed based on the heart valve surgery data. The following factors were investigated; EPV, prevalence of the event of interest, prognostic strength of the risk model, the presence or otherwise of noise or continuous variables in the data, and the degree of collinearity. The details of these simulations are now described.

3.4.1 Overview

For the dataset described in section 3.2, a true model was derived using the entire dataset (N=16,679) and a full model was specified. From the true model, a set of true regression coefficients were obtained and used to generate new responses. These were used to investigate the following factors:

§ *Strength of risk model*: to study whether EPV requirements change if the prognostic strength of the risk model varies, three levels of model strength were specified; weak, original, and strong. 'Original' refers to the models in which the coefficients are obtained from the model fitted to the heart valve surgery data.

§ *Outcome prevalence*: to assess how prevalence of the outcome can change the EPV requirements, three outcome prevalences were examined; 7%, 25% and 40%. These represent situations where the occurrence of the event of interest is low, medium or high. Note that the first level is the outcome prevalence in the heart valve surgery data (7%).

§ *Noise variables*: to learn whether the presence of noise variables in the risk model can affect the EPV requirements, a mixture of continuous and binary noise variables were simulated and included in the risk models. There were either 0, 5, or 10 noise variables.

§ *Type of predictor* : to understand whether the EPV requirements are different for models with different types of predictors, type of predictors in the model was changed;

variables were either all categorical, or a mixture of both categorical and continuous with one type dominant. Four levels were considered for this scenario; 0, 2, 4 or 7 continuous predictors out of 11 predictors.

§ *Number of predictors*: to check whether the EPV requirement should vary depending on the number of predictors in the model, different numbers of predictors were considered to be in the model. Four levels were studied for this scenario; 4 or 7, 10 or 22 predictors. These mimicked the situations whereby there were a small (4 or 7 variables), medium (10 variables) or large (22 variables) number of predictors in the risk model.

§ *Degree of collinearity*: this factor is to understand the impact of collinearity in data on the EPV requirement. The average (pairwise) Pearson correlation in the real data was 4%. Hence, two extra datasets (with about 45% and 70% (pairwise) correlation) were produced to assess this scenario.

§ *Postestimation shrinkage*: to learn how EPV requirements can change when postestimation shrinkage is to be applied, two postestimation linear shrinkage methods were used; namely, heuristic shrinkage and bootstrap. The former method is simple and can be calculated more easily than the latter (see section 2.5). In contrast, the second method is recommended but may take a long time to obtain if, for example, there is high collinearity between variables (Harrell, 2001).

§ *EPV*: nine EPV were considered: 2.5, 5, 7.5, 10, 12.5, 15, 20, 25, and 30 . This range includes all recommended EPV (see section 3.1).

For each combination of factors, 400 datasets were simulated. For each EPV, the simulated dataset was randomly split into two parts (development (80%) and validation (20%)) and the size of the development data was altered (by sampling separately from events and nonevents), but the size of the validation data remained unchanged. The risk model was fitted on the development data and the performance of it was quantified using the validation data. The predictive accuracy of each risk model was assessed by comparing the estimated performance measures with reference measures. These reference measures represent the performance of the risk models fitted on the full size development dataset. The following measures were used:

§ calibration - calibration slope,

§ discrimination - C statistic, D statistic,

§ overall performance - Brier score

Only 'full' models were used - no variable selection method was applied.

Evaluating the models

To enable comparison across different simulated scenarios, model performance was quantified relative to the corresponding reference values using

$$\text{Relative differences} = (\frac{\widehat{m}_i - m}{m}) \times 100$$

where $\widehat{m}_i$ and m are the performance measure from the ith simulated data and the reference measure.

Sample size calculation

EPV is defined as:

$$\text{EPV} = \frac{\text{(number of events)}}{\text{(number of potential variables)}} \tag{3.4.1}$$

In this study, the full model approach was taken; therefore, the number of potential variables is the number of variables in the model.

This can be rewritten as:

$$\text{EPV} = \frac{\text{(N)} \times \text{(outcome prevalence)}}{\text{(number of variable)}} \tag{3.4.2}$$

and so:

$$\text{N} = \frac{\text{EPV} \times \text{(number of variables)}}{\text{(outcome prevalence)}}. \tag{3.4.3}$$

The sample sizes required for different EPV when ten variables are in the model are given in Table 3.3.

Details of each simulation scenario is described below along with the corresponding results.

3.4.2 Model strength

Generally, the strength is evaluated by measuring how well that model can separate between prognostic groups. A strong model has the ability to classify patients into clinically useful groups (Altman and Royston, 2000). The required EPV is expected to be lower when developing models with several strong predictors. To aid better understanding of this, consider two scenarios in which the true logistic model consists

Table 3.3: Sample sizes required for each EPV and by outcome prevalence when there are ten predictor variables.

		Outcome prevalence		
		7%	25%	40%
EPV	2.5	357	100	63
	5	714	200	125
	10	1429	400	250
	12.5	1786	500	313
	15	2143	600	375
	20	2857	800	500
	25	3571	1000	625
	30	4286	1200	750

of a single binary predictor where the corresponding β equals either 2 or 0.5. Further, consider the prevalence of the binary outcome to be 10%. The C statistic for the model with the strong predictor is 0.73 and for the model with the weak predictor it is 0.56. To develop a model with this strong predictor (β=2) an acceptable C statistic (say, 0.72) can be obtained using only 10 events per variable. However, a model with the weak predictor (β=0.5) will always have low discrimination ability even when EPV exceeds 50. In other words, EPV needs to be large when developing a risk model with a weak predictor.

The effect of EPV on the performance of models of varying strength was studied. To do this, the logistic model was fitted on the entire data; the true linear predictor (η_{true}) was obtained and used to derive the new linear predictor (η_{new}). That was conducted using $\eta_{new} = A + B \times (\eta_{true} - \bar{\eta}_{true}) + \bar{\eta}_{true}$, where A sets the outcome prevalence and B shrinks or boosts the prognostic strength. The values of B were chosen to be 0.5, 1, and 2 to achieve the following desired values for the C statistic (0.61, 0.74, and 0.87, respectively) as an indicator of the strength of the model. Since a prognostically strong model can accurately differentiate patients into different risk groups and a larger C means that the model has a good discriminating ability, our method seems adequate. The original outcome prevalence (7%) is used for all three strength levels for comparability. For convenience the models corresponding to the C statistic values of 0.74, 0.61 and 0.87 will be called original, weak and strong models,

Table 3.4: Reference values of performance measures for all scenarios

Scenario	Scenario level	Calibration slope	Brier score	C statistic	D statistic
Model strength	weak	1.00	0.064	0.62	0.70
	original	1.00	0.059	0.73	1.47
	strong	1.00	0.046	0.87	3.15
Outcome prevalence	7%	1.00	0.059	0.73	1.47
	25%	1.00	0.177	0.64	0.86
	40%	1.00	0.228	0.62	0.75
Noise variables	0/5/10	1.00	0.059	0.73	1.47
Type of predictors (number of continuous variables)	0	1.00	0.059	0.73	1.52
	2	1.00	0.059	0.74	1.55
	4	1.00	0.060	0.73	1.52
	7	1.00	0.061	0.74	1.54
Number of variables in the model	4	1.00	0.062	0.72	1.42
	7	1.00	0.061	0.72	1.44
	10	1.00	0.059	0.73	1.47
	22	1.00	0.059	0.73	1.46
Presence or otherwise of collinearity	4%	1.00	0.059	0.73	1.47
	46%	0.99	0.059	0.72	1.44
	71%	1.00	0.058	0.72	1.40

respectively.

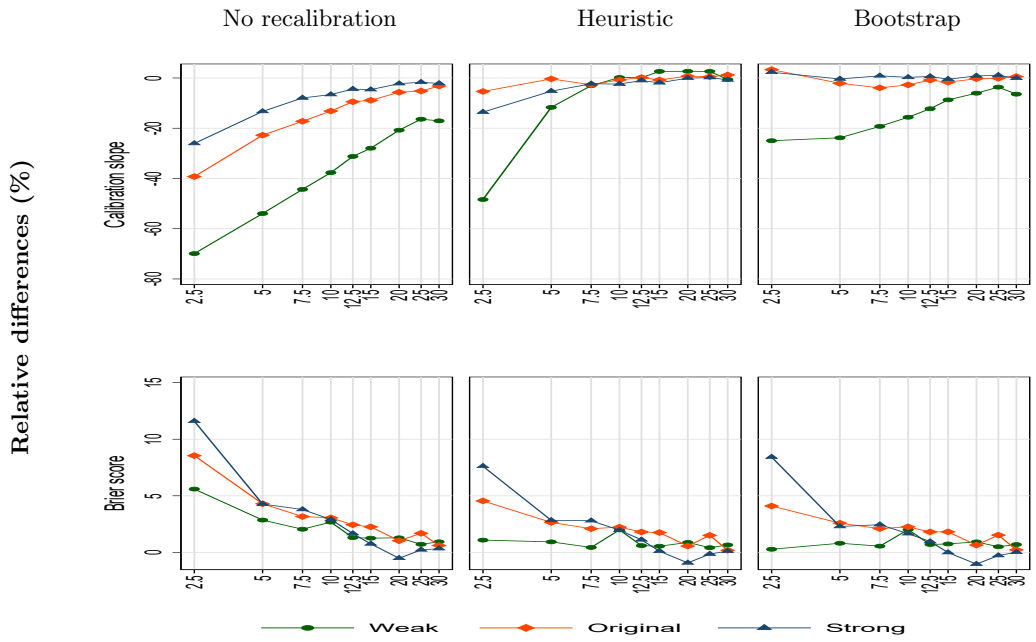
Note that the model χ^2 for the true original model was 857.1 and the interquartile range (IQR) for the weak and strong models χ^2 were (193 to 233) and (2664 to 2792), respectively.

Model strength: results

Figure 3.2 displays relative differences in the calibration slope and Brier score by model strength over EPV and Table 3.4 presents reference values for the various model strengths. From the plot, the relative differences in all measures were smallest for strong models, and largest for weak models for all EPV.

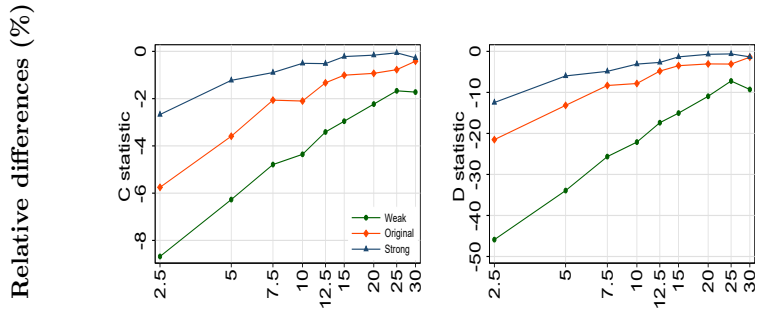
From the calibration slope plots, there was an overfitting problem in models of any strength at almost all EPV if no recalibration was applied. However, the overfitting issue was a greater concern in all EPV for weak models than strong and original models. Such models were overfitted by more than 10% at EPV $\leq$7.5 and EPV$\leq$ 10, respectively.

The overfitting issue was either resolved or alleviated by applying postestimation



The simulation standard error (SE) among all EPV, model strength and the recalibration status of models for the estimated calibration slope and Brier score were between (0.003, 0.041) and (0.000133, 0.000135), respectively.

Figure 3.2: Relative differences in the calibration slope and Brier score over EPV by strength of risk model based on 400 simulated datasets. The X axis is on the log scale.



The simulation standard error (SE) among all EPV, model strength and the recalibration status of models for the estimated C statistic and D statistic were between (0.0018, 0.0021) and (0.017, 0.018), respectively.

Figure 3.3: Relative differences in the C statistic and D statistic over EPV by strength of risk model based on 400 simulated datasets. The X axis is on the log scale.

linear shrinkage for all EPV for all models of any strength (Figure 3.2, top row). While the application of bootstrap postestimation factor removed entirely differences in strong and original models at all EPV. However, it only diminished the differences in weak models at EPV$\geq$ 15. On the other hand, by applying the Heuristic postestimation factor the differences almost completely vanished in the strong and original models and the calibration slope dramatically improved in weak models, although there were some differences at EPV= 2.5 with strong and original models and large differences at EPV$\leq$ 5 with weak models.

Looking at the Brier score plot, there was a downward trend in the relative differences by increasing EPV. It was steeper for strong models than for the original and weak models, and for models with no recalibration, compared to those in which linear postestimation shrinkage was applied. The relative differences in the Brier score were less than 4% at EPV $\geq$ 10 for all models with no recalibration and those in which linear postestimation shrinkage was applied.

Figure 3.3 displays the relative differences in the C statistic and D statistic by model strength over EPV and Table 3.4 presents the corresponding reference values for various model strengths. From the plots, we can see that the relative differences were the smallest for the strong models and the largest for the weak models by deceasing EPV. The relative differences for the C statistic were less than 2% of the reference value at EPV $\geq$ 7.5 and EPV $\geq$ 10 for strong and original models, respectively. However, the relative differences were always larger than 2% for weak models. The relative differences for the D statistic were less than 10% of the reference value at EPV$\geq$ 5, EPV $\geq$ 7.5, and EPV $\geq$ 15 for strong, original models and weak models, respectively. Note that as applying postestimation shrinkage does not change the risk order, and so C statistic and D statistic. Thus, the corresponding graphs of C statistic and D statistic were not presented.

The results of studying this scenario imply that researchers should gather information from the related literature about their predictors and the strength of their relation with the outcome of interest prior to their study. If they know that some (or one) of the predictors are strong (β is large or the odds ratio is significantly different than one), they can be assured that a sample size as small as EPV=10 may result in a reliable prediction model. However, if they know that all or most of the predictors are weak,

they should consider recalibrating their model even at EPV=30 as recommended by Ambler et al. (2011) and Steyerberg et al. (2001).

Furthermore, we note that the bias in the estimated regression coefficients decreased by increasing EPV, and bias was largest for the weak models and smallest for the strong models at all EPV (Figure A.1). As the linear predictor is the linear combination of estimated regression coefficients, a bias in the estimated regression coefficients causes an inaccurate and precise estimation for the performance measures. This explains the pattern seen in Figures 3.2 and 3.3.

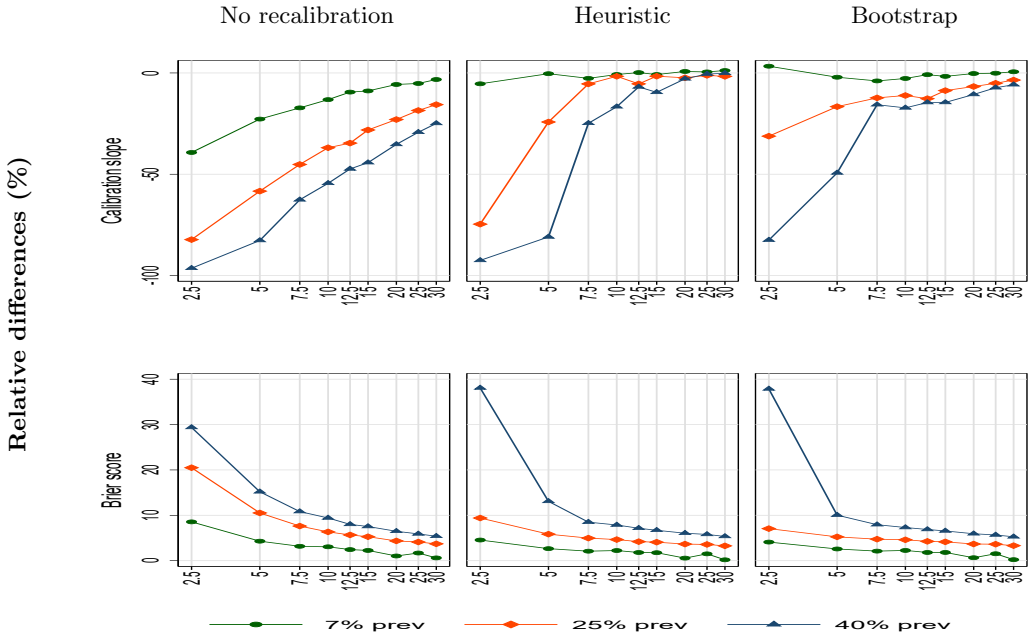
3.4.3 Outcome prevalence

The occurrence of the desired event can change from very rare to very frequent. Where the event of interest refers to endpoint such as death, heart disease, strokes, or the diagnosis of cancer. This can affect the amount of available prognostic information (Altman and Royston, 2000), where prognostic information relates to the spread of predicted probabilities. When the occurrence of the outcome of interest is relatively high (say, 40%), then a low EPV means that the total sample size is small.

The effect of EPV on the performance of risk prediction models fitted on three datasets with various outcome prevalences was studied. To simulate outcomes with different prevalences, the intercept of the true model was altered to produce new levels of outcome prevalence. The strength of all models with a different intercept was kept at similar levels. The transformed linear predictor was $\eta_{new} = A + B \times (\eta_{true} - \bar{\eta}_{true}) + \bar{\eta}_{true}$, where A and B were used to retain the outcome prevalence and strength of the model. Taking the described approach, two other series of datasets were generated, with 25% and 40% outcome prevalences.

Note that the data used throughout my thesis is from a prospective cohort study and the correct Epidemiological term for the occurrence of (new) events in this context is "incidence"hence rather than "outcome prevalence". However, we use the term "outcome prevalence" because we mean the occurrence of the event of interest in general.

Note that, as before, the model χ^2 for the true model fitted using data with the original outcome prevalence was 857.1. A simulation study was conducted to check the strength of the models. The IQR of the models' χ^2 for those fitted on datasets with 25% and 40% outcome prevalences over 400 simulated datasets were (819 to 901) and (828 to 890), respectively.



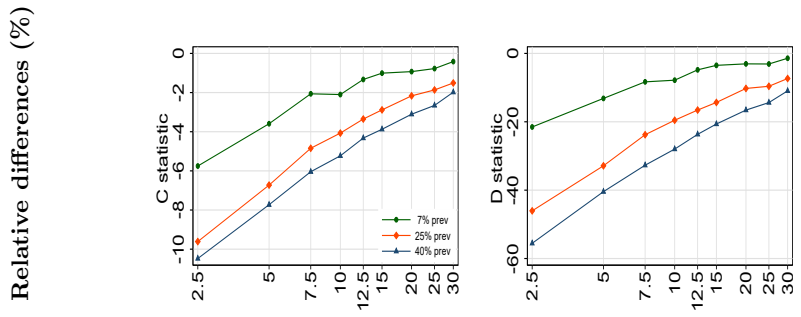
The simulation standard error (SE) among all EPV, outcome prevalence and the recalibration status of models for the estimated calibration slope and Brier score were between (0.006, 0.023) and (0.001, 0.002), respectively.

Figure 3.4: Relative differences in the calibration slope and Brier score over EPV by outcome prevalence based on 400 simulated datasets. The X axis is on the log scale.

Outcome prevalence: results

Figure 3.4 illustrates the relative differences in the calibration slope and Brier score by outcome prevalence and Table 3.4 presents the corresponding reference values by outcome prevalence.

A main point in this scenario is that as outcome prevalence increased the sample size decreased as the example in Table 3.3 demonstrates. Also, the parameter convergence problem (very large estimations for coefficients) (Heinze and Schemper, 2002) took place more in the higher outcome prevalences as a result of the smaller sample sizes (see Table 3.3). These problems occurred in 44% of simulations for all EPV (mostly EPV$\leq$ 10) and high prevalence (40%) compared to 3% for a very low EPV (2.5) when the prevalence was much lower (7%). These problems persisted up to when EPV was 30 but dramatically decreased by increasing the EPV for high prevalence (40%); in contrast, the problems mostly vanished when the EPV increased to 5 when the prevalence was



The simulation standard error (SE) among all EPV, outcome prevalence and the recalibration status of models for the estimated C statistic and D statistic were between (0.0008, 0.0011) and (0.006, 0.008), respectively.

Figure 3.5: Relative differences in the C statistic and D statistic over EPV by outcome prevalences based on 400 simulated datasets. The X axis is on the log scale.

low (7%). Results which led to a negative calibration slope were excluded from the analysis (1.7% of the entire results of the outcome prevalence).

From the matrix plot we can see that the smallest differences at all EPV were at the lowest outcome prevalence and the largest differences were at the highest outcome prevalence.

From the calibration slope plot, the values of the calibration slope were underestimated by more than 10% at EPV$\leq$10 and outcome prevalence of 7%. At the outcome prevalence of 25% and 40% differences were more than 10% at all EPV.

From the Brier score plot, the relative differences were less than 5% at EPV$\geq$ 5 and EPV$\geq$ 20 at 7% and 25% outcome prevalences, respectively. Relative differences were never 5% or less at the 40% outcome prevalence. Furthermore, the relative differences dropped at almost all EPV for the 7% outcome prevalences and at 12.5 EPV to less than 5% for outcome prevalences of 25%, when the heuristic shrinkage was applied. The application of this shrinkage made differences larger at EPV=2.5 and had almost no effect at EPV $\geq$ 5 when the outcome prevalence was 40%. The decrease in relative differences was slightly more with applying bootstrap shrinkage for all EPV at outcome prevalence of 25% especially for EPV of 5 at the outcome prevalence of 40% (5% decrease) compared to applying heuristic shrinkage.

From the C statistic plot, the relative differences declined at all outcome prevalences by increasing EPV. That decrease was steeper at the outcome prevalence of 40% than

at outcome prevalences of 25% and 7%. The relative differences were less than 2% at EPV$\geq$ 7.5, EPV$\geq$ 20, and EPV= 30 at 7%, 25%, and 40% outcome prevalences, respectively.

From the D statistic plot, the pattern of decreasing relative differences was similar to the C statistic. The relative differences were $<$ 10% at EPV$\geq$ 7.5, EPV $\geq$20 and EPV =30 in 7%, 25%, and 40% outcome prevalence, respectively.

From the results of the two previous scenarios, we note that the C statistic and D statistic are correlated. The correlation coefficient was about 0.98. Hence, only results for C statistic (the most common measure) will be presented for the rest of the scenarios.

In brief, from the results of the simulation for the scenario of outcome prevalence, it appears that the EPV needs to be at least 12.5 to develop a risk model. At this EPV the relative differences in estimated performance measures were acceptable. However, a shrinkage might be needed for large outcome prevalences.

Furthermore, from Figure A.2 we note that the bias in the estimated regression coefficients decreased by increasing EPV, and bias was largest when the outcome prevalence was large and smallest when the outcome prevalence was small at all EPV. Moreover, the linear predictor is the linear combination of estimated regression coefficients. That is, the bias in the estimated regression coefficients causes an inaccurate and precise estimation for the performance measures. This explains the pattern seen in Figures 3.4 and 3.5.

3.4.4 Presence or otherwise of noise variables

When model building, there might be some variables in the study sample which show a strong association with the outcome yet their true correlation is zero ($\beta_{true} = 0$) (Flack and Chang, 1987). This is more likely to happen when the sample size is small.

Now, let us postulate that there is already a dataset which includes some noise predictors along with the main predictors, variables which have a real relationship with the response, and that all predictors in this sample have an association with the response. Also let us assume that the model maker wishes to develop a risk model using all predictors in the sample without using variable selection methods. Now, the question is how much data is required to develop a reliable risk model using all variables? or, does the performance of a model with some noise variables change in

datasets of varying size? To address these issues we investigated the impact of 'noise' variables on the performance of a risk model and the effect such variables have on sample size requirements.

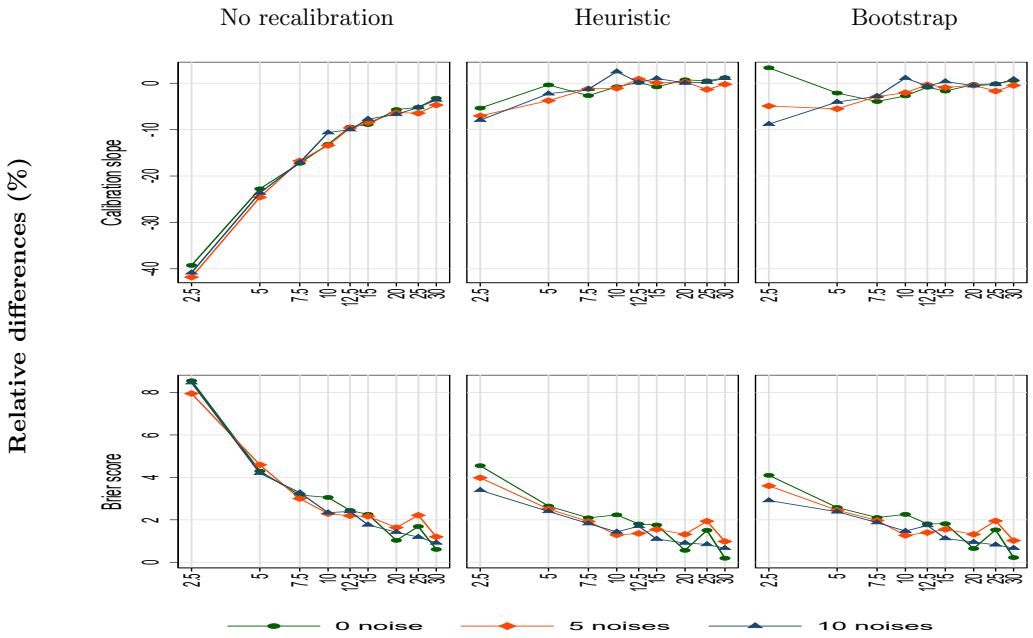
From the heart valve data, the seven predictors were chosen which have the biggest $\Delta\chi^2$, where $\Delta\chi^2$ is the decrease in the χ^2 statistic for the full model when the predictor is omitted and the model refitted. Five and ten additional noise variables were considered. The five noise variables consisted of three binary variables generated using probabilities of 0.1, 0.3, and 0.5 respectively, and two standard normal variables. The ten noise variables consisted of six binary variables generated using probabilities of 0.1, 0.2, 0.3, 0.4, 0.4, and 0.5 respectively, and four standard normal variables. These noise variables were generated completely independently from the outcome. Three risk models were developed, the first model had just the seven original predictors and the second and third models had the additional five and ten noises respectively along with those original predictors. Note that, in the simulation process, the size of the data was adjusted according to the number of noise variables in the data to retain the required EPV. For example, having five noise variables in the model with ten influential variables required us to pick about 38 events for simulated data to achieve the EPV=2.5, the simulated datasets for 0 noise model had 25 events.

Presence or otherwise of noise variables: results

This section illustrates relationships between bias in performance measures and EPV in models with a varying number of noise variables.

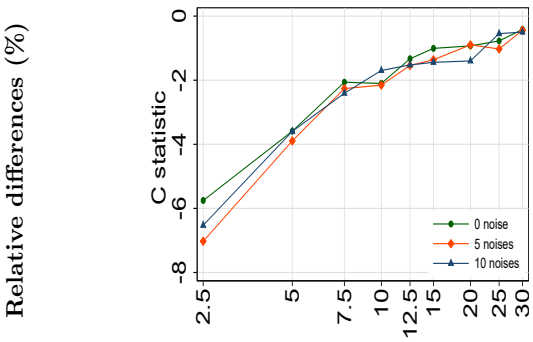
Figures 3.6 and 3.7 present the relative differences in the measures of performance quality over EPV across the number of noise variables and Table 3.4 presents reference values across various number of noises and modelling strategies. From the plot, the pattern of changes in measures of models with a large or medium number of noises were in a similar fashion to models with no noise variables.

In conclusion, noise variables did not affect the performance of the risk model for fixed EPV if we adjusted the sample size to achieve the target EPV. However, the presence of noise variables in the models required a larger sample size. If we had not increased the sample size then the presence of noise predictors would have reduced the EPV and would have resulted in a worse model.



The simulation standard error (SE) among all EPV, the presence of noise variables and the recalibration status of models for the estimated calibration slope and Brier score were between (0.002, 0.015) and (0.00006, 0.00007), respectively.

Figure 3.6: Relative differences in the calibration slope and Brier score over EPV by the presence of noise variables based on 400 simulated datasets. The X axis is on the log scale.



The simulation standard error (SE) among all EPV, the presence of noise variables and the recalibration status of models for the estimated C statistic was between (0.0003, 0.0005), respectively.

Figure 3.7: Relative differences in the C statistic over EPV by the presence of noise variables based on 400 simulated datasets. The X axis is on the log scale.

The results of this scenario suggest that a researcher needs to make a thorough review of literature about the research questions, the outcome variables, and the important predictors before commencing the study. The researcher also needs to exclude all variables which are 'speculative' in the model building stage in order to have enough EPV to develop a valid risk model where noise variables are considered to be speculative.

3.4.5 Type of predictors

In practice, medical researchers often convert continuous variables to categorical. That is because precise measurement is not always possible and it is easier for them to interpret the results of analysis (Harrell, 2001). However, categorising continuous variables may reduce the amount of information that the predictor holds.

A simulation study was conducted to investigate this issue. For this simulation study, a population of 100,000 observations was generated. The data for the population was generated according to a true logistic regression model. The true model included a continuous predictor (X_1) from a standard normal distribution with $\beta = 2$. The probability of an event (p_i) for individual i was computed using the inverse logit transformation, $p_i = \frac{1}{\alpha+\beta x_i}$. The intercept, α, was set to -3.4 in order to obtain an outcome prevalence of 10%. Then, two other predictors were produced by splitting X_1 into four (predictor X_2) and two (predictor X_3). For each observation, 1000 outcome Y_i was generated using Bernoulli distribution with probability p_i (the average outcome prevalence for the categories of X_2 were 0.33%, 5.13%, 29.24%, and 65.30% and for the categories of X_3 were 5.46%, and 94.54%). Each time, three logistic regression models were fitted using the response variable and one predictor: one with continuous predictor X_1, one with dichotomised variable X_2, and another with dichotomised variable X_3. For the model with the continuous variable X_1, the average C statistic was 0.88 compared to 0.85 and 0.75 for the model with X_2 (categorised variable with four categories) and for the model with X_3 (dichotomised variable with two categories), respectively. This implies that categorising a continuous variable may reduce the amount of prognostic information it holds.

Given the fact discussed thereof, the question is whether one needs different EPV when developing a risk model with categorical predictors comparing to when developing a risk model with continuous predictors. Hence, the heart valve surgery data was used

to address the issue. Some variables from data were transformed: some continuous predictors were categorised, and some categorical predictors were transformed into continuous predictors.

The following simulation study was carried out to investigate the above issue. To simulate this scenario, eleven predictors were chosen to be included in the model from the real data. Those were operative priority, operation sequence, valve operation, concomitant CABG surgery, age, BMI, renal failure, ejection fraction, diabetes, preoperative arrhythmias and hypertension, (see section 3.2). The first four were originally categorical. Age and BMI were recorded as continuous in the dataset, but most of the time researchers collect them as categorical. Hence, we categorised them according to common practice. The rest of the variables (renal failure, ejection fraction, diabetes, preoperative arrhythmias, and hypertension) were recorded as categorical in the dataset, but most of them are based on continuous measurements, such as ejection fraction. Therefore, the following approach was taken to produce the continuous versions of those variables.

To transform categorical predictors to continuous, a uniform noise variable of $[-0.5, 0.5]$ was added to each categorical predictor in order to jitter it across the reasonable range such that the categories of the variables did not overlap. Then, the jittered data was mapped to standard Normal distribution via order statistics for each categorical predictor. That is, $x_i = \Phi^{-1}(\frac{i}{N+1})$ where x_i, N and $\Phi^{-1}(.)$ are ith observation of the variable, number of observations and inverse standard Normal distribution function. These transformed predictors were considered to be continuous versions of the categorical predictors.

For example, consider a binary variable that is coded as 0 or 1. To jitter it, a uniform noise of [-0.5, 0.5] was added to each value, hence, the categories coded 0 and coded 1 will have values between [-0.5,0.5] and [0.5, 1.5], respectively. Then, these values were mapped to a standard Normal distribution via order statistics for this binary variable.

Subsequent to forming a continuous type of categorical variables, four models were produced using the various types of predictors: model with two continuous predictors and nine categorical predictors; model with four continuous predictors and seven categorical predictors, model with seven continuous predictors and four categorical predictors; model with 11 categorical predictors.

Table 3.5: Statistical significance of each predictor in a multivariable model fitted (using ML) on the heart valve dataset. $\Delta\chi^2$ denotes the decrease in the likelihood ratio χ^2 statistic for the model when the predictor is omitted and the model refitted. cont. means that variable is continuous.

Variables	D.F.	$\Delta\chi^2$
Age (cont.)	1	105.53
Age	4	104.55
Renal failure (cont.)	1	45.79
Renal failure	1	107.69
Ejection fraction (cont.)	1	30.19
Ejection fraction	2	41.62
Diabetes (cont.)	1	0.41
Diabetes	1	5.38
Preoperative arrhythmias (cont.)	1	29.78
Preoperative arrhythmias	2	47.99
Hypertension (cont.)	1	6.43
Hypertension	1	9.12
BMI (cont.)	1	8.9
BMI	2	16.97

Table 3.5 exhibits the effect of each form of predictors on the full model. As it can be seen from the table, transforming some of the variables to continuous resulted in a loss of information. That is the associated $\Delta\chi^2$ reduced. Therefore, the resulting models had different strengths and consequently were not comparable. The following describes how the strength of these models were equalised.

To equalise the strength of the models the model with all the original predictors (a model with two continuous predictors) was treated as the reference model. For the rest of the models, which had different strengths to the reference, prognostic indexes were boosted and then transformed in order to generate new outcomes and models refitted with new outcomes. This was conducted in the same fashion as it was performed for the model strength scenario or outcome prevalence.

Note that, as before, the model χ^2 for the true model fitted including two continuous variables was 857.1 and the interquartile range (IQR) of the models' χ^2 for those fitted including 0, four and seven variables were (888 to 968), (889, 978), and (895 to 974), respectively. Therefore, the models are now of comparable strength.

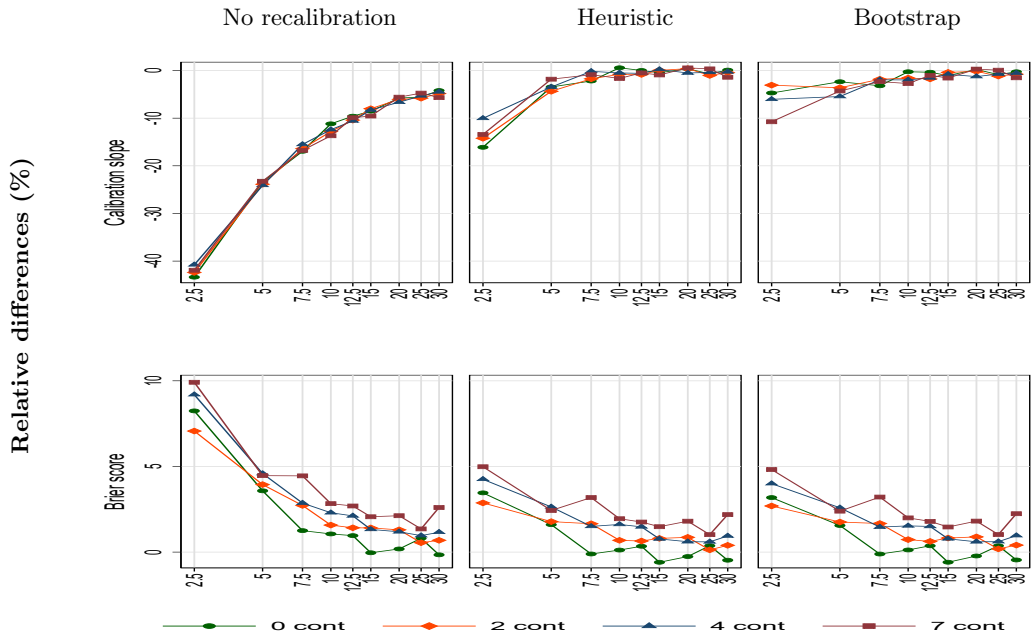
Type of predictors: results

This section illustrates relationships between performance measures and EPV in models with various type variables.

Figures 3.8 and 3.9 display relative differences in performance measures over EPV across the number of continuous variables, and Table 3.4 presents reference values across the types of variables and modelling strategies. As the figure demonstrates, there was little differences between the performance of models with different number of continuous predictors, holding EPV constant.

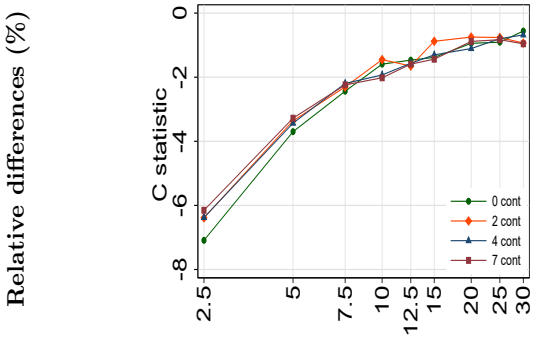
3.4.6 Number of variables in the model

Many researchers believe that if they collect as many variables that budget and time allow and include them all in the stage of developing the model, then the final model can be of satisfactory performance (Steyerberg, 2009). However, the results in section 3.4.4 (the scenario of 'Presence or otherwise of noise variables') showed that if we do not increase the sample size, the presence of the variables which are 'speculative' in the dataset reduces the EPV and results in a worse model. Therefore, it is recommended that when EPV is low (or the sample size is small), one should just include influential



The simulation standard error (SE) among all EPV, the type of predictors and the recalibration status of models for the estimated calibration slope and Brier score were between (0.025, 0.001) and (0.00005, 0.00006), respectively.

Figure 3.8: Relative differences in the calibration slope and Brier score over EPV by the type of predictors based on 400 simulated datasets. The X axis is on the log scale.



The simulation standard error (SE) among all EPV, type of predictors and the recalibration status of models for the estimated C statistic were between (0.0003, 0.0004), respectively.

Figure 3.9: Relative differences in the C statistic over EPV by type of predictors based on 400 simulated datasets. The X axis is on the log scale.

variables (based on literature and common practice) in the model. In other words, all variables should be relevant and worth investing. To evaluate the relation between the number of variables in the risk model and the EPV, the following simulation study was set up.

To simulate outcomes, models with 22 variables, ten variables, seven variables, and four variables were obtained by fitting logistic regression models using the following variables. These were the categorical (five categories) and continuous versions of age at surgery, operative priority (three categories), renal problem (two categories), operation sequence (two categories), valve operation (three categories), ejection fraction (three categories), concomitant CABG surgery (two categories), respiratory disease (two variables), preoperative arrhythmias (three categories), diabetes (two categories), and year of operation (six categories). The calculated $\Delta\chi^2$ for these variables can be seen in Table 3.1 in section 3.2. To increase the comparability of the models, the strength of all models was equalised. This was performed using $\eta_{new} = A + B \times (\eta_{true} - \bar{\eta}_{true}) + \bar{\eta}_{true}$, where A and B were to retain the outcome prevalence and strength of the model and η_{new} and η_{true} were the new and true linear predictors.

Note that, as before, the model χ^2 for the true model fitted using ten original variables was 857.1 and the IQR of the models' χ^2 for those fitted using four, seven, and 22 variables were (807 to 895), (816, 889), and (802 to 893), respectively.

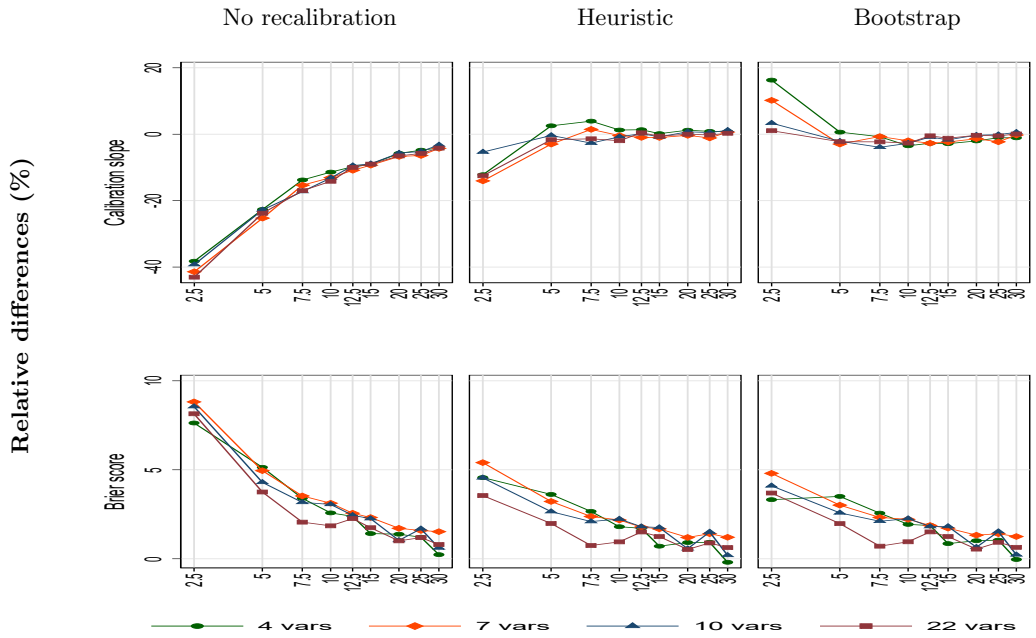
Number of variables in the model: results

This section illustrates relationships between performance measures and EPV in models with different number of variables.

Figures 3.10 and 3.11 demonstrate relative differences in the measures of performance quality over EPV across the numbers of variables in the model, and Table 3.4 displays the reference values for various number of variables and different modelling strategies. It is evident from the figures that including a large number of variables in the risk model did not affect the performance of the risk models as long as the EPV was held constant.

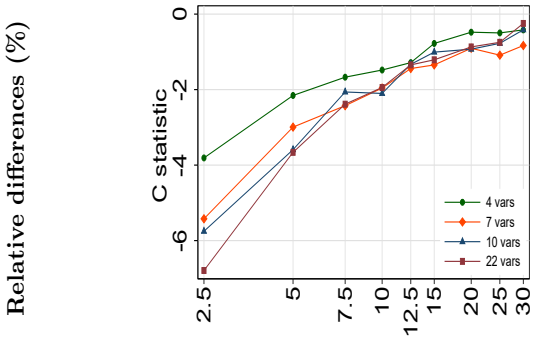
3.4.7 Degree of collinearity

In the modelling stage, it quite often happens that one or more of the predictors are capable of predicting other predictor(s) perfectly or partially. This is called collinearity.



The simulation standard error (SE) among all EPV, the number of variables and the recalibration status of models for the estimated calibration slope and Brier score were between (0.002, 0.521) and (0.00005, 0.00008), respectively.

Figure 3.10: Relative differences in the calibration slope and Brier score over EPV by the number of variables based on 400 simulated datasets. The X axis is on the log scale.



The simulation standard error (SE) among all EPV, the number of variables and the recalibration status of models for the estimated C statistic was between (0.0003, 0.0006), respectively.

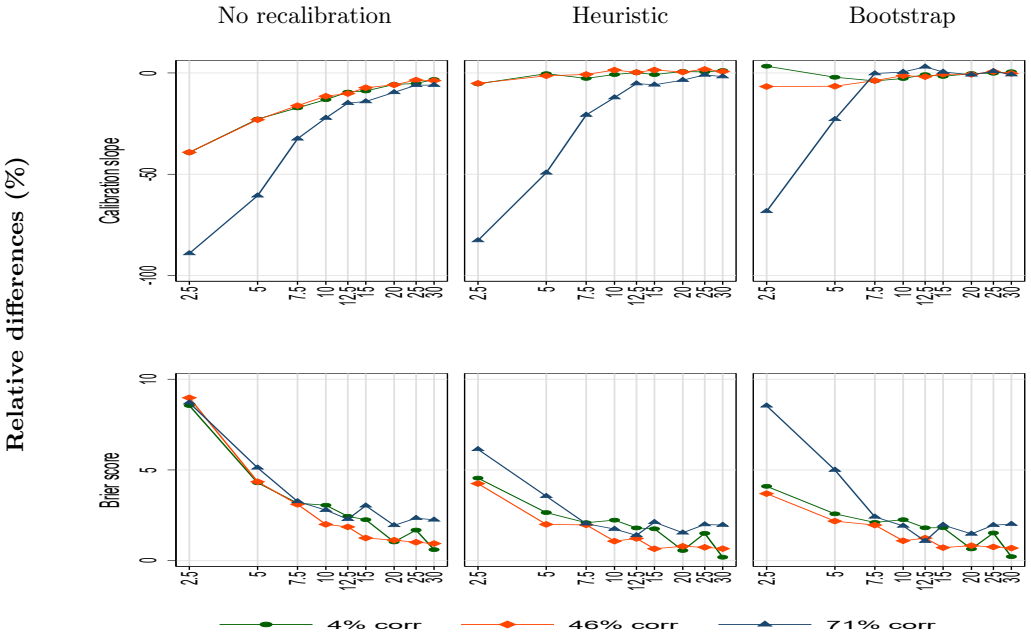
Figure 3.11: Relative differences in the C statistic over EPV by the number of variables based on 400 simulated datasets. The X axis is on the log scale.

If collinearity is perfect, the regression coefficients are indeterminate (they are not definitively or precisely determined) and their standard errors are infinite. If collinearity is not perfect the regression coefficients, although determinate, cannot be estimated with great precision or accuracy.

Since the estimation of coefficients can be problematic at presence of collinearity, this may affect the development of risk prediction models. In particular, more events might be required to develop a risk model as the dataset does not hold enough information. Therefore, we performed a simulation study to determine whether the EPV requirement changes when developing a risk model where there is a collinearity problem. The following method was used for the simulation study.

The following approach was used to change the original correlation between variables of real data, retaining original observations. The correlation between the predictors in the heart valve surgery data was changed and used to simulate new outcomes. This process is elucidated below. Firstly, true regression coefficients, β^{true}s, were estimated using the full size dataset. Then, seven multivariate standard normal variables with a specific correlation level were generated, one per original variable in the heart valve surgery data. Let $(z_1, z_2, z_3, \ldots, z_7)$ be multivariate standard normal variables and $(X_{age}, X_{valve}, X_{EjectFrac}, X_{cabg}, X_{prior}, X_{sequence},$ and $X_{renal})$ are seven variables from the heart valve surgery data. Therefore, there were seven pairs of variables, one from the simulated variables and one from heart valve surgery dataset; (z_1, X_{age}), (z_2, X_{valve}), $(z_3, X_{EjectFrac})$, $(z_4, X_{EjectFrac})$, (z_5, X_{cabg}), (z_6, X_{prior}) and (z_7, X_{renal}). Then, observations within each heart valve variable were sorted according to the order of its counterpart in the simulated multivariable normal variable. For example, consider pair (z_1, X_{age}); we first sort z_1 in ascending order (that is, $z_{11} < z_{12} < z_{13} < ... < z_{1n}$, where n is the number of observations), then, sort X_{age} only so that $X_{age\ 1} \leq X_{age\ 2} \leq X_{age\ 3} \leq \cdots \leq X_{age\ n}$. The process was repeated for all the pairs regardless of the type of variable. Table 3.6 presents the old and new pairwise Pearson correlation between variables of heart valve surgery data. From the table the goal was accomplished.

Ultimately, the new linear predictor was obtained by linearly combining the true regression coefficients and seven predictors with a new level of collinearity; that is, $\eta_{new} = \sum \beta_j^{true} X^{(j)}$ where β_j^{true} denotes that the true estimated regression coefficient corresponds to variable j, and is used to simulate new outcomes. To increase the



The simulation standard error (SE) among all EPV, the degree of collinearity and the recalibration status of models for the estimated calibration slope and Brier score were between (0.002, 0.081) and (0.00005, 0.00008), respectively.

Figure 3.12: Relative differences in the calibration slope and Brier score over EPV by the degree of collinearity based on 400 simulated datasets. The X axis is on the log scale.

comparability of the models, the process of adjusting for both the strength of the linear predictor and the outcome prevalence was also executed using the same approach as before to retain models with similar features; that is, the outcome prevalence was the same in all datasets and models had similar strength.

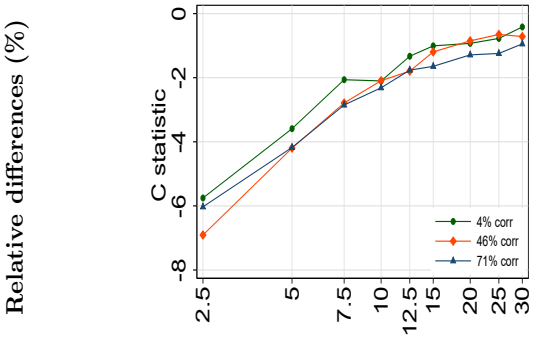
Note that the model χ^2 for the true model included variables with 0.04 average pairwise collinearity was 857.1. The IQR of the models' χ^2 for those included variables with 0.46 and 0.71 average pairwise collinearity which were (819 to 909) and (811 to 896), respectively.

Presence or otherwise of collinearity: results

Figures 3.12 and 3.13 illustrate the relative differences in the measures of performance quality over EPV across the amount of collinearity, and Table 3.4 displays reference values for collinearity levels and modelling strategies. As it is evident from the plots,

Table 3.6: Correlation matrix of predictors in datasets with different degrees of collinearity. Data size in all stages = 16679.

	The obtained correlation						
	Age at surgery	Concomitant CABG surgery	Renal problem	Ejection fraction	Operative priority	Operation sequence	Valve operation
Age at surgery	1						
Concomitant CABG surgery	0.23	1					
Renal problem	0.02	-0.01	1		Real Data		
Ejection fraction	0.05	0.12	0.09	1			
Operative priority	-0.01	0.02	0.13	0.19	1		
Operation sequence	-0.14	-0.12	0.04	0.06	0.09	1	
Valve operation	-0.06	-0.08	0.04	0.04	0.01	0.15	1
Average pairwise Pearson correlation	0.04						
Age at surgery	1						
Concomitant CABG surgery	0.49	1					
Renal problem	0.30	0.33	1				
Ejection fraction	0.50	0.51	0.41	1			
Operative priority	0.48	0.50	0.42	0.55	1		
Operation sequence	0.38	0.42	0.44	0.49	0.50	1	
Valve operation	0.50	0.50	0.40	0.54	0.53	0.47	1
Average pairwise Pearson correlation	0.46						
Age at surgery	1						
Concomitant CABG surgery	0.66	1					
Renal problem	0.40	0.40	1				
Ejection fraction	0.70	0.88	0.67	1			
Operative priority	0.65	0.88	0.64	0.89	1		
Operation sequence	0.50	0.56	0.78	0.71	0.74	1	
Valve operation	0.69	0.88	0.63	0.98	0.89	0.69	1
Average pairwise Pearson correlation	0.71						



The simulation standard error (SE) among all EPV, the degree of collinearity and the recalibration status of models for the estimated C statistic was between (0.0003, 0.0006), respectively.

Figure 3.13: Relative differences in the C statistic over EPV by the degree of collinearity based on 400 simulated datasets. The X axis is on the log scale.

collinearity did not effect the performance of risk models as long as the amount of it was medium or lower.

From the calibration slope plots, medium (46%) and low (4%) levels of collinearity did not change the quality of the risk models, holding the EPV constant. However, model performance was negatively affected at high (71%) levels of collinearity at EPV $\leq$ 15. This is because the regression coefficients of the model are not estimated accurately when both sample size is small and collinearity is high (see A.6 for bias in estimated regression coefficients). Furthermore, if the level of collinearity among variables in the model was around medium or lower, applying postestimation linear shrinkage almost completely removed overfitting at all EPV.

From the C statistic graph, the discrimination ability of the risk models was only slightly influenced by the amount of collinearity.

From the Brier score plot, this measure was influenced by the amount of collinearity in a similar pattern to the calibration slope. However, the overall performance of the risk models with highly correlated variables improved more noticeably with applying the heuristic shrinkage factor than employing the bootstrap method at small EPV. The Brier score showed that the heuristic shrinkage method is best at small EPV. Further investigation of bootstrap and heuristic shrinkages at the presence of EPV and collinearity scenarios also suggested that the bootstrap method shrinks the predictions

more than the heuristic method when correlation is 71% which is the opposite of what happens when the correlation is 46%. This pattern persists when the EPV is as high as 20.

In brief, it seems there should be at least 13 EPV to develop a model when variables are correlated at a level of medium or less. Moreover, a model developer needs a higher EPV, or needs to apply shrinkage when the collinearity level is more than medium. One practical point in this scenario is that models in which there is a high collinearity problem may appear with a very low (say, < 0.5) or negative calibration slope, especially at small EPV. That is because the estimated regression coefficients are inaccurate and imprecise when there is a high collinearity problem. In our simulation, these problems occurred in 19% of simulations at all EPV (mostly EPV$\leq$ 10) when collinearity was high (71%) compared to 3% at very low EPV (2.5) when collinearity was medium (45%) or much lower (4%). These problems persisted up to when EPV was 20 but dramatically decreased as EPV increased for high (71%) collinearity. In contrast, the problems mostly vanished when EPV reached 5 for medium (46%) or low (4%) collinearity. The results which led to a negative calibration slope were excluded from the analysis (0.65% of entire results of collinearity scenario). Moreover, 5.2% of the models when collinearity was large did not converge.

3.5 Conclusion

The focus of this chapter was on the sample size requirements for developing a reliable risk model with binary outcome. Harrell et al. in 1984 suggested that the performance of a risk model is a function of the number of events per variable, rather than sample size (Harrell et al., 1984). Also, he claimed that when EPV is less than ten the performance of a risk model deteriorates. Although he did not provide any supporting evidence for this suggestion, model-developers commenced using that threshold as a *rule of thumb* when developing a risk model regardless of the aim of constructing the model (Harbarth et al., 2000, Judith et al., 2002, Lassnigg et al., 2004, Jonas and Johnny, 2005, Voerman et al., 2007, Akins et al., 2008, Stone et al., 2011, William et al., 2013). The investigations in this chapter were started with a case study using the heart valve surgery data to answer the question of whether performance of a risk model is influenced by EPV.

Utilising four very common performance measures of the C statistic, D statistic, Calibration slope and Brier score in the case study, it was observed that the performance of risk models increased by moving from small to large EPV. That is, the EPV has an impact on the performance of a risk model. In the next step, another question was raised that whether EPV is the only factor which can affect the performance of model. A simulation study was set-up to address this issue. A number of scenarios were investigated; namely, the strength of risk model, the outcome prevalence, the presence of noise variables, the number of continuous predictors, the number of variables in the model, the presence of multicollinearity, and application of post-estimation shrinkage factor.

As evident from the results of the simulations in this chapter, as little as five EPV can be enough when developing a risk model including a few strong predictors where the aim is either that of prediction and estimation (see section A). Moreover, shrinkage can remove the optimism in the model when the model is strong at EPV equal to 5.

As the simulation study revealed, EPV needs to be at least 30 where the outcome prevalence is as high as 40% if the focus of developing a risk model is on prediction. This value reduced to ten , when the outcome prevalence reduced to 7% for the same circumstance. Applying shrinkage at the outcome prevalence of 40% and of 7% could reduce the EPV requirements to 10 and 2.5.

In addition, the presence of noise variables or more continuous variables were not associated with the performance of a risk model, if EPV is held constant. However, bias in the estimation of regression coefficients was lower in small EPV where there were more noise variables or more continuous variables in the risk model (see section A). Also, involving a further number of variables in the model did not affect the performance of it, holding EPV constant. The amount of overfitting declined dramatically subsequent to applying shrinkage where the EPV was as small as five at all number of variables in the model.

Additionally, the results of this chapter revealed that the performance of a risk model is only associated with the amount of multicollinearity when collinearity is high in the light of the EPV. Moreover, EPV should be at least 15 for developing a reliable unshrunk risk model with high collinearity (71%). The EPV requirement will drop to 7.5 when applying the bootstrap shrinkage and to 12.5 when using the Heuristic shrinkage.

Therefore, the EPV does not need to be large, say five or ten, when there are strong predictors in the risk model or when the total sample size is large;that is, outcome prevalence is small. Other factors such as the number of variables, the number of continuous variables, the number of noise variables do not have an impact on the performance of the risk model, holding the EPV constant. Although, applying the shrinkage factor improved the performance of a model for EPV between five and ten. Additionally, high multicollinearity (of $> 46\%$) deteriorates the performance of the model and EPV needs to be at least 15 to develop a risk model with acceptable performance.

In closing, when the collinearity is not high even EPV equal to five can lead to a reliable risk model as long as there are a number of strong predictors in the model and/or the total sample size is large. Moreover, the results of this simulation study agree with Harrell's *rule of thumb* (Harrell et al., 1984) in the sense that if there is no prior knowledge about the strength of the predictors there should be at least ten EPV when developing risk models. However, if one knows from literature that one or some variables have strong relationships with the outcome, having even EPV equal to five can result in satisfactory risk models. Furthermore, the strength of the true model can be important in the performance of studied models in accordance with this simulation study. Note that the strength of the true model in this study, based on model's χ^2, was large, except at the scenario of the strength of the risk model which was smaller

for the weak model and larger for the strong model. The simulated model may tolerate high collinearity at a small EPV if the corresponding true model is strong.

Moreover, based on this simulation study we also agree with Ambler et al. (2011) that the shrinkage method should be used even when EPV exceeds 30. However, the discrimination ability of the model may not improve in some scenarios such as those in our study.

This study was based on heart valve surgery data; other scenarios or datasets should be explored.

Chapter 4

Sample Size Requirements to Validate a Risk Model Using Independent Binary Data

The sample size requirements to obtain accurate predictive performance measures when developing a risk model using independent binary outcome has already been explored (see Chapter 3). The question of how many observations are required to investigate the validity of a risk model is now addressed.

Risk prediction models have an important role in the medical setting. Hence, these models should be developed carefully and be validated with great caution so that they can safely be used in other patients. It has been shown that risk models validated using small datasets may show exaggerated performance (Harrell (2001); Vergouwe et al. (2002); Vergouwe et al. (2005); Peek et al. (2007); Steyerberg (2009); Collins et al. (2014); Collins et al. (2015)). In other words, the performance of the risk model in small validation dataset may appear much better than the actual performance is in larger datasets. As a result, these validated models should not be trusted to be used in practice. For an example of such models with exaggerated performance, see Chamogeorgakis et al. (2009) in which a modified Thoracoscore, to predict in-hospital mortality after general thoracic surgery, was validated using 155 patients, of which just 8 died. The reported C statistic value was very high (0.95 with 95% confidence interval 0.91 to 0.99). As another example, see Brusselaers et al. (2009) where a data set of only 119 patients with only one outcome event was used to validate a risk prediction

model. The reported C statistic was absurd (1.00 with 95% confidence interval 1.00 to 1.00). Therefore, it is crucial to take the issue of number of events into account when validating a risk model in a medical setting.

This chapter discusses the sample size required to validate a risk prediction model in the context of the independent binary outcome. The chapter is structured as follows. All available research in this area is provided in the literature review, section 4.1. We show that the precision of performance measures depends on sample size in Section 4.2. A case study, in Section 4.3, was conducted to give an insight into the issue of sample size, or more specifically the number of events. A simulation study to explore whether outcome prevalence can affect the validation of risk models is discussed in Section 4.4. Finally, we summarise our findings and place them in the context of the existing literature in section 4.5.

4.1 Number of events in validating a risk model: A Review

There is a wide range of literature discussing the importance of the validation of risk prediction models (Royston and Altman, 2013, Collins et al., 2014). However, the design requirements for conducting a validation study have been little explored.

A key aspect when planning to validate a risk prediction model is to calculate the required sample size. As discussed in chapter 3, Harrell et al. (1996) suggested the performance of a risk prediction model is a function of the number of events.

Harrell et al. (1996)

Harrell et al. (1996) recommended that there should be at least 100 events in the validation data in order to successfully validate a risk model. No supporting evidence has been provided.

Vergouwe et al. (2005)

They conducted a simulation study to determine the required number of events to detect relevant changes in the performance of logistic regression model. In their study, Vergouwe et al. (2005) used real data consisting of 544 patients, who were treated with

chemotherapy for metastatic testicular germ cell cancer. In their data, 254 patients (47%) had the event of interest (histology of retroperitoneal lymph nodes).

They evaluated three various validation scenarios in which the studied model was invalid but validation datasets had a similar case-mix to the development datasets, and one validation scenario in which the tested model was valid, but the validation dataset had a different case-mix to the development dataset. Those scenarios are as follows. *First*, a situation where the predicted probabilities of the model in the validation dataset were systematically too high or too low compared to true proportions, which might happen when an important factor is missing from the developed risk model (Steyerberg et al., 2004). *Second*, a case in which the developed model was overoptimistic in the validation data and its predictions were too extreme. This situation can arise as a result of inadequate shrinkage of the regression coefficients in the development stage (Altman and Royston, 2000). *Third*, a circumstance in which the estimated regression coefficients were imprecise or the definition of the predictor variables in the development data was different than those in the validation dataset, leading to the wrong regression coefficients for the validation samples. *Finally*, a case in which the developed model is valid but the dataset included more homogeneous patients. That is the validation sample had different case-mix. In technical terms, it means that the variances of explanatory variables in the validation sample were smaller than those in the development data.

Vergouwe et al. used the calibration curves and the Hosmer-Lemeshow test to quantify the calibration ability of the model, the C statistic to measure discrimination and both the Brier score and Nagelkerke's R^2 to quantify overall performance (see section 2.5).

They concluded that a validation sample with at least 100 events and 100 nonevents is necessary to detect large differences in model performance with 80% power. They additionally reported that the sample size required to detect a decrease in the discrimination ability of a model for samples with low outcome prevalence (less than or equal to 10%) is larger.

Vergouwe et al. (2005) only studied three sample sizes with two outcome prevalences. Furthermore, they investigate a case where both the risk model is valid and the case-mix in the validation data is similar to that in the development data.

Peek et al. (2007)

They conducted a validation study to compare the performance of four models on a large dataset. They also investigated the effect of the sample size on the validation results. The logistic prediction models were the Simplified Acute Physiology Score II (SAPS II) (Le Gall et al., 1993), the Acute Physiology and Chronic Health Evaluation II (APACHE II) (Knaus et al., 1985), and the Mortality Probability Models II (MPM0 II and MPM24 II) (Lemeshow et al., 1993). These models are commonly used to predict the probability of in-hospital mortality in intensive care units (ICUs).

They used a dataset of 42,139 patients (of whom 19.8% died in hospital) from the National Intensive Care Evaluation (NICE) registry in Netherlands. The outcome prevalence in their data was similar to that in the datasets used to develop SAPS II, APACHE II and MPM II (21.8%, 19.7% and 20.8%, respectively). Peek et al. (2007) used the C statistic, Brier score, calibration slope, and the Hosmer-Lemeshow test to evaluate the performance of the model and varied the sample size by randomly sampling from the entire dataset.

In their study, Peek et al. (2007) found that considerable sample sizes such as 1000 patients (20% events) were required to compare and externally validate those prediction models.

However, in their external validation study to evaluate the effect of sample size Peek et al. (2007) did not study the effect of various outcome prevalence.

Collins et al. (2015)

In their external validation study, they employed The Health Improvement Network (THIN) data, consisting of information from 2 million patients from primary care records held at general practice surgeries around the UK, to assess the influence of sample size on the performance of three sex-specific Cox regression prediction models.

The prediction models were QRISK2 (Hippisley-Cox et al., 2008), Cox Framingham (**?**) and QDScore (Hippisley-Cox et al., 2009). While the first two prediction models were developed to predict the 10-year risk of developing cardiovascular disease, the last was constructed to predict the 10-year risk of developing type 2 diabetes. The outcome prevalences of data in which QRISK2, Framingham, and QDScore were developed were 6%, 15% and 3%, respectively, compared to 5% and 3% which were the prevalence of

developing cardiovascular disease and type 2 diabetes, respectively, in the THIN data. In their resampling study, Collins et al. (2015) investigated whether the estimation of C index, D statistic, R^2_D, ρ^2_{OXS}, Brier score and calibration slope was unbiased and precise, where R^2_D, ρ^2_{OXS} are two R^2-type measures (Choodari-Oskooei et al., 2012a,b) used with survival data. Collins et al. (2015) varied the sample size by stratified sampling from events and nonevents groups according to the outcome prevalence in THIN data.

Collins et al. (2015) found that the estimation of all six performance measures varied largely when the number of events was smaller than 100. They also reported that the mean standardized biases in the performance measures were larger than 10% when the number of events was less than 75. No explanation on why there was bias was reported. Thus, Collins et al. (2015) concluded that to externally validate a risk prediction model there should be a minimum of 100 events, preferably 200 in a validation dataset.

Collins et al. (2015) only studied one dataset with the small outcome prevalence to investigate the effect of sample size on a validation study. Moreover, the prevalence of developing cardiovascular disease in their study (THIN) was lower than that in a dataset employed to develop QRISK2 and Framingham.

Jinks et al. (2015)

They derived formulae based on the D statistic to calculate the required sample size for risk prediction models using time-to-event data. Inspired by Armitage et al. (2001) for comparison of the means of two independent groups with equal within-group variance, they assumed that either there is a previous study (with e_1 events) in which the D statistic has been estimated (D_1 with variance of σ_1) or the researcher has a target D statistic in mind.

Thus, with the assumption that a previous study existed, they supposed that a researcher wished to validate the estimate of the `D` statistic for the model in a new study (which has e_2 events). They also assumed that $\texttt{D}_2$ and σ^2_2 were the estimates of D and its variance in that new study. Since the standard error of the differences between D_1 and D_2 did not explicitly include e_1 and e_2, they assumed $\lambda = \sigma^2_1 e_1 = \sigma^2_2 e_2$, where λ was a model- and disease-specific structural constant which could be either obtained from the previous study or estimated using an approximation incorporating a value of `D` and the proportion of censoring (*cens*) in the data using $\lambda = 2.66 + 1.26\,\texttt{D}^{1.9} - 1.65(\texttt{D} \times \texttt{cens})^{1.3}$. They developed this equation using simulated datasets.

4. SAMPLE SIZE REQUIREMENTS TO VALIDATE A RISK MODEL USING INDEPENDENT BINARY DATA

Jinks et al. (2015) found that $e_2 = \lambda \left[\left(\frac{\delta}{z_{1-\alpha}+z_{1-\beta}} \right)^2 - \sigma_1^2 \right]^{-1}$ could be used to detect differences (δ) in D between the first and second studies with α significance level and $1-\beta$ power, where z_x was the x-quantile of the standard normal distribution. Moreover, they suggested that $e_2 = \lambda \left[\left(\frac{w}{z_{1-\alpha}} \right)^2 - \sigma_1^2 \right]^{-1}$ could be used to obtain a required sample size for multivariable prediction models in time-to-event data, based on the precision of the estimate of D in a new study in terms of a confidence interval of width $2w$.

For cases where there was no previous study, they assumed that there was a fixed target value of $\mathtt{D}^*$ with zero uncertainty ($\sigma^2 = 0$). Their previous formulae, by substituting the new assumption, became $e_2 = \lambda \left[\frac{\delta}{z_{1-\alpha}+z_{1-\beta}} \right]^{-2}$ and $e_2 = \lambda \left[\frac{w}{z_{1-\alpha}} \right]^{-2}; \qquad w > 0,$ respectively.

As an example, Jinks et al. (2015) considered the study by Collette et al. (2008), which was conducted to compare three existing staging systems for advanced liver cancer. The estimated D statistic for the CLIP prognostic model was 1.01 ($\sigma = 0.09$), using time-to-event dataset of 538 patients with 7% censoring. To validate this model using a new data, one can employ the information in the paper, and use Jinks equation to obtain $\lambda (= e_1\sigma_1^2 = 502 \times 0.09^2 = 4.1)$. Based on Jinks formulae there should be 558 events in the validation dataset when the validation study is used to detect differences of $\delta = 0.25$ in D between the first and second studies with one-sided $\alpha = 0.05$, 90% power. If the aim is to specify a 95% confidence interval with 0.4 width for D, there should be 391 events in the dataset.

Now assume that the aim is to add a prognostic factor to the CLIP model, which is believed to improve the prognostic ability of the model (to $\mathtt{D} = 1.3$). There is no previous study, thus, λ can be estimated using a target value of D (=1.3) and the censoring proportion (expected to be 10%) in the dataset. Thus, the estimated λ is 4.62, and for a non-inferiority validation study to detect differences of $\delta = 0.25$ in D between the first and second studies with one-sided $\alpha = 0.05$, and 90% power, based on Jinks' equation, there should be 633 events in the validation dataset.

Jinks et al. (2015) only studied validating a risk model using time-to-event outcomes based only on the D statistic.

Summary

To sum up, there are some guidelines on how many events are required to validate risk prediction models. The rule of 100 events to validate a risk model was suggested by Harrell et al. (1996) without evidence. Later, in 2005, Vergouwe et al., using a simulation study, recommended that there must be at least 100 events and 100 non-events in validation data to be able to detect some types of model invalidity with a specific power using the binary outcomes. Peek et al. (2007) suggested using at least 200 events to validate a risk model in the context of binary outcome. Moreover, Collins et al. (2015) suggested using at least 100 events (preferably 200 events) when validating a reliable risk model developed using time-to-event outcomes.

There are different recommendations on how many events are required to validate a risk model. This leads to the consideration that not all model-makers trust those rules. In fact, in a review paper, Collins et al. (2014) found that only 41 papers out of 78 had 100 number of events or more in their validation dataset. Thus, this study is to ascertain the required number of events to validate a correctly-specified model using datasets with various outcome prevalences. But, first, we analytically illustrate that the precision of performance measures depends on the sample size.

4.2 Precision of the performance measures and sample size

When validating a risk model, the key point is that the estimated performance measure should have high precision. In other words, the standard error of the estimated measures should be small. In this section, it will be shown how the variance of calibration slope, D statistic, and C statistic depends on sample size, and thus the number of events.

Variance of calibration slope

To calculate the calibration slope (CS), we fit a logistic model on data using an outcome variable and the estimated linear predictor ($\hat{\eta}$) as the only predictor, where the outcome and $\hat{\eta}$ come from the validation dataset. That is

$$log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_{CS}\hat{\eta}_i, \tag{4.2.1}$$

where β_{CS} is calibration slope. The coefficients of this logistic model are estimated by $(H^TWH)^{-1}H^TZ$ where

$$H = \begin{pmatrix} 1 & \hat{\eta}_1 \\ \vdots & \vdots \\ 1 & \hat{\eta}_n \end{pmatrix},$$

$$W = \begin{pmatrix} \hat{p}_1(1-\hat{p}_1) & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{p}_n(1-\hat{p}_n) \end{pmatrix}$$

Z is a vector with ith element $y_i - \hat{p}_i$. Moreover, $\hat{\eta}_i$ is the estimated prognostic index $(\hat{\beta}X)$ for the ith patient from the validation data with n patients, where the coefficients β are estimated using development data but the predictor values (x) are from the validation datasets and $\hat{p}_i$ is estimated risk obtained using $\hat{\eta}_i$.

Let us suppose that we estimate the coefficients of the model in equation 4.2.1 using the validation data and $\tilde{\eta}_i$ and $\tilde{p}_i$ are the estimated prognostic index and risk for ith patient from the validation data. We now make the assumption that the risk model is perfectly estimated. That is, $\hat{\eta} = \tilde{\eta} = \eta$ and $\hat{p} = \tilde{p} = p$ where η and p are the true prognostic index and risk. Moreover, we assume that $\eta_i \sim N(\mu, var(\eta))$ where μ and $var(\eta)$ are the mean and variance of η. In fact, the prognostic index is likely to follow the normal distribution as the dimension of the parameter vector (βs) increases based on the central limit theorem (Choodari-Oskooei et al., 2012c).

It can be written that $\eta_i = log(\frac{p_i}{1-p_i})$ in which rearranging will give $\frac{p_i}{1-p_i} = e^{\eta_i}$ and so $p_i = \frac{e^{\eta_i}}{e_i^{\eta}+1}$.

Furthermore, w_i is also a function of η_i. That is,

$$\begin{aligned} w_i &= p_i(1-p_i) \\ &= (\frac{e_i^{\eta}}{e_i^{\eta}+1})(\frac{1}{e_i^{\eta}+1}) \\ &= \frac{e_i^{\eta}}{(e_i^{\eta}+1)^2}. \end{aligned}$$

Given that the first derivative of w_i is $\frac{dw_i}{d\eta_i} = -\frac{e_i^{\eta}(e^{\eta_i}-1)}{(e^{\eta_i}+1)^3}$, the w_i around $\eta_i = \bar{\eta}$ ($\bar{\eta}$ is the mean value of prognostic index within the validation data) can be approximated

using the Taylor series expansion as follows.

$$\begin{aligned} w_i &= \frac{e^{\bar{\eta}}}{(e^{\bar{\eta}}+1)^2} - \frac{e^{\bar{\eta}}(e^{\bar{\eta}}-1)}{(e^{\bar{\eta}}+1)^3}(\eta_i - \bar{\eta}) \\ &= A + B\,\eta_i, \end{aligned} \tag{4.2.2}$$

where

$$A = \frac{e^{\bar{\eta}}}{(e^{\bar{\eta}}+1)^2}\,(1 + \bar{\eta}\,\frac{e^{\bar{\eta}}-1}{e^{\bar{\eta}}+1})$$

$$B = \frac{-e^{\bar{\eta}}\,(e^{\bar{\eta}}-1)}{(e^{\bar{\eta}}+1)^3}.$$

The variance of the calibration slope (β_{CS}) can be written as

$$\texttt{Var}(\beta_{CS}) = \frac{\sum w_i}{\sum w_i(\sum w_i\eta_i^2) - (\sum w_i\eta_i)^2}. \tag{4.2.3}$$

Substituting 4.2.2 in the variance formula (equation 4.2.3), we have

$$\texttt{Var}(\beta_{CS}) \approx \frac{\sum A + \sum B\eta_i}{(\sum A + \sum B\eta_i)(\sum A\eta_i^2 + \sum B\eta_i^3) - (\sum A\eta_i + \sum B\eta_i^2)^2}. \tag{4.2.4}$$

The first three moments of the distribution of H the prognostic index can be used to estimate $\sum \eta_i$, $\sum \eta_i^2$ and $\sum \eta_i^3$ which are $\texttt{E}(\eta) = \bar{\eta} = \frac{1}{n}\sum \eta_i$, $\texttt{E}(\eta^2) = \frac{1}{n}\sum \eta_i^2 = \bar{\eta}^2 + Var(\eta)$, and $\texttt{E}(\eta^3) = \frac{1}{n}\sum \eta_i^3 = \bar{\eta}^3 + 3\,\bar{\eta}\,Var(\eta)$. Therefore, $Var(\beta_{CS})$ (equation 4.2.4) can be approximated as

$$\frac{n(A + B\bar{\eta})}{n^2(A + B\bar{\eta})[A\,(\bar{\eta}^2 + Var(\eta)) + B\,(\bar{\eta}^3 + 3\,\bar{\eta}\,Var(\eta))] - n^2\,[(A\bar{\eta} + B\,(\bar{\eta}^2 + Var(\eta)))^2]} \tag{4.2.5}$$

where $\bar{\eta}$ and $var(\eta)$ can be obtained using the validation data. The equation 4.2.5 can be simplified as

$$\frac{1}{n}.\frac{A + B\bar{\eta}}{(A + B\bar{\eta})[A\,(\bar{\eta}^2 + Var(\eta)) + B\,(\bar{\eta}^3 + 3\,\bar{\eta}\,Var(\eta))] - (A\bar{\eta} + B\,(\bar{\eta}^2 + Var(\eta)))^2}. \tag{4.2.6}$$

That is, the variance of the calibration slope is inversely proportional to the sample size n.

Variance of D statistic

The D statistic is calculated by fitting logistic regression to data consisting of the outcome variable (Y_i) and z_i, where $z_i = \sqrt{\frac{\pi}{8}}\,\Phi^{-1}\left(\frac{i-3/8}{n+1/4}\right)$ in which i is the rank order of the prognostic index (η), where $\Phi^{-1}(.)$ denotes the inverse standard normal distribution function. The slope from this regression is the D statistic. In other words, one calculates calibration slope using η_i, and transforms it to calculate D statistic. Therefore, one can approximate the D statistic by calculating the product of calibration slope and $\frac{\pi}{8}$. For example, the D statistic is approximately equal to 0.31 for the calibration slope of 0.78, or the D statistic is equal to 0.39 for when there is perfect calibration and the calibration slope equals one.

Likewise, one can approximate the standard error of D statistic by calculating the product of the standard error of the calibration slope and $\frac{\pi}{8}$. For instance, if the standard error of calibration slope is 0.32, the standard error of D statistic approximately equals 0.13.

Thus, like the variance of calibration slope, the variance of D statistic is inversely proportional to the sample size n.

Variance of C statistic

Under the normality assumption of prognostic index (η_i), the C statistic and D statistic are closely related. Based on this assumption, an analytical relationship between C statistic and D statistic is derived as follows.

Let us assume that $\eta_i|Y_i = 1 \sim N(\mu_1, \sigma^2)$ with $P(Y_i = 1) = p_1$ and $\eta_j|Y_j = 0 \sim N(\mu_0, \sigma^2)$ with $P(Y_j = 1) = p_0$. Further, assume that the conditional distribution of η given $Y = y$ is normal with the mean μ_y and variance $2\sigma^2$. This formulation corresponds to linear discriminant analysis (LDA) (Anderson, 1958), which is equivalent to logistic regression model (Efron, 1975). In LDA, one assigns subject i with prognostic score η_i to the population who had experienced the event with probability $P(Y_i = 1|\eta_i)$. This probability can be expressed in terms of a logistic model as

$$P(Y_i = 1|\eta_i) = \frac{1}{1 + exp(-(\beta_0 + \beta_\eta \eta_i))}, \tag{4.2.7}$$

where $\beta_0 = -log\frac{p_1}{p_0} + \frac{1}{2}\frac{(\mu_1^2 - \mu_0^2)}{2\sigma^2}$ and $\beta_\eta = \frac{\mu_1 - \mu_0}{2\sigma^2}$.

Standardising the prognostic indexes, η_1 and η_0, and then multiplying them by $\frac{\pi}{8}$ gives the terms $Z'^{(1)}$ and $Z'^{(0)}$, which are distributed as $N(0; \frac{\pi}{8})$.
That is: $Z'^{(1)} \sim N(\frac{\mu_1-\mu_y}{\sigma\sqrt{2}}, \frac{\pi}{8})$ and $Z'^{(0)} \sim N(\frac{\mu_0-\mu_y}{\sigma\sqrt{2}}, \frac{\pi}{8})$.

These formulations also correspond to the LDA with the transformed variables $Z'^{(1)}$ and $Z'^{(0)}$, and can be expressed in terms of a logistic regression model for the binary outcome variable Y with Z' as a predictor:

$$P(Y_i = 1|z'_i) = \frac{1}{1 + exp(-(\beta_0 + \beta_{z'} z'_i))}, \tag{4.2.8}$$

Therefore, the D statistic is the coefficient of Z', $\beta_{z'}$, in the above model and can be estimated approximately by analogy to β_η as

$$D \approx (\frac{8}{\pi}) \left(\frac{\mu_1 - \mu_0}{\sqrt{2\sigma^2}}\right) = \frac{8}{\pi}\Phi^{-1}(C).$$

Therefore,

$$C \approx \Phi(\frac{\pi}{8} D).$$

In order to obtain the variance of C statistic, the variance of $\Phi(\frac{\pi}{8} D)$ should be calculated. That is,

$$Var(C) \approx Var(\Phi(\frac{\pi}{8} D)).$$

To do so, the delta method can be employed which is

$$Var(C) \approx Var(D) \left(\frac{d\Phi(\frac{\pi}{8} D)}{dD}|_{D=\hat{D}}\right)^2, \tag{4.2.9}$$

where $\hat{D}$ can be estimated in the validation dataset.

The first derivative of $\Phi(\frac{\pi}{8} D)$ is $\frac{\pi}{8}\,\phi(\frac{\pi}{8} D)$, where $\phi(.)$ denotes the standard normal density function. Thus, the equation 4.2.9 is equivalent to

$$Var(C) \approx Var(D) \left(\frac{\pi}{8}\,\Phi(\frac{\pi}{8}\,\hat{D})\right)^2. \tag{4.2.10}$$

That is, like the variance of D, the variance of C statistic is inversely proportional to the sample size n.

For example, under the assumption of perfect calibration and given that variance of D statistic is known to be 0.13, the variance of C statistic approximately equals 0.018.

Validating the accuracy of the proposed formulas

Simulation was used to validate the derived formulas for estimating the standard error of the performance measures.

A set of 200 validation datasets (N=800) based on a prognostic index with mean μ and standard deviation of σ was simulated. The outcome variables were simulated from these prognostic indices. We allowed the prognostic index to have either a normal or log-normal distribution. Note that its assumed that the model to be validated is the correct model. The three measures of calibration slope, D statistic and C statistic were then obtained using the outcomes and prognostic indices. The observed standard errors of the performance measures were averaged over 200 simulated datasets and were compared with the expected values which were calculated using the derived formulae (Table 4.1).

As can be seen, the estimated standard errors for the calibration slope are almost as good as the observed standard errors for both distribution. However, the standard errors of D statistic, and consequently C statistic, are not as good as the observed standard errors when the standard deviation of the prognostic index is large.

Table 4.1: The expected and observed standard errors of performance measures. The observed standard errors were averaged over 200 simulations.

Distribution* of η	Equivalent median risk (Min, Max)		Standard error of		
			Calibration slope	D statistic	C statistic
N(μ= 2, σ = 0.25)	87% (77%, 95%)	expected	0.445	0.175	0.025
		observed	0.442	0.176	0.031
LN(μ= 2, σ = 0.25)			0.453	0.175	0.031
N(μ= 2, σ = 0.5)	87% (70%, 99%)	expected	0.236	0.093	0.013
		observed	0.223	0.178	0.029
LN(μ= 2, σ = 0.5)			0.249	0.176	0.028
N(μ= 1, σ = 0.25)	72% (61%, 91%)	expected	0.321	0.126	0.018
		observed	0.327	0.130	0.023
LN(μ= 1, σ = 0.25)			0.341	0.130	0.023
N(μ= 1, σ = 0.5)	71% (54%, 98%)	expected	0.164	0.064	0.009
		observed	0.170	0.136	0.022
LN(μ= 1, σ = 0.5)			0.200	0.133	0.021
N(μ= 0.5, σ = 0.15)	62% (55%, 80%)	expected	0.486	0.191	0.028
		observed	0.490	0.118	0.021
LN(μ= 0.5, σ = 0.15)			0.505	0.118	0.021
N(μ= 0.5, σ = 0.2)	61% (53%, 82%)	expected	0.365	0.143	0.021
		observed	0.371	0.118	0.021
LN(μ= 0.5, σ = 0.2)			0.386	0.118	0.021

N(μ,σ): normal distribution with the mean μ and standard deviation σ.
LN(μ,σ): Log-normal distribution with the mean μ and standard deviation σ.

4.3 Case study

From section 4.2, the precision of performance measures have a nonlinear relationship with the sample size. A case study was conducted to empirically present that the precision of the estimated performance measures on real validation data can be affected by the number of events. This case study was carried out using the heart valve surgery data.

4.3.1 Method

We employed all predictors in Ambler's model (Ambler et al., 2005), in total 14 predictors equivalent to 16 degrees of freedom, and fitted the logistic risk model on the data from the first five years (the development data in Ambler's paper). The fitted model

was validated on the second part of the data (data from the years 2000 to 2005, which was the validation data in Ambler's paper). The estimated performance measures on the validation data were: D statistic=1.76, C statistic=0.77, calibration slope=1.11 and Brier score=0.050. That is, it seems that the agreement between the predicted and observed outcomes is not good (calibration slope greater than one), but this model was successful in terms of distinguishing patients from different risk groups.

Therefore, the model was recalibrated by transforming the predicted log-odds using 50% of the validation dataset, a linear function found to be appropriate. The recalibration improved the fit of the model, calibration slope was equal to one (obtained using the rest of the validation data). Thereafter, we validated this recalibrated model on validation samples of different sizes sampled from the original validation data. The size of the validation datasets were adjusted by sampling separately without replacement from events and non-events (retaining the outcome prevalence of the heart valve data, 0.6%). For each number of events, 200 samples were selected. For each sample, performance measures were calculated based on the recalibrated model.

4.3.2 Results

Table 4.2: Mean value (standard deviation) of the predictive performance measures by number of events over 200 external samples.

Number of Events	C statistic	D statistic	Calibration Slope	Brier Score
500	0.77 (0.006)	1.75 (0.052)	1.00 (0.027)	0.454 (0.0003)
300	0.77 (0.010)	1.76 (0.087)	1.00 (0.046)	0.453 (0.0005)
250	0.77 (0.012)	1.76 (0.099)	1.00 (0.050)	0.454 (0.0006)
200	0.77 (0.013)	1.75 (0.113)	1.00 (0.060)	0.454 (0.0007)
150	0.77 (0.017)	1.75 (0.138)	1.00 (0.074)	0.454 (0.0007)
125	0.77 (0.018)	1.78 (0.151)	1.00 (0.081)	0.453 (0.0009)
100	0.77 (0.024)	1.76 (0.188)	1.00 (0.094)	0.454 (0.0010)
75	0.77 (0.025)	1.76 (0.202)	1.00 (0.107)	0.454 (0.0011)
50	0.77 (0.032)	1.76 (0.269)	1.00 (0.147)	0.454 (0.0016)

We expected the precision of estimation for measures in external samples to increase by increasing the sample size.

Table 4.2 shows the precision of predictive performance measures was affected by the number of events. The standard deviation of the measures across samples was

smallest for the largest number of events, and largest for the smallest number of events in the validation sample.

4.3.3 Conclusion

To sum up, this case study also showed that the number of events in the validation samples has an effect on the precision of the estimated calibration and discrimination measures and overall performance measures. The larger the number of events in the validation sample, the higher the precision of the estimated validation measures. This was indeed what we showed in section 4.2.

Further investigation was conducted to understand the reasons for the trends seen in Table 4.2. The effect of the number of events on the separation between $\hat{\eta}^{(0)}$ and $\hat{\eta}^{(1)}$ is illustrated in Figure 4.1, where $\hat{\eta}^{(0)}$ and $\hat{\eta}^{(1)}$ are the linear predictors that correspond to the nonevent and event groups, respectively. As can be seen, the variation within $\hat{\eta}^{(0)}$s and $\hat{\eta}^{(1)}$s was least at the largest number of events compared to the smallest number of events. This indeed was the pattern observed in Table 4.2; the standard deviation of the measures was the largest at the smallest number of events. Additionally, in our analytical work in section 4.2, we showed that the precision of the performance of measures increases with increasing sample size.

This case study was conducted by sampling without replacement from the real data. Also, the outcome prevalence was fixed at that observed in the original heart valve surgery data. Therefore, a simulation study was performed by varying the outcome prevalence in the validation sample to ascertain how many events are required to achieve satisfactory precision in the validation measures.

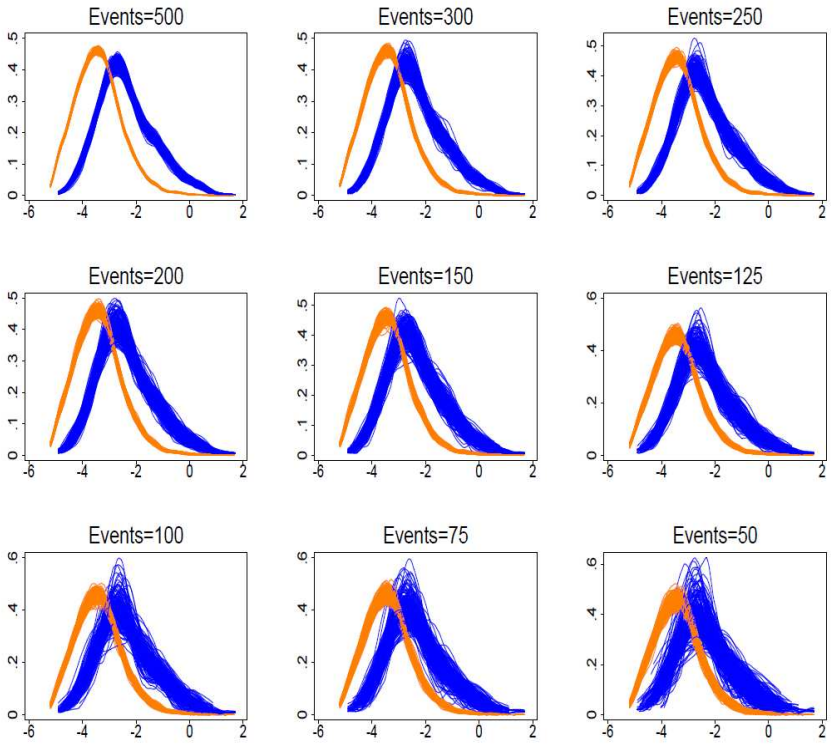


Figure 4.1: Separation between $\hat{\eta}^{(0)}$ and $\hat{\eta}^{(1)}$ by number of events. $\hat{\eta}^{(0)}$ and $\hat{\eta}^{(1)}$ are linear predictors corresponding to nonevent and event groups, respectively.

4.4 Simulation study

We required to assess whether the precision of these measures can be worsen by the number of events at different outcome prevalences when validating a correct risk model.

A number of simulation studies were performed using the heart valve surgery data. All investigations in this chapter started with assuming that the risk model had been perfectly estimated in the development data (see section 4.2). Details of these simulations are now described.

4.4.1 Overview

A model was derived using the entire heart valve surgery data with 10 pre-specified variables. The linear predictor was obtained for all patients in the data and used to generate new outcomes to investigate the following factors:

§ *Different risk profile in validation set*; to learn whether the sample size requirements should change if the risk profile varies. Three levels of low, medium, and high were considered for the risk profile.

§ *Number of events*; the following number of events were examined; 25, 50, 75, 100, 125, 150, 200, 250, 300 and 400.

For each combination of these factors, the following approach was repeated 2000 times. The linear predictor variable was sorted in ascending order, and split into three equal groups: low-, medium-, and high-risk groups. The proportion of risk groups was manipulated to change risk profile in the simulated data. This was achieved by sampling the required proportion from each group (see section 4.4.2). For each number of events, the simulated data was sorted by outcome and the first observations were selected to be in the studied sample until the required numbers of events had been observed, and the quality of the predictions were then quantified using the calibration slope, C statistic and Brier score. The results for the D statistic are not presented for this simulation study since they are correlated with those from the C statistic. The performance measures from the reduced validation datasets were compared with those from the full size validation datasets using percent relative differences. Detailed information regarding the simulation scenarios are now described.

4.4.2 Different risk profile in validation set

In reality, the prevalence of some disease might be smaller or larger than others, or patients of one country may be highly at risk of developing a disease than other countries. For instance, heart failure is a major public health issue in the USA, with over 5.8 million sufferers compared to over 23 million worldwide (Bui et al., 2011). To reflect such situations, datasets with different risk profiles were produced and utilised to validate the risk model.

To do so, the true model was fitted on the entire heart valve surgery data (N=32,839). For each number of events, three datasets were simulated (with low, medium and high risk profile), and the required number of events and nonevents were selected by separate random sampling from them. To change the risk profile, the true linear predictors were estimated and sorted in ascending order. The linear predictors were split into three groups of low, medium and high risk such that the first third of patients who had the lowest linear predictors constituted the low-risk group, and the second and third sections of the remaining patients with medium and large linear predictors formed the medium- and high-risk groups, respectively. Then, samples without replacement for the low-risk profile were produced by selecting 60% of the patients from the lowest risk group, 33% of those from the middle, and the remaining 7%, from the highest risk group category. These proportions were about 7%, 33% and 60% from low- medium- and high-risk groups, respectively, to form the high-risk profile samples. Furthermore, 33% from each risk group were selected to create the medium-risk profile samples.

Figure 4.2 illustrates the distribution of the linear predictors in one of the full size samples across different risk profiles. The average outcome prevalence in the low, medium and high risk profile samples were 3.6%, 6.4% and 9.2%, respectively.

4.4.3 Generating new outcomes

The logistic regression prediction model was derived using the entire heart valve surgery data and all pre-specified predictors (see section 3.2). For each patient, the linear predictor, $lp = -5.95 + 0.036age + 0.36Mitral + 0.69Aortmitr + 0.41cabge + 0.96Renal + 0.21FairEjec + 0.70PoorEjec + 0.51Urgent + 1.60Emergency + 0.67Sequence$ was obtained from the prediction model and used to simulate new outcomes. The predictor values and outcomes were combined to make new datasets. The generated new datasets

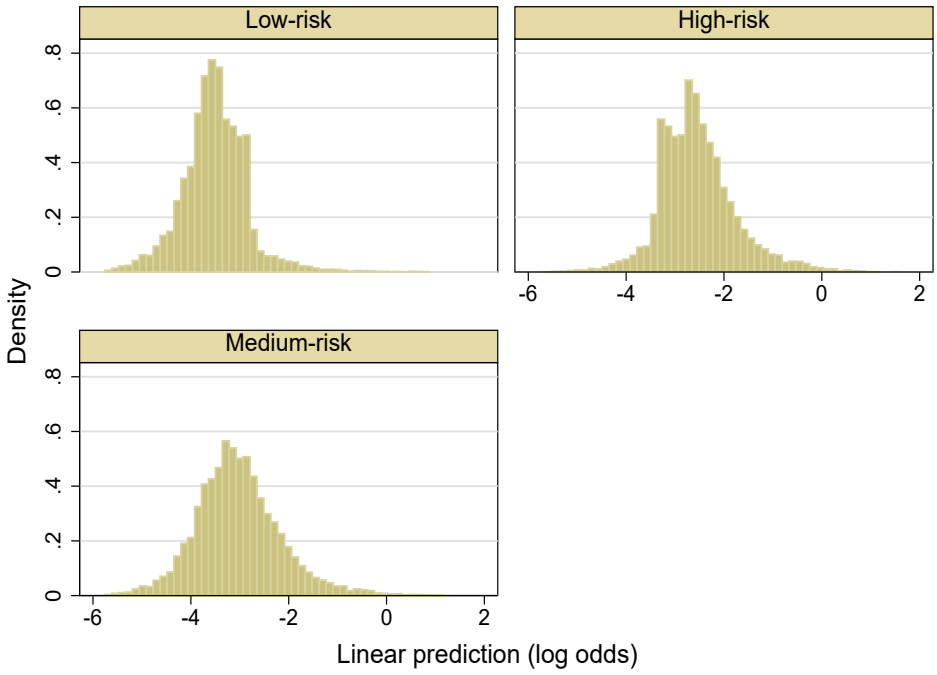


Figure 4.2: Distribution of linear predictors in samples with varying risk profile, N = 32,839.

were referred to as the validation samples. This method mimics the situation where the validation sample originates in the same underlying population as the development dataset. The same approach was taken to create outcomes for patients in the low-risk and high-risk profile samples.

4.4.4 Results

Figure 4.3 displays percent relative differences in the estimated performance measures by number of events in the validation samples over different risk profiles based on 2000 simulated datasets, and Table 4.3 presents the reference values and their standard values obtained based on 2000 full size validation data. Note that reference values were obtained by calculating the performance measures using the entire validation datasets. Furthermore, standard error of the reference measures are the standard deviation of measures across 2000 validation datasets.

Table 4.3: The reference values of the performance measures and their standard deviations based on 2000 simulated datasets.

Risk profile	Performance measure		
	C statistic	Brier score	Calibration slope
Low	0.69 (0.0098)	0.033 (0.0011)	1.00 (0.0444)
Medium	0.74 (0.0071)	0.055 (0.0013)	1.00 (0.0299)
High	0.72 (0.0062)	0.076 (0.0014)	1.00 (0.0281)

As can be seen from the graph, the precision of the measures increases by increasing number of events for all risk profiles. In fact, the interquartile range of relative differences in all performance measures decreased by increasing the number of events at differing levels of risk profile.

Additionally, it seems that only the number of events affect the precision of performance measures.

To sum up, according to the interquartile range, the precision of performance measures seemed acceptable when the number of events were at least 75. At this number of events, the percent root mean squared differences of calibration slope was 15, of C statistic was 3.5, and of Brier score was 1. Thus, we suggest that there should be at least 75 events when validating a risk model using independent binary outcome. The percent root mean squared differences is calculated by squaring the subtraction of the estimated measure from the reference measure and multiplying the result by 100.

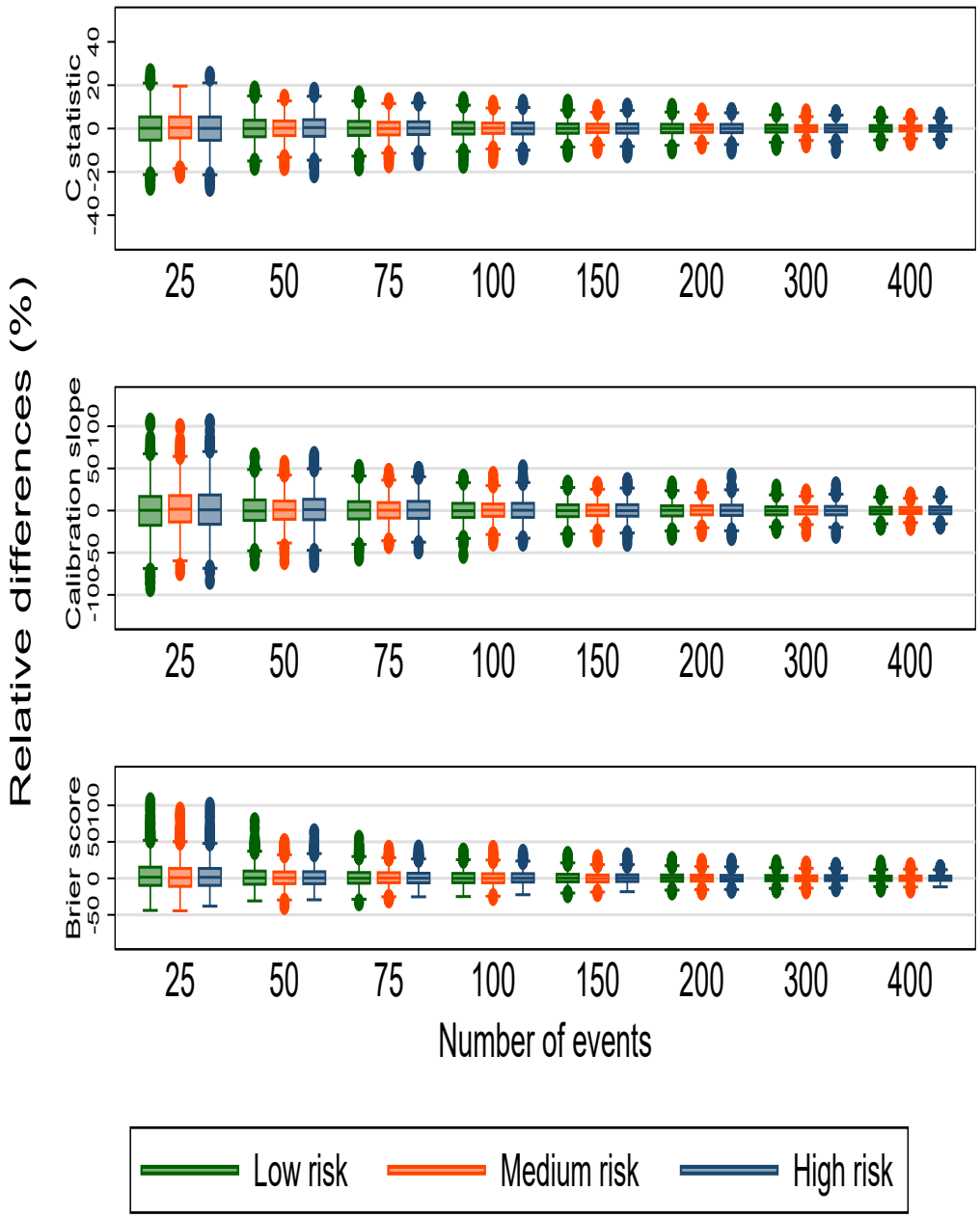


Figure 4.3: Percent relative differences in the estimated performance measures in the validation samples by number of events for different risk profiles based on 2000 simulated data. The standard error range of simulation for the C statistic, calibration slope and Brier score among all number of events and risk profiles were (0.0001, 0.0012), (0.0006, 0.0059) and (0.00004, 0.00031), respectively.

.

4.5 Conclusion

The objective of this chapter was to analytically investigate the relationship between the precision of the performance measures and the sample size, and to ascertain the required number of events when validating a correctly-specified risk prediction model with independent binary outcome.

With regard to the importance of the risk prediction models in medical settings, other than these models should be developed carefully, they should be validated with great care such that they could satisfactory perform in other new settings. Moreover, risk models validated using small datasets cannot be trusted to be used in practise (Harrell, 2001, Vergouwe et al., 2002, 2005, Steyerberg, 2009). Therefore, taking account of the issue of sample size is crucial when validating a risk model in medical setting.

It is suggested to use at least 100 events when validating a risk model (Harrell et al. (1996); Vergouwe et al. (2005); Peek et al. (2007); Collins et al. (2014)). However, none of the studies investigated the relationship between the risk profile and precision of the performance measures in the light of the number of events.

Thus, we first analytically showed that the precision of the calibration slope, D statistic and C statistic have a nonlinear relation with the sample size.

We also conducted a case study to empirically assess this dependency. The results were in line with the analytical observations. A further question arose over the need for a different number of events in validation datasets if the outcome prevalence differed from what was observed in the heart valve surgery data. A simulation study was conducted to address this issue.

The results of our simulations showed that the required number of events is not different for different risk profiles. We found that with at least 75 events a researcher can validate a model satisfactorily .

Chapter 5

Sample Size Requirements for Developing a Risk Model using Binary Clustered Data

We have already studied the required number of events to develop a risk prediction model using datasets in which the assumption of the independence of observations holds (see chapter 3).

However, data may be of the clustered structure (Twisk, 2006). This is the case in a lot of medical research. For example: patients who are registered in their local surgery are clustered within surgeries; or measures repeatedly taken from the same patient or variable over long periods of time in longitudinal studies (Robertson et al., 2013) are clustered within patients or variables; as well as the data from different studies of the same sort evaluated in a meta analysis, are clustered within studies (Thayyil et al., 2010).

In clustered data, observations inside the same cluster are correlated in some features compared to those from different clusters (Beitler and Landis, 1985, Hedeker et al., 1991, Kreft and Leeuw, 1998, Ambler et al., 2005, Robertson et al., 2013). For instance, the results of treatments for patients from the same surgery due to the treatment policy in the surgery, or due to the unique characteristic of each patient in longitudinal studies which appears in each measurement, or due to the fact that each study has been conducted by a certain method which might somewhat differ from other studies. The treatment result for patients from different surgeries, or the measurements of each

patient, or the resulting information from each study are correlated within their own clusters.

Thus, as the assumption of independence of observations is violated in clustered data, the standard EPV calculation may no longer apply, nor the standard modelling techniques. That is because those techniques cannot accommodate the dependence among subjects within the same cluster and must be adjusted for the presence of clustering. However, one might choose to fit a standard risk model with fixed effects for clusters (see section 2.3.2), in which case one could use a standard EPV calculation.

The objective of this chapter is to determine the required sample size when developing a risk prediction model using clustered binary data. The chapter includes the following sections. The literature is reviewed in section 5.1. Section 5.2 describes the dataset used in this chapter and Chapter 6. Section 5.3 is devoted to a case study conducted to give an insight into the issue of events per variable when with clustered binary data. In Section 5.4, a simulation study in which the issue of EPV is further investigated is reported. In Section 5.5, the recommended sample size was compared with the one used in the common practice. Finally, Section 5.6 discusses findings and provides some recommendations.

5.1 EPV in developing a clustered risk model: A review

In general, the usual approach to calculate sample size for clustered data is one that uses standard methods to find the required sample size, and then multiplies the results by a design factor (Simpson et al., 1995), where design factor is given by $1+(m-1)\times \texttt{ICC}$ in which m and `ICC` are average cluster size and 'intra cluster' correlation coefficients, respectively. This approach was developed to use when the primary interest was significant testing or precision of estimated regression coefficients to estimate the amount that the standard errors of parameters are underestimated (Kish, 1965). It is worth noting that the standard modelling approach assumes that observations are independent and that each observation adds to our information from the study population. However, when observations are clustered within centres, those from the same cluster provide similar information from the studied population. Thus, when choosing the modelling approach, it is important to take clustering into account in order to obtain the correct standard errors for the estimated parameters.

This approach may not be appropriate to take when developing a risk prediction model as there is less interest in individual covariate effects. Rather, the main focus is likely to be quantifying the ability of the model to predict outcomes for future patients, or to separate patients from different groups, as Copas (1983) observed that "a good model may include variables which are 'not significant', exclude others which are, and may involve coefficients which are systematically biased". Thus, the best risk prediction model may not be produced by only basing sample size decisions on the significance or unbiased estimation of model coefficients.

In literature, there are some simulation studies which have been conducted to investigate the required sample size to accurately estimate fixed- and random-effect coefficients and their corresponding standard errors (SE) when using multilevel linear regression models in the context of continuous outcome (Maas and Hox, 2004, 2005, Bell et al., 2008, 2010). These papers are reviewed here since the estimated coefficients of the risk model are linearly combined to constitute the linear predictors which are directly or indirectly used in calculation of performance measures.

Maas and Hox (2004, 2005)

They conducted a simulation study to determine the influence of different cluster sizes on the accuracy of fixed- and random-effect parameters estimations and their standard errors in multilevel linear regression models using continuous outcome. The fixed-effect parameters referred to the intercept and regression slopes at two levels and the random-effect parameters were error terms for subject level and cluster level. Maas and Hox assessed the number of clusters, cluster size and ICC.

Maas and Hox (2004, 2005) discovered that the estimation of both fixed- and random-effect parameters have negligible bias (less than 5%) in all scenarios. They also found that 95% confidence interval coverage for estimated parameters is influenced by the cluster sizes less than the number of clusters, and the estimation of standard errors for the random-effect parameters were biased downward when there were 50 clusters or fewer of sizes of 30 or less in the datasets. Moreover, they reported that ICC did not have an effect on 95% confidence interval coverage

They did not investigate the effect of those three scenarios on accuracy and precision of estimation and predictions of a multilevel logistic regression model.

Moineddin et al. (2007)

They conducted a simulation study to examine the effect of sample and cluster sizes on the estimation of the fixed- and random-effect parameters and their standard error in two-level logistic regression models. In their simulation study, they examined all the scenarios which Maas and Hox (2005) have studied plus outcome prevalence.

Moineddin et al. found that while the estimation bias of fixed-effect parameters was very low (less than 4%) when the size of the clusters increased (to 30 observations) as well as the number of clusters (to 50 clusters), the standard error of random-intercept was consistently underestimated when with a cluster size of at least 30, regardless of the number of clusters (the bias was less than 4%, though, when with 100 clusters of size 50). Furthermore, they reported that bias was smaller at low outcome prevalence for both fixed- and random-effect parameters. Above all, Moineddin et al. (2007) confirmed that ICC had no effect on the estimation of the fixed-effect parameters, although relative bias for the random-intercept decreased by increasing ICC.

Moineddin et al. (2007) also discovered that while ICC had no effect on 95% confidence interval coverage for the fixed-effects parameter, the 95% confidence interval coverage decreased by increasing ICC for the random-intercept parameter. What is more, they verified that while the standard error of fixed-effect parameters was almost always accurate in all scenarios, the standard error of random-effect parameters was always biased (the convergence of 95% confidence interval was lower than nominal level in all scenarios).

Moineddin et al. (2007) recommended using at least 50 clusters of size 50 when developing a two-level logistic risk model in order to produce valid (accurate and precise) estimation for parameters. They also recommended that low outcome prevalence requires larger cluster sizes in order to have at least one event in each cluster.

They did not evaluate the effect of listed scenarios on predictions of multilevel logistic models.

Paccagnella (2011)

They conducted a Monte Carlo simulation study to assess the accuracy of regression coefficients estimates and their standard errors when using a two-level random-intercept

logistic model. As in previously reviewed studies, they investigated the following scenarios; the number of clusters, (unequal) cluster sizes and ICC.

They reported that the estimation of fixed-effect parameters were accurate when there were at least 10 clusters in the dataset. They also found that the estimates of standard errors of fixed-effect parameters were accurate when there were at least five clusters in the dataset; however, the estimates of standard errors of random-effect parameters were never accurate. Paccagnella (2011) also reported that the estimates of fixed- and random-effect parameters were not influenced by ICC.

Overall, the results of this work by Paccagnella (2011) agree with the findings of Maas and Hox (2004, 2005) and Moineddin et al. (2007): that the accurate estimates for fixed-effect of fixed-effect parameters can be achieved at even 10 clusters, but the standard error of those estimates were too small and standard error of random-effects parameters were always underestimated. These researchers did not study the effect of the number of clusters, cluster sizes and ICC on prediction.

Recommendations for estimation of regression coefficients

Summing up the recommendations, there should be at least 50 clusters of size at least 30 observations in the dataset when developing a multilevel linear regression model using continuous outcome. This allows us to achieve an accurate estimation for fixed- and random-effect parameters and a precise estimate for fixed-effect regression coefficients. However, a studied dataset should consist of at least 50 clusters of size at least 50 observations to accomplish the same aim when developing a multilevel logistic regression model using binary outcomes.

Nevertheless, there is still an unsolved problem. How many observations are required to obtain accurate predictions when developing multilevel regression model using either continuous or binary outcome? To answer this question, in a small part of their study, Bouwmeester et al. (2013) examined the effect of sample size on predictions using clustered binary outcome. This study is reviewed below.

Bouwmeester et al. (2013)

They conducted a simulation study to compare the ability of a random-effect logistic model (in producing accurate median predictions (p_0) or cluster-specific predictions (p_u) (see section 2.3)) with the ability of standard logistic regression models (in producing

accurate standard predictions (see section 2.3)) in both development and validation datasets. The number of clusters and ICC were the scenarios of study. Bouwmeester et al. assessed both apparent and external performance (see section 2.4) of the models in terms of discrimination and calibration, using the C statistic, calibration slope and calibration-in-the-large (see section 2.5).

Bouwmeester et al. (2013) reported that models developed using datasets with a small number of clusters (5 clusters) performed worse compared to those developed using datasets with a medium number of clusters (50 clusters) regardless of ICC. However this issue of sample size was not the main focus of the publication, and so investigation of this aspect was limited in scope.

Wynants et al. (2015)

In their simulation study, Wynants et al. (2015) looked into the effect of EPV on predictive performance of the random-intercept logistic models, as well as the estimation of their coefficients. They evaluated the following scenarios; the number of clusters, cluster size, ICC and EPV.

Wynants et al. only studied median predictions, (P_0) (see section 2.3) and compared overall and within-cluster calibration and the C statistic in simulated data with those from source population. Note that the overall measures evaluate the performance of the models in population level, and within-cluster measure is a weighted combination of cluster-specific measures which quantify the performance of the model in cluster level.

From their study, Wynants et al. (2015) reported that the largest estimation bias was in the lowest EPV at all ICCs. For example, at ICC=20% the estimation bias of their example coefficient was about 10% at EPV=5 compared to 2% at EPV=50. They also found that the standard error of random-intercept was often underestimated, but the amount of underestimation was less at large EPV. For example, at ICC=20% bias was -15% at EPV=5 and it dropped to -5% at EPV=50. They also noted that the values of both the overall and within-cluster C statistic and calibration slope went up by increasing EPV at all ICC levels. For example, at ICC=20% bias in within-cluster C statistic was around 3% at EPV=5 compared to about 0 at EPV=50, and bias in the calibration slope was about 28% at EPV=5 compared to 3% at EPV=50. We also found this relationship between EPV and performance of the risk model in our study, reported at section 3.4 of this dissertation. Wynants et al. also reported that models

developed in larger samples perform well, holding EPV constant. Of note, we also found similar results to their last finding in our study by studying different outcome prevalence scenario, reported at section 3.4.3 of this dissertation.

They recommended using at least 10 EPV (calculated by incorporating the number of fixed- and random-effect parameters) to fit a predefined prediction model when using clustered binary outcomes.

Wynants et al. only investigated the issue of the number of events using median predictions.

EPV summary

To sum up, there is no consensus on how much sample size is required to develop a multilevel risk model with accurate or nearly accurate predictions. It has been recommended to use 50 clusters of size 30 to develop a two-level linear risk model (Maas and Hox (2004); Maas and Hox (2005)) and to use 50 clusters of size 50 to develop a two-level logistic risk model (Moineddin et al., 2007) with accurate estimation of fixed- and random-effect parameters, as well as accurate estimation of fixed-effect standard error. All of those studies only addressed the estimation of fixed- and random-effect parameters and their standard errors and did not investigate the impact of sample size on predictive performance of the models.

There is not enough study on how much data is required to develop a reliable multilevel risk model for prediction. In an independent binary setting, it has been suggested to use EPV of at least 10 when developing a risk prediction logistic model (Harrell et al., 1984). Wynants et al. (2015) also suggested to use EPV of at least 10 when developing a random-intercept logistic model with reasonable performance. However, they only studied median predictions obtained from only incorporating the estimates of fixed-effects parameters. Therefore, the following questions are yet to be answered. Are cluster-specific predictions (p_u) and/or marginal predictions (p_m) affected by EPV? What other factors apart from EPV are important to accomplish an acceptable performance from random-intercept logistic models?

Therefore, this study was conducted to ascertain the required number of events to develop a risk prediction model with acceptable performance. We illustrated the dependency of two-level random-intercept risk prediction model performance on EPV in our case study. This was conducted using three common performance measures;

calibration slope, C statistic and Brier score. We also performed a simulation study examining different scenarios, including various number of clusters, cluster size, ICC and EPV. We evaluated the performance of the random-intercept logistic risk model using median prediction and cluster-effect predictions. We also examined the performance of the marginal models (see section 2.3) as well as fixed-effect logistic models in the studied scenarios in relation to EPV.

5.2 Data

Heart valve surgery data was used to conduct studies in chapter 5 and 6. This data will be described now.

The entire heart valve surgery dataset was employed (N=32,839) in chapter 5 and 6. The dataset includes the information of patients who are clustered within health centres. There were 30 clusters with the cluster sizes ranged from 298 patients to 2007 patients (cluster size mean and standard deviation were 1279 and 404, respectively). The overall outcome prevalence was 6.36%. The outcome prevalence in clusters ranged from 0.67% to 12.43% (average prevalence over clusters was 6.03 and the standard deviation of it was 1.93). The intra-cluster correlation (ICC) coefficient was calculated using the method of analysis of variance (ANOVA) (Chakraborty et al., 2009) to be 6%. A set of seven predictors, all with prognostic information, were chosen to be used in this study (see section 3.2).

The Maximum Likelihood (ML) random-intercept model logistic regressions with two levels was fitted on the entire dataset (N=32,839) and used to investigate the importance of each predictor. The process was conducted as follows. For each predictor, two models was fitted, one with all predictors included and one excluding the predictor. The decrease in χ^2 ($\Delta\chi^2$) was then calculated for each predictor (Table 5.1).

The variance parameter of the random effects, σ_u^2 was estimated as 0.05. This corresponds to an $\texttt{ICC} = \sigma_u^2/(\sigma_u^2 + \frac{\pi^2}{3}) = 0.015$, indicating a weak correlation between patients within a centre, after accounting for the fixed predictors. It transpired that, fitting the random-intercept model for this multi-centre data using standardised variables, there were four moderate (the estimated coefficients were between 0.2 and 0.5) and three weak (the estimated coefficients were less than 0.2, inclusive) predictors.

Table 5.1: Importance of the predictors in a random-intercept model (estimated using maximum likelihood) fitted for the heart valve data.

Variables	Category	D.F.	$\hat{\beta}^*$	$\Delta\chi^2$
Operative Priority	elective	2		416.03
	urgent		0.264	
	emergency		0.351	
Age		1	0.476	171.39
Operation Sequence	First	1		122.74
	Second & more		0.266	
Renal failure	no	1		118.22
	Cr>200 & dialysis		0.202	
Ejection Fraction	good (>49)	2		108.14
	fair (30-49)		0.110	
	poor (<30)		0.208	
Valve Operation	aortic	2		64.56
	mitral		0.149	
	aortic+mitral		0.182	
Concomitant CABG Surgery	no	1		54.37
	yes		0.214	
Full model		10		"$LR\chi^2$= 1631.65 "

$\hat{\beta}^*$ the estimated coefficient for standardised variables.
$\Delta\chi^2$ denotes the decrease in χ^2 statistic for the model when the predictor is omitted and the model refitted.
D.F. denotes the number of parameters included in the model for that predictor.
Prevalence of category refers to the percentage of subjects with the condition.

5.3 Case study

From the previous section the available literature is not enough to determine the required sample size to develop a multilevel risk prediction model using clustered binary outcomes.

Therefore a case study was conducted to understand how the accuracy of predicted risks produced using random-intercept regression models can be affected by EPV. This case study was carried out using the heart valve surgery data (see section 5.2). All types of predictions were studied (see section 2.3).

5.3.1 Method

To conduct a case study, the first five years from the heart valve surgery data was used for development, and the rest for validation. For each EPV, the following method was repeated 200 times. Firstly, the size of development dataset was varied to achieve the desired EPV (5, 10, 20, 50 and 116) by separately sampling without replacement from events and nonevents. These numbers of EPVs included all recommended EPVs in the previous studies. The EPV was calculated by dividing the number of events into 10 (the number of variables in the model). Secondly, a random-intercept logistic regression model was fitted on development datasets of varying size and used to produce three types of predictions (cluster-specific (P_u), median (p_0) and marginal (p_m) predictions (see section 2.3)). Finally, the overall performance measures (C statistic, D statistic, calibration slope and Brier score) were calculated on validation data for each type of prediction.

We only use overall measures in both this chapter and chapter six as overall and within-cluster performance measures (see section 2.6) perform similarly (Wynants et al. (2015), Bouwmeester et al. (2013)).

5.3.2 Results

Table 5.2 presents the mean value (and standard error) of performance measures by EPV and types of predictions over 200 samples where types of predictions were cluster-specific (p_u), median (p_0) and marginal (p_m) predictions.

From the table, the performance of random-intercept logistic models were affected by EPV. The discrimination ability of the models decreased as EPV decreased according

to C statistic and D statistic. The C statistic obtained using any type of predictions dropped from 0.77 for EPV=116 to 0.73 for EPV=5. Moreover, the D statistic obtained using p_0 or p_m fell from 1.72 when EPV was 116 to 1.49 when EPV was 5, but this statistic decreased from 1.75 for EPV=116 to 1.51 for EPV=5 when obtained using p_u.

Furthermore, models underfitted in higher EPV and overfitted in lower EPVs, according to the calibration slope. This is discussed further shortly. The calibration slope dropped from 1.18, 1.20 or 1.19 for EPV=116 to 0.89, 0.90 or 0.90 for EPV=5 when using p_0, p_m and p_u, respectively.

In addition, with regard to Brier score, the overall performance of the models was slightly affected by EPV regardless of the type of predictions. This is due to low outcome prevalence in our dataset.

Furthermore, the results for all types of predictions were similar, since the clustering effect in heart valve surgery data was weak.

Table 5.2: Mean value and standard error of each performance measure by EPV and type of prediction obtained from a random-intercept logistic model where types of predictions were cluster-specific, median and marginal predictions.

			Performance quality measures in validation data					
		EPV	**C(SE)**	**D(SE)**	**CS(SE)**			**BS(SE)**
					naive	updated	optimism corrected	
Types of prediction		116	0.77	1.72	1.18	0.98	0.96	0.0499
		50	0.76 (0.002)	1.7 (0.011)	1.17 (0.046)	0.98 (0.003)	0.96 (0.002)	0.05 (0.0001)
	P_0	20	0.76 (0.005)	1.66 (0.033)	1.12 (0.081)	0.98 (0.004)	0.95 (0.004)	0.0502 (0.0002)
		10	0.75 (0.009)	1.6 (0.066)	1.05 (0.114)	0.98 (0.005)	0.94 (0.006)	0.0505 (0.0003)
		5	0.73 (0.043)	1.49 (0.135)	0.9 (0.16)	0.99 (0.04)	0.98 (0.04)	0.0511 (0.0006)
		116	0.77	1.72	1.2	0.98	0.97	0.0499
		50	0.76 (0.002)	1.7 (0.011)	1.18 (0.047)	0.98 (0.002)	0.96 (0.003)	0.05 (0.0001)
	P_m	20	0.76 (0.005)	1.67 (0.034)	1.13 (0.081)	0.98 (0.004)	0.95 (0.004)	0.0502 (0.0002)
		10	0.75 (0.009)	1.6 (0.066)	1.06 (0.114)	0.98 (0.005)	0.95 (0.006)	0.0505 (0.0003)
		5	0.73 (0.043)	1.49 (0.135)	0.91 (0.162)	0.99 (0.03)	0.98 (0.01)	0.0511 (0.0006)
		116	0.77	1.75	1.19	0.99	0.98	0.0498
		50	0.77 (0.001)	1.74 (0.011)	1.18 (0.045)	0.99 (0.0009)	0.99 (0.002)	0.0498 (0.0001)
	P_u	20	0.76 (0.005)	1.69 (0.036)	1.13 (0.08)	0.99 (0.004)	0.99 (0.004)	0.05 (0.0002)
		10	0.75 (0.009)	1.62 (0.066)	1.06 (0.114)	0.99 (0.006)	0.98 (0.006)	0.0504 (0.0003)
		5	0.73 (0.044)	1.5 (0.136)	0.91 (0.16)	0.98 (0.02)	0.99 (0.04)	0.051 (0.0006)

P_0: median prediction. P_m: marginal prediction. P_u: cluster-specific prediction. C: C statistic. D: D statistic. CS: calibration slope. BS: Brier score. SE: standard error of the measure over 200 simulations. Updated refers to which the calibration slope was obtained from predictions of the recalibrated-updated model. Optimism corrected refers to the calibration slope which was obtained from predictions of the bootstrap-optimism-corrected risk model.

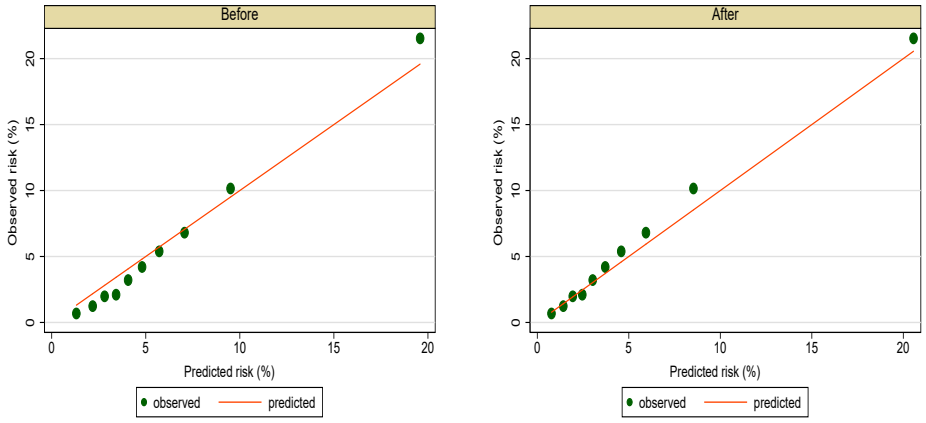


Figure 5.1: Observed and predicted log odds of mortality in validation data (n=16,160) before (left plot) and after (right plot) updating. Perfect calibration is represented by the red line through the origin.

Further analysis

We have used the original data to conduct the case study and as it can be seen in Table 5.2, models are misspecified at large EPVs (calibration slope was greater than one). In practice, in such situations, one should look for the reasons which caused the model to be underfitted (or overfitted) and try to fix it. Therefore, we inspected the data used for development and validation of the prediction model. The results of our inspection for EPV of 116 are reported here.

As a first step, we examined the outcome prevalence in both development and validation datasets. The outcome prevalence in the first five years of heart valve surgery data which was used to develop the risk prediction model was 6.97%, but it was 5.73% in validation data. We plotted the observed risk against predicted risk for a first impression of validity of the risk prediction model for the last four years (2000-2003) of heart valve surgery data (Figure 5.1 - left plot). We noted that the observed mortality rate in high-risk group patients was larger and in low-risk group patients somewhat lower than those predicted. This may be attributed to the slight difference in outcome prevalence.

The validity was further assessed by fitting a random-intercept model in both development and validation datasets and comparing the regression coefficients between

Table 5.3: Differences in estimated Logistic regression coefficients in the development and validation data and P-value for the test for "whether the effect of predictors in validation data differ significantly". All the results are presented for only sample with EPV of 116 .

Variable	**Categories**	**differences* in coefficients**	**Is the effect in val. different? (P-value)**
Age		- 0.0073	No (0.679)
Valve Operation	Aortic		No (0.625)
	Miteral	0.0768	
	Aortic&miteral	- 0.0634	
Concomitant CABG Surgery	No		No (0.875)
	Yes	- 0.0579	
Renal Failure	No		No (0.186)
	Cr>200&dialysis	-0.3321	
Ejection Fraction	Good		No (0.526)
	Fair	- 0.0902	
	Poor	- 0.2595	
Operative Priority	Elective		No (0.411)
	Urgent	- 0.1765	
	Emergency	- 0.1416	
Operation Sequence	First		No (0.293)
	Second&more	- 0.2430	

* estimated coefficient in development data minus estimated coefficient in validation data. val.: validation data

them (Table5.3). The coefficients when EPV was 116 are reasonably similar, although the coefficient of Mitral Operation is somewhat smaller in validation data and those of the rest of the variables larger.

As can be seen from Figure 5.1 (left plot) and Table 5.3, there is a clear need for updating the fitted model. To do so, we implemented the updating method for the standard logistic regression model suggested by Steyerberg (2009) to update the random-intercept logistic regression model which was developed on the entire development data. The steps taken for this purposes are described here.

The cluster-specific linear predictor (η_1) was obtained for patients in validation data using the model developed on development dataset. To recalibrate the model, the intercept and overall calibration slope were updated by fitting a new random-intercept logistic model on the validation data using η_1 as the only predictor. Thereafter, the updated cluster-specific predictor was used to produced predictions and to compute the overall calibration slope. This was performed for cluster-specific linear predictors (obtained incorporating the effect of both fixed- and random-effect variables). Then, median and marginal linear predictors were derived from those. The process was conducted for all samples with EPV of 5 to EPV of 116.

We plotted the observed risk against expected risk for an updated model when EPV was 116 (Figure 5.1 (right plot)). As noted, performance of the prediction model after updating slightly improved in low- and high-risk groups but slightly deteriorated in medium-risk group.

Furthermore, to verify if any of the coefficients needs to be re-estimated in validation data, we used validation dataset to update (the intercept and slope of) the developed risk prediction model and also test whether the predictors had an effect that was clearly different in validation dataset. We performed likelihood ratio tests of model extensions in a forward stepwise fashion and extended the revised model until all differences in predictive effects have $p > 0.05$ for each predictor. We did not find statistically significant deviations from overall recalibrated values (see Table 5.3) which means there is no need for re-estimation of individual coefficients.

We reported the effect of updating the model on the values of calibration slope (Table 5.2). As can be seen in Table 5.2, while the average value of the naive calibration slope obtained using p_0 or p_m were 1.17 and 1.18, respectively, dropping to 0.98 after updating when EPV was 116. The average value of it obtained using p_u was 1.19

dropping to 0.99. We updated all models fitted in development datasets of varying size and recalculated the calibration slopes. From the table, the calibration slope for EPVs of 50 to 10 obtained using p_u stayed the same as it was for EPV of 116, but the empirical standard error of the calibration slope increased by decreasing EPV. The calibration slope for EPV of 5 was slightly different than it for EPV of 116. However, its empirical standard error was larger. The calibration slope obtained using p_0 and p_m changed in a similar fashion to that just described by decreasing EPV.

Moreover, as we used validation datasets for both updating the models and computing the calibration slope, the apparent estimate of calibration slope may be severely optimistic. Therefore, we studied the internal validity of all the updated risk prediction models using the bootstrap method. To do so, each developed model was updated and its related calibration slope was calculated using 200 bootstrap validation samples. The optimism-corrected calibration slope for that model was obtained by averaging calibration slope values over 200 bootstrap samples. Table 5.2 displays the results of the study. From the table, while the optimism-corrected calibration slopes obtained using p_0 and p_m were somewhat smaller than updated ones, those obtained using p_u were quite similar to the updated ones.

5.3.3 Discussion

This case study showed that the performance of the random-intercept risk prediction models were affected by EPV. The performance of risk prediction model was good when the EPV was large compared with those when the EPV was small, regardless of the type of predictions. Moreover, the results for median, marginal and cluster-specific predictions were similar because the clustering effect was weak in our data.

Further investigation on the result of simulation showed that the distribution of linear predictors for patients with the event of interest overlapped largely compared with that for patients without event at lower EPV. The separation between $\hat{\eta}_u^{(0)}$ and $\hat{\eta}_u^{(1)}$ by EPVs among 200 samples illustrated in Figure 5.2, where $\hat{\eta}_u^{(0)}$ and $\hat{\eta}_u^{(1)}$ are linear predictors, corresponds to patients who survived and those who died, respectively, and was obtained using both fixed- and random-effects of random-intercept models. For each panel (each EPV), the overlap of all 200 $\hat{\eta}_u^{(0)}$ and $\hat{\eta}_u^{(1)}$ was overlaid to judge the separation by decreasing EPV value. Furthermore, the separation between $\hat{\eta}_m^{(0)}$ and $\hat{\eta}_m^{(1)}$ and between $\hat{\eta}_0^{(0)}$ and $\hat{\eta}_0^{(1)}$ appeared to be approximately equal to those just described

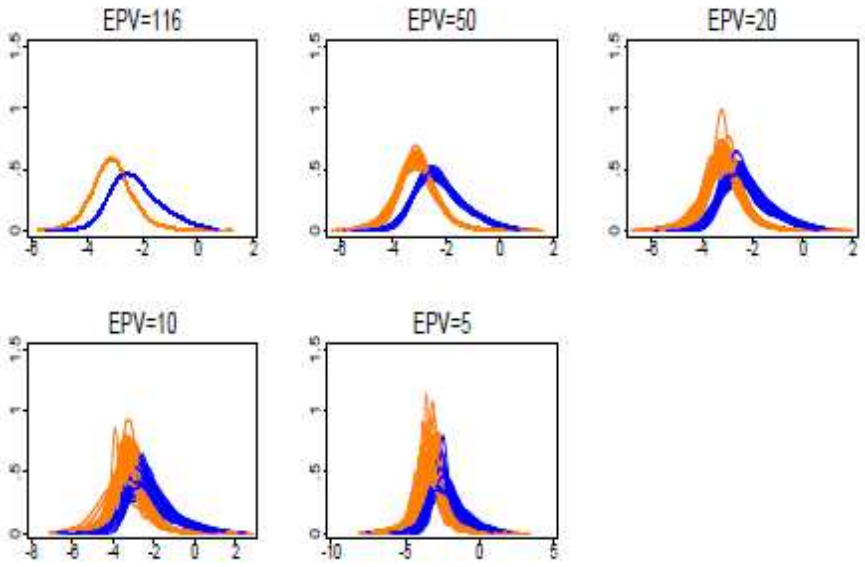


Figure 5.2: Separation between $\hat{\eta}_u^{(0)}$ and $\hat{\eta}_u^{(1)}$ by EPVs. $\hat{\eta}_u^{(0)}$ and $\hat{\eta}_u^{(1)}$ are linear predictors corresponding to patients who survived and those who died, respectively. The orange (and blue) lines correspond to the distributions of linear predictors for patients with (and without) event of interest over 200 simulated datasets.

by decreasing EPV, although the results are not shown here. This is expected because the clustering effect in heart valve surgery data used for this case study is not strong.

The questions that we are also interested in addressing in this chapter are whether the EPV is the only factor that affects the performance of the multilevel risk model, more specifically a random-intercept logistic model, or whether there are other factors alongside EPV, such as a different number of clusters, various cluster sizes or ICCs, which can change the quality of performance. Several simulation studies were conducted and various characteristics of the heart valve surgery data modified such as ICC, number of clusters and cluster size. Then the performance of the models were quantified. That brings us to the next section in which the quality of the random-intercept and fixed-effect and marginal (standard) models were studied using diverse scenarios with samples

including different EPVs.

5.4 Simulation study

The case study illustrated how the performance of the risk models are affected by EPV. Nevertheless, it is not obvious whether the EPV effect can disappear by changing some of the characteristics of clustered data. To understand how the accuracy of performance measures can be influenced when developing risk models using clustered datasets of different features and varying EPV, a number of simulation studies were conducted. These simulations were based on heart valve surgery data. In these studies, different factors were investigated; EPV, ICC, number of clusters, cluster size, type of predictions and type of the model. Details are elaborated here.

5.4.1 Overview

Simulation studies were conducted to investigate the influence of EPV on performance of logistic risk models when fitted using clustered binary data. To do so, a source population was produced and used to develop the true model using seven predictors (10 variables). True linear predictors were derived and employed to study the performance of risk models in relation to EPV in the presence of the following scenarios.

§ *Intra-cluster-correlation coefficient* (ICC): to learn whether the EPV requirements should change with increasing ICC, the intra-class correlation (ICC) coefficient was modified. Four levels were considered; 0%, 5%, 10% and 20%. These values represent four common clustering levels in multicenter prediction research (Adams et al., 2004).

§ *Number of clusters*: to evaluate the influence of EPV on the performance of multilevel risk prediction models in the presence of various numbers of clusters, different numbers of clusters were examined; 7, 14, 21, 28, 30, 42 and 70.

§ *Cluster sizes*: to examine the impact of EPV on the accuracy of predictions obtained from multilevel risk prediction models in correlation with different cluster sizes, various cluster sizes were studied. The average cluster sizes were 7, 14, 21, 28, 30, 42 and 70.

§ *Modelling approach*: to examine whether the effect of EPV on the performance of each type of model differs, three very common modelling approaches were considered; random-intercept, fixed-effect and marginal (standard) logistic regression (see section 2.3).

§ *Type of predictions*: three types of predictions were studied; median (P_0), cluster-specific (P_u) and marginal (P_m) predictions (see section 2.3).

§ *EPV*: the following values were studied depending on the available number of events in samples; 5, 10, 15, 20, 30 , 50. The EPV calculation was based on the number of parameters in the random-intercept model (11 parameters; 10 fixed- and one random-effect parameters).

§ *Sub-studies*: to study all factors thereof, three sub-studies were designed; a) the number of clusters and cluster sizes were fixed, but outcome prevalence altered to change EPV, b) the number of clusters and outcome prevalence were fixed, but the cluster sizes and EPV were changed, and c) the cluster sizes and outcome prevalence were fixed, but the number of clusters and EPV were varied.

For each combination of factors, the following steps were carried out 400 times. Firstly, the required number of clusters was sampled without replacement from the source population. Secondly, the resulting sample was divided into development and validation; 80%:20%. Thirdly, the size of the development sample was altered by separate sampling without replacement of the desired number of events and nonevents (the proportion of events and nonevents were determined by the required EPVs using equation 3.4.3). Lastly, risk models were fitted on development samples, and all types of predictions (p_u, p_m and p_0) of death for all patients in validation samples were obtained. These were used to measure the performance of the models in related validation samples and compare them with their true values by calculating the percent relative differences (see section 3.4.1). Note that the true values were obtained using cluster-specific predictions, which were produced using a true model. Three overall performance measures were used; calibration slope, C statistic and Brier score.

We ought to illustrate the sampling procedure with the use of an example. To do so, let the source population have 5 clusters of sizes 50, 45, 80, 40 and 55. In addition, assume that one wishes to study samples consisted of 3 clusters and thus chooses 3 clusters without replacement from the source population. To proceed with this example, let the hypothetical clusters be of 80, 55 and 45 sizes. The selected clusters, thus, make a sample of size 180 observations. One will divide the sample with portions of 80% (144 observations) for development and 20% (36 observations) for validation and follow the simulation protocol thereof.

5. SAMPLE SIZE REQUIREMENTS FOR DEVELOPING A RISK MODEL USING BINARY CLUSTERED DATA

The studied scenarios and the features of source dataset are summarised in Table 5.4. Now, the details of the simulation study are explained.

Table 5.4: The features of studied scenarios (in the development datasets) and the source datasets.

	Source data	Sample				Model		
	ICC(%)	J	$\bar{n}_j$ $(\overline{min}, \overline{max})$	N	p	K	type	EPV
(1)	0, 5, 10, 20	30	30 (12, 62)	900	0.06	11	RE, FE	5
	0, 5, 10, 20	30	30 (12, 62)	900	0.12	11	RE, FE	10
	0, 5, 10, 20	30	30 (12, 62)	900	0.18	11	RE, FE	15
	0, 5, 10, 20	30	30 (12, 63)	900	0.24	11	RE, FE	20
	0, 5, 10, 20	70	70 (27, 164)	4900	0.02	11	RE, FE	10
	0, 5, 10, 20	70	70 (28, 163)	4900	0.03	11	RE, FE	15
	0, 5, 10, 20	70	70 (28, 165)	4900	0.04	11	RE, FE	20
	0, 5, 10, 20	70	70 (27, 163)	4900	0.07	11	RE, FE	30
	0, 5, 10, 20	70	70 (28, 164)	4900	0.11	11	RE, FE	50
(2)	0, 5, 10, 20	30	7 (1, 16)	210	0.26	11	RE, FE	5
	0, 5, 10, 20	30	14 (5, 31)	420	0.26	11	RE, FE	10
	0, 5, 10, 20	30	21 (8, 45)	630	0.26	11	RE, FE	15
	0, 5, 10, 20	30	28 (11, 59)	840	0.26	11	RE, FE	20
	0, 5, 10, 20	30	42 (18, 86)	1260	0.26	11	RE, FE	30
	0, 5, 10, 20	30	70 (32, 144)	2100	0.26	11	RE, FE	50
(3)	0, 5, 10, 20	7	30 (17, 47)	210	0.26	11	RE, FE	5
	0, 5, 10, 20	14	30 (14, 54)	420	0.26	11	RE, FE	10
	0, 5, 10, 20	21	30 (13, 59)	630	0.26	11	RE, FE	15
	0, 5, 10, 20	28	30 (12, 62)	840	0.26	11	RE, FE	20
	0, 5, 10, 20	42	30 (11, 67)	1260	0.26	11	RE, FE	30
	0, 5, 10, 20	70	30 (10, 74)	2100	0.26	11	RE, FE	50

ICC: Intra-cluster-correlation coefficient. J: the number of clusters. $\bar{n}_j$: average cluster size. N: sample size. p: outcome prevalence. K: the number of variables. type: the type of the model. RE: random-effect. FE: fixed-effect. EPV: events per variable.

(1) Fixed number of clusters and cluster size, varying outcome prevalence.

(2) Fixed number of clusters and outcome prevalence, varying cluster size.

(3) Fixed cluster size and outcome prevalence, varying number of clusters.

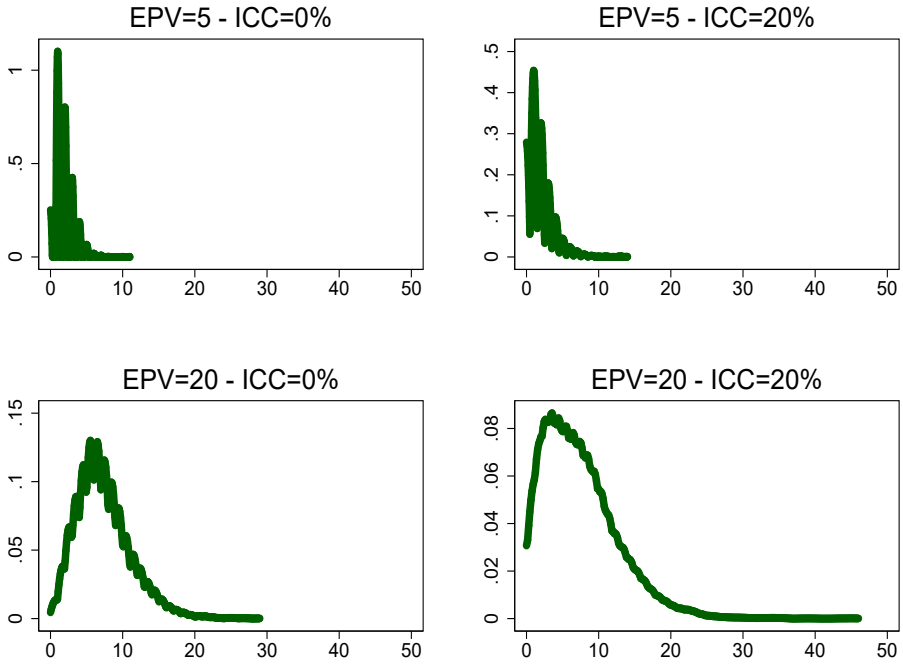


Figure 5.3: The distribution of events across clusters in five simulated datasets for ICCs of 0% and 20% and EPVs of 5 and 20. This figure corresponds to a fixed number of clusters and cluster sizes and varied outcome prevalence scenario, N=900. Note that there are 30 clusters of size average 30 in each of the simulated datasets.

Source populations

The source populations were generated as follows. The size of the heart valve surgery data was doubled to N=65,678 in order to increase the number of available clusters. We decided to use a two-step sampling procedure in order to reduce the size of our sample. That is, we first sampled the required number of clusters, then selected the required number of observations.

A set consisted of 150 λ exponentiated random values from Normal (mean 6 and standard deviation 0.33) distribution was produced. The Poison distribution with the mean of λ_i was used to determine the size of ith cluster, where $i = 1, \ldots, 150$. Note that as $\sum_{i=1}^{150} \lambda_i < 65,678$ the aforementioned approach led to 151 clusters with sizes

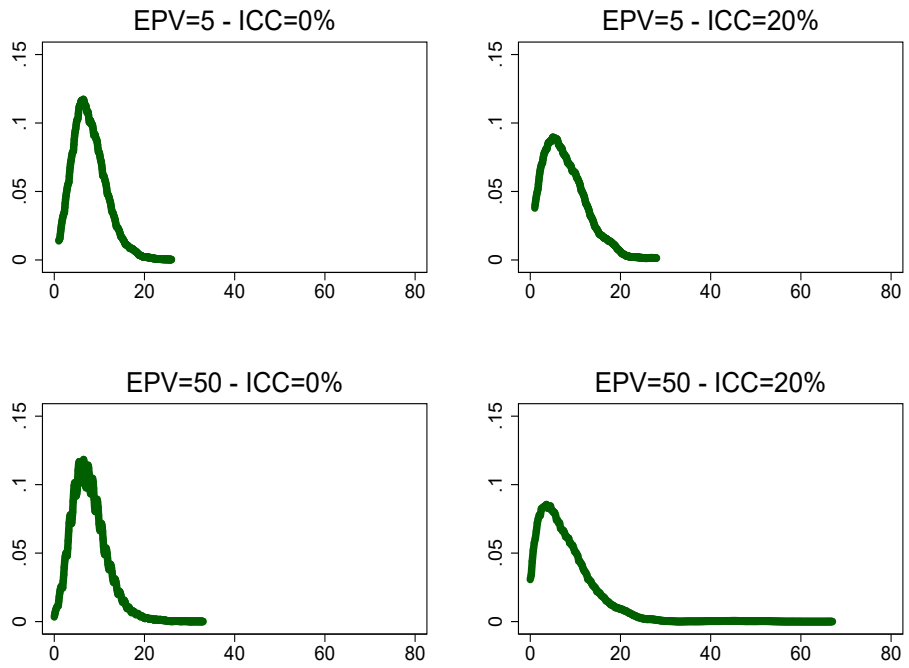


Figure 5.4: Distribution of events across clusters in five simulated datasets for ICCs of 0% and 20% and EPVs of five and 50. This figure corresponds to a fixed cluster sizes. Note that there are seven clusters of average size 30 when EPV is five and 70 clusters of average size 30 when EPV is 50 in each of the simulated datasets.

between 192 and 1143 observations. Also, note that the use of different λ in the process of producing source population ensured that the distribution of cluster sizes bear a close resemblance to the real population, in particular the heart valve surgery data.

Intra-cluster-correlation coefficient and outcome prevalence

The following approach was taken to change ICC and outcome prevalence in the source dataset. The true linear predictor (η_{true}) was used to alter ICC using

$$\eta_{new} = A + \eta_{true} + \sigma_u \times u. \tag{5.4.1}$$

where A is used to change outcome prevalence, η_{new} denotes new linear predictors, σ_u is standard deviation of random-intercept and u is a random standard normal variable. The threshold formula, given by ICC $= \frac{\sigma_u^2}{\sigma_u^2+\pi^2/3}$ (Twisk, 2006), was employed to calculate σ_u for desired ICC. Based on the formula, σ_u^2 should be 0, 0.17, 0.37 and 0.82 to achieve ICC of 0%, 5%, 10% and 20%, respectively.

The outcome prevalence in the heart valve surgery data was 6.36% and was changed to 26% for the purpose of the simulation studies. With the data in hand, the outcome prevalence of 26% ensured that all values in the simulation design setting (the number of clusters, cluster sizes, sample sizes (N)) are a whole number. A in equation 5.4.1 was set to be 1.85, 1.80, 1.76 or 1.67 to modify the outcome prevalence in the heart valve surgery data for source datasets with ICC of 0%, 5%, 10% and 20%.

Generating new outcome

For each patient, the probability of death ($p_{ij\,u}$) was computed using a new linear predictor (η_{new}) and applying the inverse logit transformation. Thereafter, Y_{ij} were simulated using Bernoulli distribution with a probability of $p_{ij\,u}$.

Distribution of events over clusters

The distribution of events within clusters in 200 simulated datasets for the smallest (5 or 10) and largest (20 or 50) number of EPV and for the smallest (0%) and largest (20%) ICC were presented in Figures 5.3 to 5.4. Figure 5.3 corresponds to the cases that the simulated datasets had 30 clusters of size 30 and Figure 5.4 relates to the cases that the simulated datasets have increasing numbers of clusters with EPV where clusters always included 30 observations.

As can be seen, there were more sparse clusters when EPV and ICC were both small compared to when EPV was low and ICC was large. However, the distribution of events followed binomial distribution (slightly right skewed) for the large EPV and small ICC and it became heavily and positively skewed when ICC increased for the same EPV (Figure 5.3). A similar pattern was observed for the scenario of 70 clusters of average size 70 and the scenario of a fixed number of clusters (figures not shown). In contrast, from Figure 5.4, the distribution of the events over clusters followed binomial distribution only when both EPV and ICC were small. It became positively skewed as either EPV or ICC or both increased.

Sample size calculation and type of model

As discussed earlier, the EPV was calculated based on the number of parameters in the random-intercept model for the following two reasons. First, one needs a large sample size to fit a fixed-effect logistic model compared to a random-intercept logistic model. For example, let us assume that there are ten predictors and 70 clusters in the dataset. Let also assume that the outcome prevalence is 3%. By definition, to have ten events per variable, one needs a sample of size about 27,000 observations to fit a fixed-effect logistic model, compared to about 3,700 observations to fit a random-intercept logistic model. That is because there are 80 variables in the former model, compared to 11 variables in the latter. Second, in practice, that is not always possible to collect a large sample.

Table 5.5 presents the actual EPV by type of model for each nominal EPV. It is assumed that there are ten predictors, and also there are either 70 or 30 clusters in the dataset. It is also assumed that the outcome prevalence is 3%.

Table 5.5: The actual EPV by the type of model for each nominal EPV. The number of predictors is ten. The outcome prevalence is 3%.

		Actual EPV	
Number of clusters	Current EPV	Marginal model	Fixed-effect model
70	5	5.5	0.7
	10	11	1.4
	15	16.5	2.1
	20	22	2.8
	30	33	4.1
	50	55	6.9
30	5	5.5	1.4
	10	11	2.8
	15	16.5	4.1
	20	22	5.5
	30	33	8.3
	50	55	13.8

As can be seen from the table, the actual EPV for the marginal models are larger than the current EPV. On the contrary, the actual EPV for the fixed-effect models

are very much smaller than the current EPV. For example, if the nominal EPV is ten, the corresponded actual EPV is 1.4 when using a fixed-effect logistic model with 70 clusters.

Sample size calculation and sub-studies

The three sub-studies are discussed in detail here. The first sub-study was designed by fixing the number of clusters and cluster sizes but altering the outcome prevalence to change the EPV. For example, the outcome prevalence was altered from 2% to 11% to change the EPV from 10 to 50 when there were 70 clusters of average size 70 in the dataset.

The second sub-study was designed by fixing the number of clusters and outcome prevalence, but changing the cluster sizes and EPV. That is, the sample size was increased with increasing EPV. For example, while the number of clusters and outcome prevalence were always 30 and 26%, respectively, the sample sizes increased from 210 to 2100 when the EPV increased from five to 50.

The last sub-study was designed by fixing the average cluster size and outcome prevalence, but altering the number of clusters and EPV. As per the previous sub-study, in this sub-study, the sample size was increasing by increasing the EPV. For example, while the average cluster size and outcome prevalence were always 30 and 26%, respectively. The number of observations in the datasets increased from 210 to 2100 while the EPV increased from five to 50.

We give an example here to explain the differences between the sub-studies when EPV was identical (see Table 5.4). Let assume that the EPV was ten, the three sub-studies were different in the following ways . In the first sub-study at $N = 4900$, the outcome prevalence was 2%, and there were 70 clusters of an average size 70 in data when EPV was ten. In contrast, in the last two sub-studies, the outcome prevalence was 26% and there were 420 observations in the dataset for the same EPV. Furthermore, while there were 30 clusters of an average size of 14 in the second sub-study, there were 14 clusters of an average size of 30 in the last sub-study when the sample size was 420 (and EPV=10).

5.4.2 Results

The results of all simulations are displayed here. The risk models converged in more than 99.99% of the simulated datasets.

Calibration slope

Figure 5.5 presents the relative differences in the calibration slope in percent by EPV and ICC over types of predictions on 400 simulations.

From the graphs, EPV had an effect on the calibration slope at all ICCs when using any type of predictions. That is, the relative differences in the calibration slope decreased by increasing EPV at all ICCs and types of predictions. The calibration slopes obtained using cluster-specific predictions produced from random-intercept models improved at most by increasing EPV at all scenarios.

Furthermore, the magnitude of the EPV effect differs depending on the ICC, type of prediction and type of model fitted on the dataset: the calibration slope obtained using any type of predictions from random-intercept, or marginal models, and those obtained using median predictions from fixed-effect logistic models, were influenced by ICC in presence of EPV. However, the calibration slopes obtained using cluster-specific predictions of the fixed-effect model were not affected by ICC.

When there were fixed number of clusters in the datasets and the size of those clusters increased by increasing EPV, the calibration slopes obtained using any types of predictions from any type of models were very similar when the EPV was greater than 15 specially at small ICCs (say, less than 20%).

Additionally, the calibration slopes obtained using cluster-specific predictions from fixed-effect models had the largest relative differences among the others, especially when there were large number of clusters in the development data, due to the small actual EPV in the dataset (see Table 5.5). For instance, the differences in the calibration slope were the largest in both sub-studies of 30 clusters of an average size of 30 and 70 clusters of an average size of 70 at almost all EPV and ICC.

The differences in calibration slopes was also large at almost all EPV and ICC for sub-study of fixed cluster size (last row of the Figure 5.5). That is probably because there were so many parameters in the fixed-effect model to be estimated for the available sample size and EPV (see chapter 3). That is, EPV was not defined based on the

number of parameters in the fixed-effect logistic model (see Table 5.5). The differences in calibration slope obtained using cluster-specific predictions of a fixed-effect model, however, decreased as EPV and ICC increased when there were fewer number of clusters in the development data for sub-study of fixed cluster-size. In fact, the percent relative differences were very similar to those obtained using the rest of the predictions in this sub-study.

Summing up, EPV needs to be at least 20 in order to develop a logistic regression model with at least 0.8 calibration slope using clustered binary outcome. One will need even larger EPV to develop a hierarchical model with a better (closer to perfect) calibration slope. EPV needs to be defined based on the parameters that appear in the fixed-effect logistic model if one prefers to fit a such model.

C statistic

Figure 5.6 displays relative differences in C statistic in percent by EPV and ICC over any types of predictions and models based on 400 simulations.

As can be seen, the relative differences in C statistic decreased as EPV increased at all ICCs and types of predictions.

Furthermore, the magnitude of percent relative differences in the C statistic for similar EPV was different depending on the ICC. That is, the percent relative differences in the C statistic obtained using cluster-specific predictions of fixed-effect models were the largest for ICC of 0%, but those obtained using all the other predictions were the smallest and decreased at the same level for the same ICC by increasing EPV.

Also, the relative differences in C statistic obtained from cluster-specific predictions of random-intercept models were the smallest at all ICCs, but those obtained from cluster-specific predictions of fixed-effect models were the second smallest at ICCs greater than 5%.

The relative differences in the C statistic obtained from various predictions differed by increasing ICC such that those related to cluster-specific predictions of random-intercept model remained the smallest at all ICC, but those related to median and marginal predictions showed increasing deterioration by increasing ICC. In fact, the C statistic obtained from median and marginal predictions were the smallest at most or all of the EPV when ICC was greater than 5%.

The C statistic obtained with median (regardless of type of the model) and marginal predictions increased rapidly by increasing EPV for the sub-study of fixed number of clusters, compared to those for sub-study of fixed-cluster size, specially when ICC was large (say, greater than 10%). That is because the size of clusters increased by increasing EPV for the sub-study of fixed number of clusters, compared to the sub-study of fixed cluster size in which the number of clusters increased by increasing EPV.

Additionally, the C statistic, holding EPV constant, calculated using marginal predictions was identical to those obtained from median predictions of fixed- or random-effect models. That is probably because these predictions either do not take clustering into account (marginal predictions) or do not use the cluster effect (median predictions).

In brief, given that the true values for 0%, 5%, 10% and 20% are 0.72, 0.74, 0.76 and 0.79, respectively, if one wishes to use cluster-specific predictions, EPV needs to be at least 20 in order to achieve a model with C statistic of 0.7 when using clustered binary data. However, if one wishes to use median predictions out of random- or fixed-effect models or marginal predictions when ICC is 20%, one cannot produce a satisfactory model (with C statistic of 0.7) with larger EPV.

Brier score

Figure 5.7 presents percent relative differences (%) in Brier score by EPV and ICC over types of predictions based on 400 simulations.

As can be seen, EPV had an effect on relative differences in Brier score for all ICCs and types of predictions. The pattern was similar to those in the calibration slope and C statistic.

As for the calibration slope and C statistic, the Brier score was influenced by ICC. The pattern was similar to those in the calibration slope and C statistic.

To sum up, given that the true values for ICC of 0%, 5%, 10% and 20% are 0.17, 0.16, 0.16 and 0.15, if one wishes to have a model with 0.18 Brier score, one needs EPV of at least 10 events per variable.

Further investigation

Further investigation was carried out in order to learn the reason behind the ordering pattern observed in relative differences in calibration slope, C statistic and Brier score observed in Figures 5.5 to 5.7. The comparison between cluster-specific and median

predictions by sub-studies from a random-intercept logistic model was presented in Figures 5.8, and from fixed-effect logistic model was displayed in Figure 5.9. From the Figures, while the random-intercept prediction model can distinguish better between high- and low-risk patients as EPV and ICC increases, the fixed-effect prediction model was not affected by ICC.

5.5 Comparing the recommended sample size with the one used in common practice

We have reviewed in section 5.1 that the common method used in practice is that one calculates the sample size by taking a standard approach for independent binary outcome and then multiplies the results by a design factor, $1 + (m - 1)\times$ICC.

Besides, we have shown that one needs at least 12.5 EPV when developing a risk model using independent binary outcomes, evaluating samples with various outcome prevalences (see section 3.4.3). EPV of 12.5 events per variable is equivalent to sample size of about 481, when there are 10 potential variables to be included in the model and the outcome prevalence is 0.26. Thus, if the average cluster sizes and ICC are 30 and 20% respectively, given the traditional approach, the required sample size to develop a random-intercept logistic model is 3264 calculated by $((1 + (30 - 1)0.20) \times 481)$.

However, the results of our simulations in this chapter showed that one needs EPV of 20 (equivalent to $\frac{20\times11}{0.26} = 846$) for the same situations to develop a random-intercept logistic model using clustered binary outcomes.

5.6 Conclusion

The focus of this chapter was on the sample size requirements when developing risk models using clustered binary outcomes. Moineddin et al. (2007) suggested there should be at least 50 clusters of size 50 with more than one event per cluster when the objective of the study is to accurately estimate fixed- and random-effect parameters, and have precise estimations for fixed-effect parameters to develop a two-level risk model using binary clustered outcomes. Wynants et al. (2015), examining only median predictions (p_0), suggested that EPV should be at least 10 when developing a random-intercept prediction model using clustered binary outcome.

5. SAMPLE SIZE REQUIREMENTS FOR DEVELOPING A RISK MODEL USING BINARY CLUSTERED DATA

The available literature are not sufficient for the situation where the aim of the study is to develop a random-intercept risk prediction model using clustered binary outcome in order to predict for patients from clusters which were or were not in the development dataset. Thus, we investigated whether the performance of a random-intercept logistic model is influenced by EPV in the case study.

In our case study, the three common performance measures of the calibration slope, C statistic and Brier score increased by increasing EPV, which was a sign that EPV has an effect on the performance of the random-intercept logistic risk model. Therefore, our further investigation continued to address the issue of whether EPV can still play a vital role in the performance of the model if other factors such as the number of clusters, their size and ICC varies from what is observed in our real data. A simulation study was set to address this issue.

Based on the results of our simulations in this chapter, we suggest that , in general, EPV should exceed 20 when developing a random-intercept and/or a marginal risk prediction model. If one wishes to use cluster-specific predictions of a random-intercept model, EPV needs to be at least 15 when developing the model. Moreover, we recommend adjusting EPV (including the parameters for the centre indicators) when developing a fixed-effect model.

Moreover, from the results of our simulations in this chapter, it is recommended to use random-intercept logistic models rather than fixed-effect logistic models when the ICC is large (say, greater than 10%), and the aim is to use cluster-specific predictions. As in our simulation study, the random-intercept logistic model produced better cluster-specific predictions compared to the fixed effect logistic model when the ICC was large.

Additionally, predictive performance measures obtained from marginal predictions were similar to those obtained from median predictions in our simulation study.

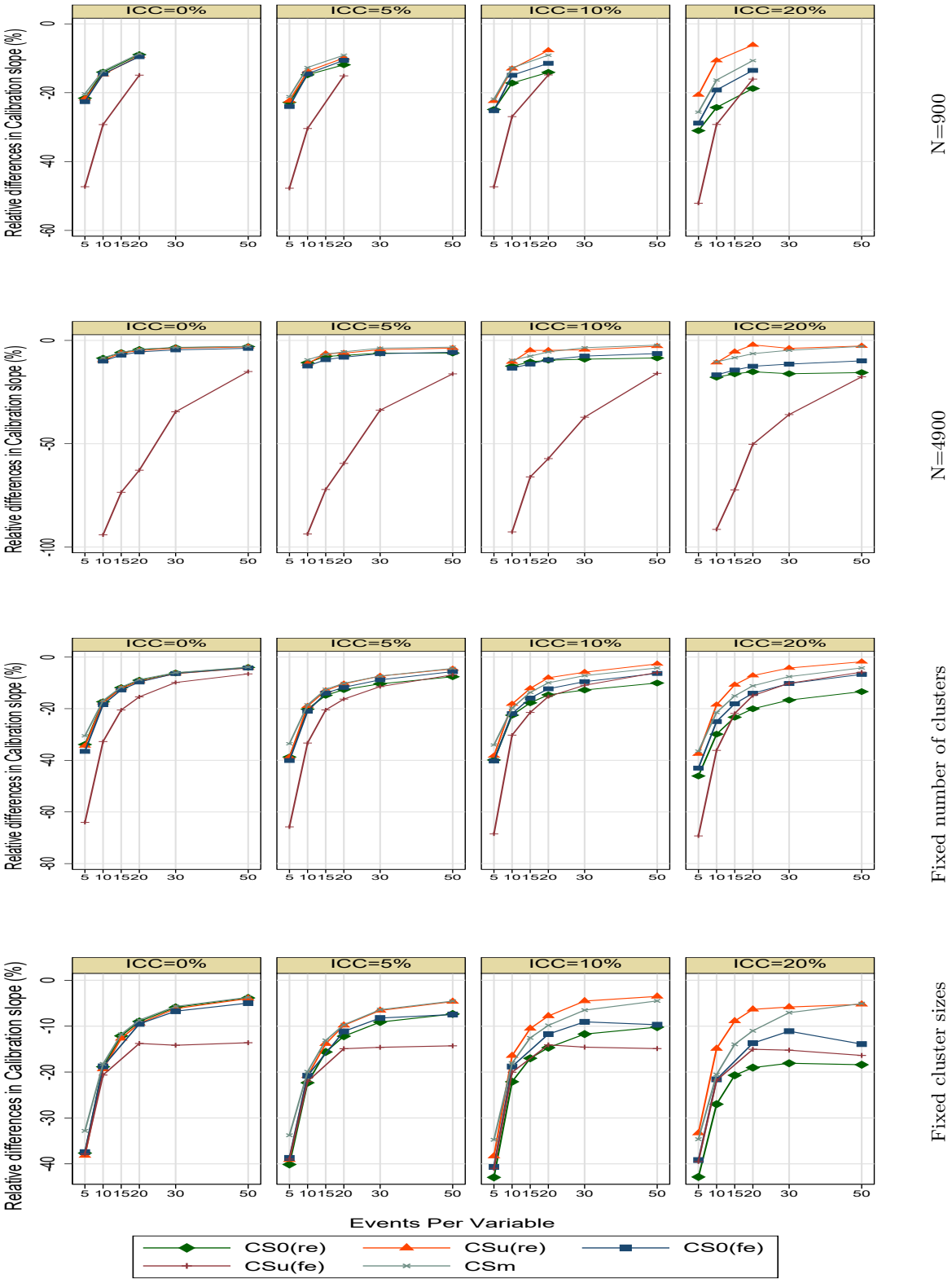


Figure 5.5: Percent relative differences in the calibration slope by EPV and ICC over types of predictions based on 400 simulations. The standard error of the simulation for the calibration slope among all simulated scenarios were (0.002, 0.015). $CS_{0(re)}$: calibration slope obtained using median predictions of random-effect model. $CS_{u(re)}$: calibration slope obtained using cluster-specific predictions of random-effect model. $CS_{0(fe)}$: calibration slope obtained using median predictions of fixed-effect model. $CS_{u(fe)}$: calibration slope obtained using cluster-specific predictions of fixed-effect model.CS_m: calibration slope obtained using marginal predictions.

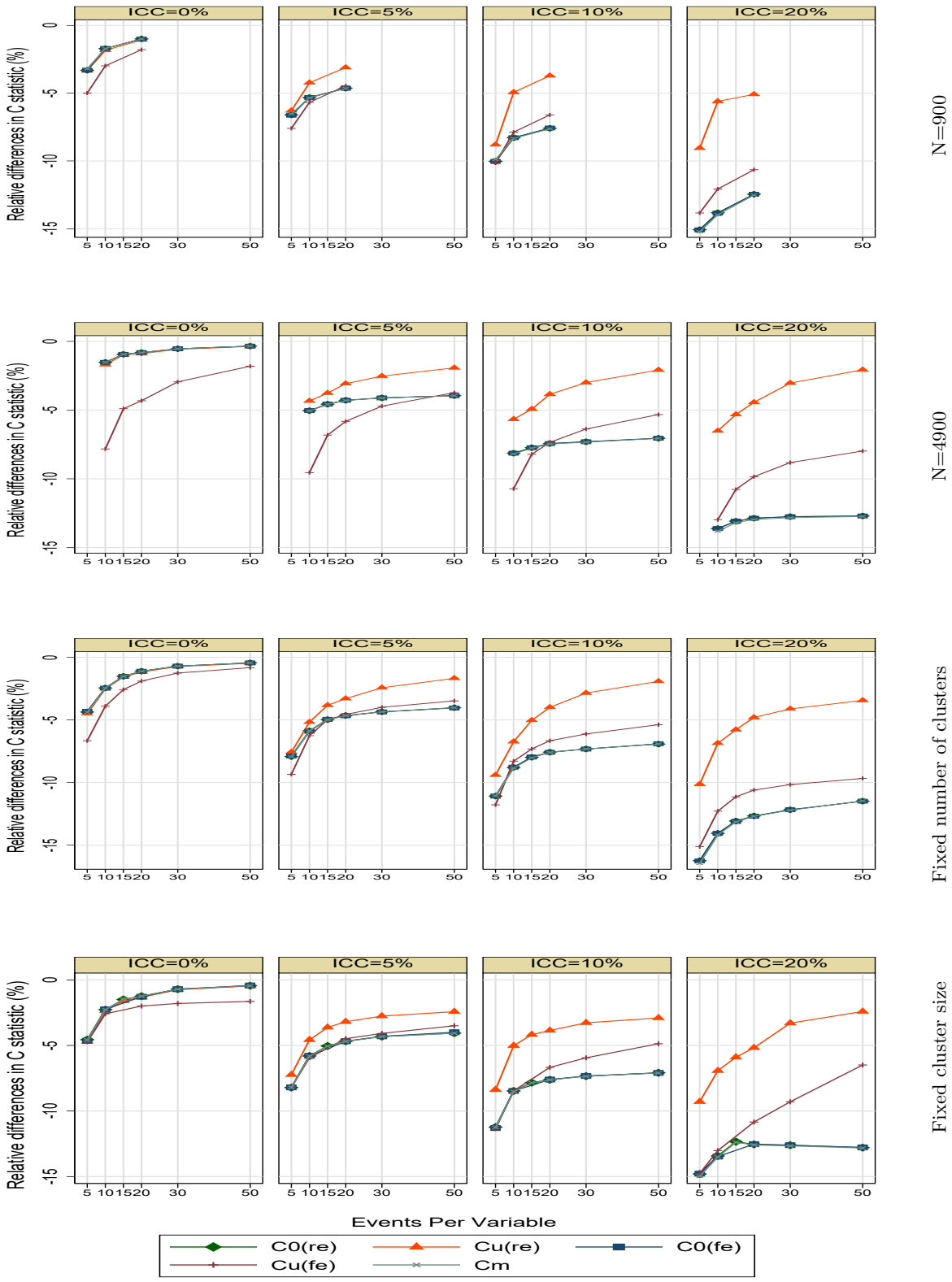


Figure 5.6: Percent relative differences in the C statistic by EPV and ICC over types of predictions based on 400 simulations. The standard error of the simulation for the C statistic among all simulated scenarios were (0.0004, 0.0020). $C_{0(re)}$: C statistic obtained using median predictions of random-effect model. $C_{u(re)}$: C statistic obtained using cluster-specific predictions of random-effect model. $C_{0(fe)}$: C statistic obtained using median predictions of fixed-effect model. $C_{u(fe)}$: C statistic obtained using cluster-specific predictions of fixed-effect model.C_m: C statistic obtained using marginal predictions.

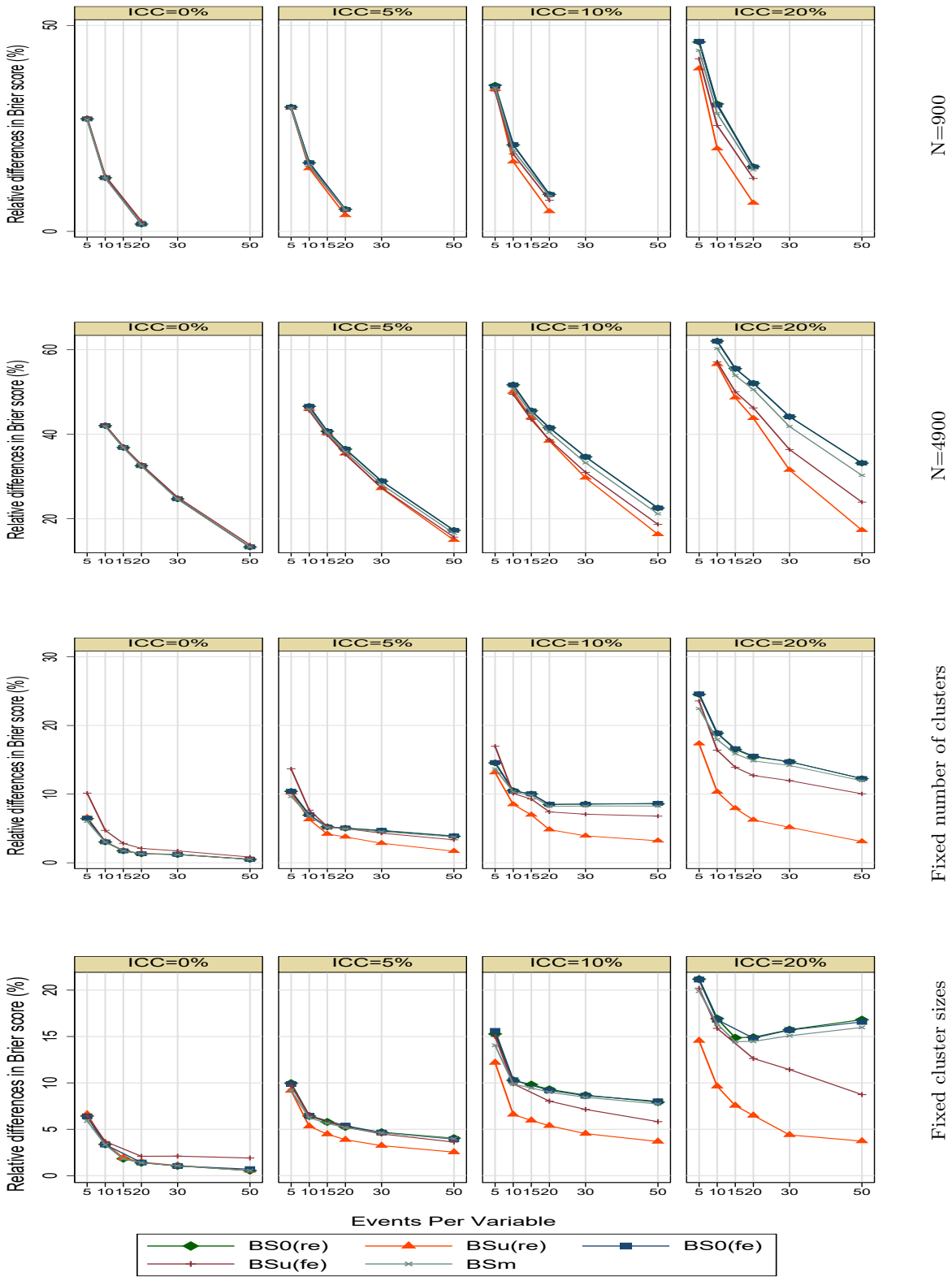


Figure 5.7: Percent relative differences in the Brier score by EPV and ICC over types of predictions based on 400 simulations. The standard error of the simulation for the Brier score among all simulated scenarios were (0.002, 0.015). $BS_{0(re)}$: calibration slope obtained using median predictions of random-effect model. $BS_{u(re)}$: Brier score obtained using cluster-specific predictions of random-effect model. $BS_{0(fe)}$: Brier score obtained using median predictions of fixed-effect model. $BS_{u(fe)}$: Brier score obtained using cluster-specific predictions of fixed-effect model.BS_m: Brier score obtained using marginal predictions.

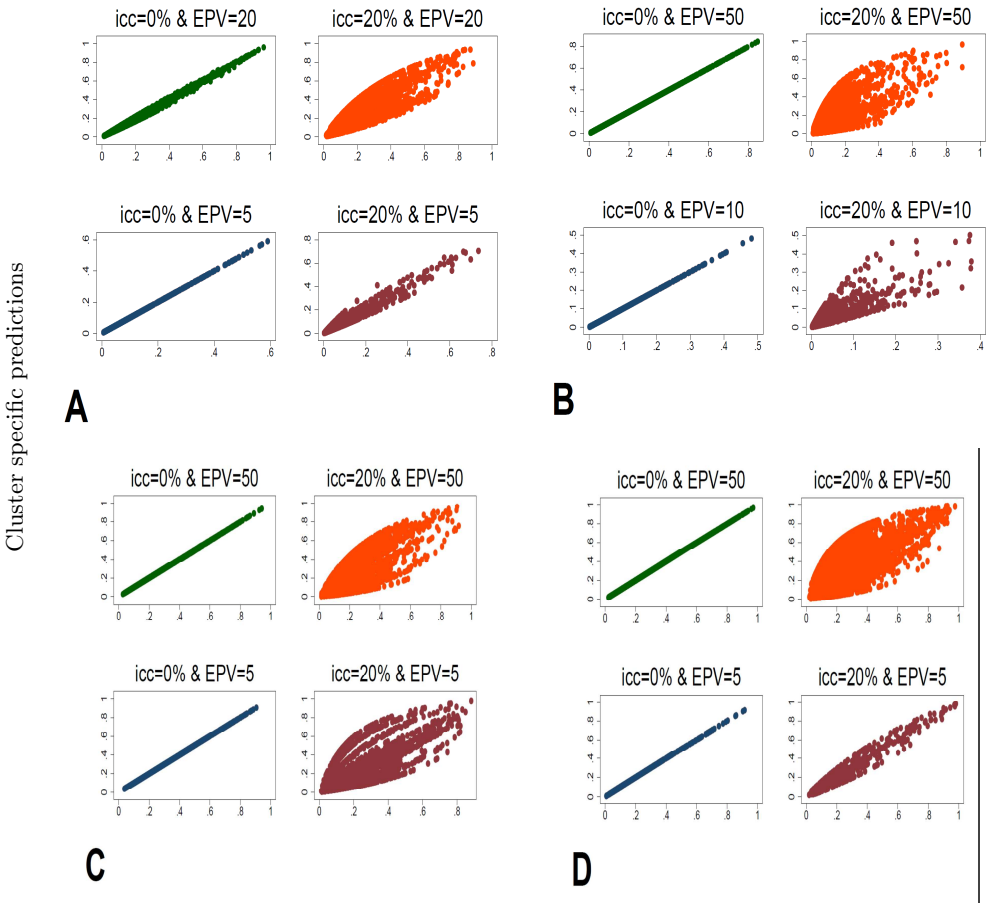


Figure 5.8: Comparing cluster-specific predictions with median predictions from a random-intercept model in one simulated dataset for EPV of 5 or 10 and 20 or 50 and ICC of 0% and 20%. A & B: fixed number of clusters and cluster sizes and varied outcome prevalence (N_A=900 & N_B=4900). C: Fixed number of clusters. D: fixed cluster sizes.

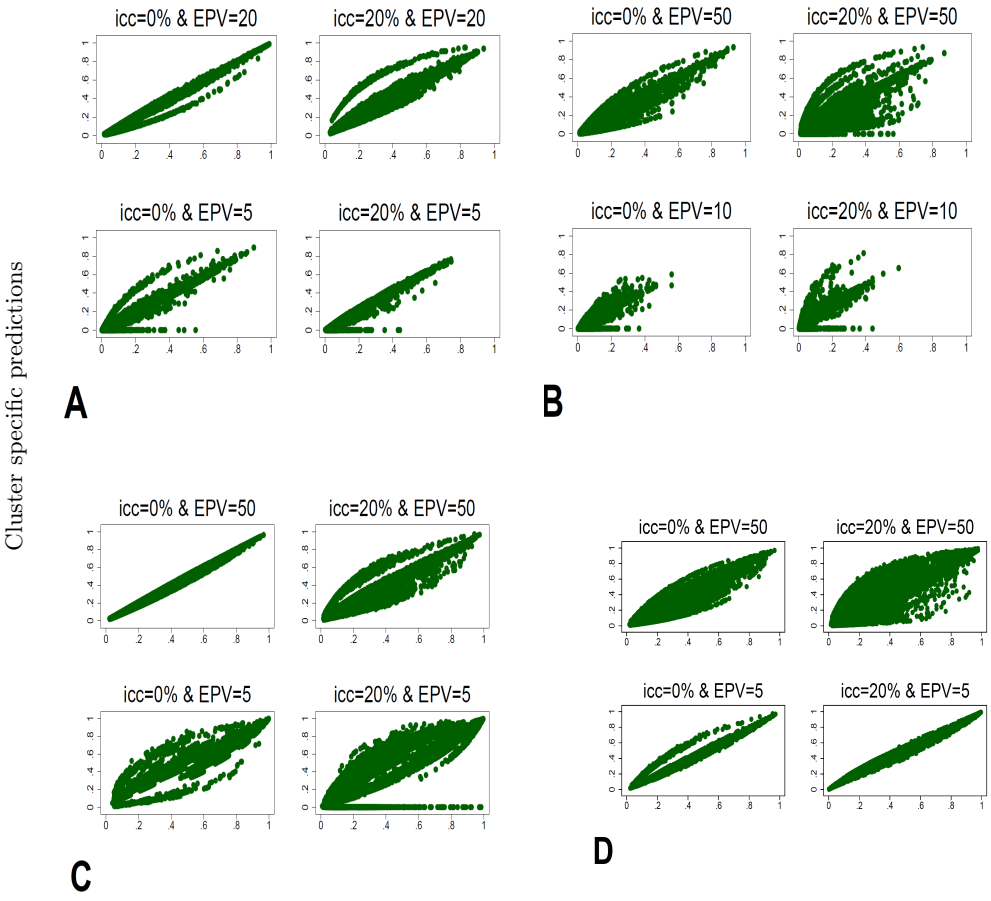


Figure 5.9: Comparing cluster-specific predictions with median predictions from a fixed-effect model in one simulated dataset for EPV of 5 or 10 and 20 or 50 and ICC of 0% and 20%. A & B: fixed number of clusters and cluster sizes and varied outcome prevalence (N_A=900 & N_B=4900). C: Fixed number of clusters. D: fixed cluster sizes.

Chapter 6

Sample Size Requirements for Validating a Risk Model Using Clustered Binary Data

The sample size requirements to validate a reliable risk model using independent binary outcome were discussed in Chapters 4. This chapter focuses on what sample size is required to determine whether or not a multilevel risk prediction model is valid.

Clustered data is very common in medical research. For example, the subjects of the study are clustered within centres such as hospitals, general practices and clinics. In such situations, observations within clusters are correlated with each other and are independent from those patients who are in other clusters. A random-effect model is often used when the studied data is clustered.

The chapter includes the following sections: Section 6.1 explores the available and relevant literature; Section 6.2 provides a case study to provide insight on the issue of the number of events in the context of clustered binary outcomes; Section 6.3 outlines the results of a simulation study, which was conducted to investigate further the problems highlighted in the case study; Section 6.4 discusses the findings and makes recommendations based on these.

6.1 The number of events in validating a clustered risk model: A review

According to our literature search, to date there have been no studies or guidelines regarding the number of events required to validate a reliable risk prediction model when using clustered binary outcomes.

6.2 Case study

A case study was conducted to understand how performance measures related to the random-intercept logistic models can be affected by the number of events.

6.2.1 Method

A similar methodology to the one outlined in chapter 5 (section 5.3) was used. In brief, the first five years of heart valve surgery data was used to develop the model and the rest was used for validation. The size of the validation dataset was varied to obtain the desired number of events by separately sampling without replacement from events and nonevents. A random-intercept logistic regression model was fitted on the full size development dataset and used to obtain three types of predictions for patients in the validation data. The predictions were cluster-specific predictions (p_u), median predictions (p_0) and marginal predictions (p_m) (see section 2.3). For each type of prediction, measures of performance were calculated on the validation data using the calibration slope, C statistic and Brier score. All procedures were repeated 200 times for each combination of factors. Five values of the number of events were studied: 50, 100, 200, 500 and 1047. This list includes previously recommended values, where 1047 is the number of events in the full size validation dataset.

There were 30 clusters of sizes ranged from 298 to 2007 patients. The intra-cluster correlation (ICC) coefficient was about 6% (see section 5.2).

The performance measures estimated using the full size validation data were the

calibration slope = 1.12 (obtained using cluster specific predictions, 1.10 obtained using median or marginal predictions), the C statistic=0.77 (obtained using all types of predictions), and Brier score=0.050 (for all types of predictions). The calibration slope suggests that this model was underfitted. The model was therefore recalibrated using part (50%) of the validation data. A linear transformation of the predicted log-odds was found to be appropriate. The recalibration improved the fit of the model and produced a calibration slope of 1.00 in the half of the validation data not used for recalibrating the model. That is, the agreement between the predicted and observed outcomes is good and the model performs well in terms of separation of different risk groups.

6.2.2 Results

Table 6.1 presents the mean value and standard deviation of the performance measures by number of events and types of predictions over 200 samples.

As can be seen, the precision of the predictive performance measures computed using predictions of random-intercept logistic models were affected by the number of events in the validation dataset. The precision of the C statistic in validation data declined as the number of events decreased (standard deviation increased from 0.006 for 500 events to 0.032 for 50 events) for all types of predictions. The precision of the calibration slope in the validation datasets was also influenced by the number of events (standard deviation increased from 0.029 for 500 events to 0.143 for 50 events). The precision of the Brier score was also affected by the number of events (standard deviation increased from 0.0004 for 500 events to 0.0016 for 50 events).

Furthermore, the precision of the predicted performance measures were affected at almost the same rate for all types of predictions. That might be expected due to the low clustering level in the heart valve surgery data (see section 5.2).

Table 6.1: Mean value and standard deviation (SD) of each performance measure by the number of events and type of predictions obtained from random-intercept logistic models in the validation datasets over 200 samples.

Types of prediction	Nominated events	Mean performance measures (SD)		
		Calibration slope	**C statistic**	**Brier score**
p_0	1047	1.00	0.77	0.050
	500	1.00 (0.029)	0.77 (0.006)	0.052 (0.0004)
	200	1.00 (0.066)	0.78 (0.015)	0.052 (0.0007)
	100	1.00 (0.094)	0.78 (0.021)	0.052 (0.0010)
	50	1.00 (0.143)	0.78 (0.032)	0.052 (0.0016)
p_m	1047	1.00	0.77	0.050
	500	1.00 (0.028)	0.77 (0.006)	0.052 (0.0004)
	200	1.00 (0.066)	0.78 (0.015)	0.052 (0.0007)
	100	1.00 (0.099)	0.78 (0.022)	0.052 (0.0010)
	50	1.00 (0.143)	0.78 (0.033)	0.052 (0.0016)
p_u	1047	1.00	0.77	0.050
	500	1.00 (0.029)	0.77 (0.006)	0.052 (0.0004)
	200	1.00 (0.066)	0.78 (0.015)	0.052 (0.0007)
	100	1.00 (0.094)	0.78 (0.021)	0.052 (0.0010)
	50	1.00 (0.143)	0.78 (0.032)	0.052 (0.0016)

p_0: median predictions.
p_m: marginal predictions.
p_u: cluster-specific predictions.

6.2.3 Discussion

From the results of this case study, it is evident that the precision (but not the bias) of the performance measures obtained using any type of predictions was affected by the number of events.

Further investigation of the results of simulation was carried out in order to understand why only the standard deviation was dependent on the number of events. We therefore plotted the distribution of the cluster-specific linear predictors of patients with ($\hat{\eta}_u^{(1)}$) and without ($\hat{\eta}_u^{(0)}$) the event of interest for all 200 datasets and the number of events (Figure 6.1).

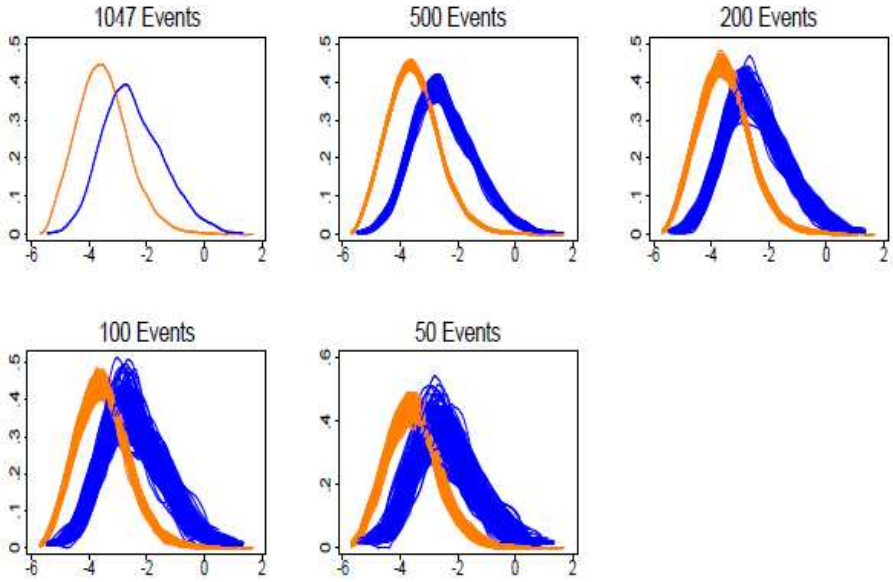


Figure 6.1: Separation between $\hat{\eta}^{(0)}$ and $\hat{\eta}^{(1)}$ by number of events. $\hat{\eta}^{(0)}$ and $\hat{\eta}^{(1)}$ are linear predictors, corresponding to nonevent and event groups, respectively. Orange lines present the distribution of linear predictors for patients without the event of interest and the blue lines correspond to the distribution of linear predictors for those with the event of interest.

As can be seen, the separation between $\hat{\eta}_u^{(1)}$ and $\hat{\eta}_u^{(0)}$ was similar for different numbers of events, but the variation within $\hat{\eta}_u^{(1)}$ and $\hat{\eta}_u^{(0)}$ increased as the number of events decreased.

Furthermore, the separation and variation for the linear predictors obtained using marginal-effects and/or fixed-effects appeared to be similar to those for cluster-specific effects (results not shown). This is probably because the clustering effect in the heart valve surgery data used in this case study is not strong.

However, there are other issues to address in this chapter, such as whether the number of events is the only factor that impacts the performance of measures when

validating a risk model using clustered binary outcomes or whether there are other influential factors coupled with the number of events. Various simulation studies were therefore conducted, using the heart valve surgery data, altering certain characteristics of the validation datasets, such as the number of clusters and ICC. The performance of validation measures was then examined in validation data.

6.3 Simulation study

Table 6.2: Features of the studied scenarios and source datasets to validate a random-intercept logistic model with eleven parameters.

	Source dataset	Sample				Number of
	ICC(%)	J	$\bar{n}_j$ (min, max)	N	p	events
(1)	0, 5, 10, 20	30	30 (12, 62)	900	0.06	50
	0, 5, 10, 20	30	30 (12, 62)	900	0.12	100
	0, 5, 10, 20	30	30 (12, 63)	900	0.24	200
	0, 5, 10, 20	30	30 (12, 63)	900	0.37	300
	0, 5, 10, 20	70	70 (27, 163)	4900	0.01	50
	0, 5, 10, 20	70	70 (27, 164)	4900	0.02	100
	0, 5, 10, 20	70	70 (28, 165)	4900	0.04	200
	0, 5, 10, 20	70	70 (27, 163)	4900	0.07	300
	0, 5, 10, 20	70	70 (28, 164)	4900	0.11	500
(2)	0, 5, 10, 20	30	7 (1, 16)	210	0.26	50
	0, 5, 10, 20	30	14 (5, 31)	420	0.26	100
	0, 5, 10, 20	30	28 (11, 59)	840	0.26	200
	0, 5, 10, 20	30	42 (18, 86)	1260	0.26	300
	0, 5, 10, 20	30	70 (32, 144)	2100	0.26	500
(3)	0, 5, 10, 20	7	30 (17, 47)	210	0.26	55
	0, 5, 10, 20	14	30 (14, 54)	420	0.26	100
	0, 5, 10, 20	28	30 (12, 62)	840	0.26	200
	0, 5, 10, 20	42	30 (11, 67)	1260	0.26	300
	0, 5, 10, 20	70	30 (10, 74)	2100	0.26	500

ICC: Intra-cluster-correlation coefficient. J: the number of clusters. n_j: cluster size. N: sample size. p: outcome prevalence.

(1) Fixed number of clusters and cluster size, varying outcome prevalence.

(2) Fixed number of clusters and outcome prevalence, varying cluster size.

(3) Fixed cluster size and outcome prevalence, varying number of clusters.

As can be seen from the case study, the performance of validation measures in the validation dataset is influenced by the number of events. However, it is not obvious whether the number of events can still be influential if, for example, the number of clusters in the dataset is different to that observed in the heart valve surgery dataset. A number of simulation studies were conducted to address this question. These simu-

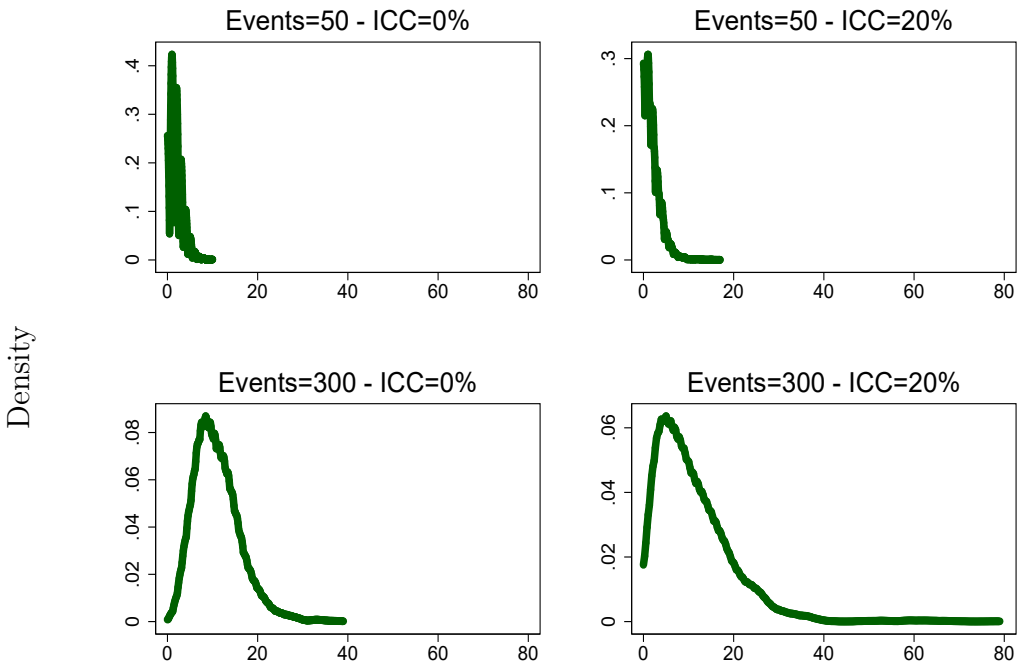


Figure 6.2: Distribution of events across clusters in five simulated datasets for ICCs of 0% and 20% and the number of events of 50 and 300. This figure corresponds to a fixed number of clusters and cluster sizes and varied outcome prevalence scenario, N=900. Note that there are 30 clusters of an average size of 30 in each simulated dataset.

lations were based on the heart valve surgery data. In these studies, different scenarios were investigated varying the number of events, ICC, number of clusters and the size of clusters. The methodology and scenarios used in the simulation study were similar to those described in chapter 5.4 with the following three exceptions. First, we divided the dataset in half (one half for development and the other for validation). Second, unlike chapter 5, the development data was always 50% of the heart valve surgery data, but the size of the validation sample varied according to the required number of events. Finally, there were always 151 clusters in the development dataset, but the number of

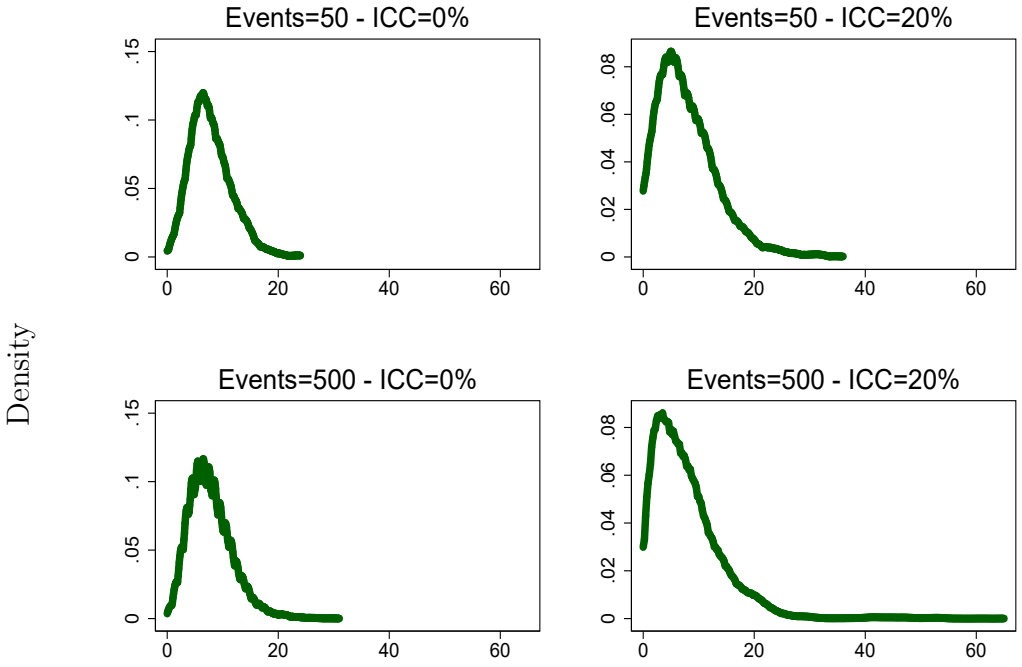


Figure 6.3: Distribution of events across clusters in five simulated datasets for ICCs of 0% and 20% and the number of events of 50 and 500. This figure corresponds to the fixed cluster sizes. Note that when there were 50 events, there are seven clusters of an average size of 30; and when there were 500 events, there are 70 clusters of an average size of 30 in each simulated dataset.

clusters in the validation dataset was varied according to the scenario. The features of the simulated datasets are summarised in Table 6.2, and the distribution of events within clusters in the simulated validation datasets is shown in Figures 6.2 to 6.3.

In Figure 6.2, we see that the simulated validation datasets are similar to the development datasets in chapter 5. To explain further, there were more sparse clusters when the number of events was low and the ICC was small, compared to when the number of events was low and the ICC was large. However, the distribution of events followed the binomial distribution when the number of events was high and the ICC was

low, but it became heavily skewed to the right when the number of events remained the same but ICC was high. A similar pattern was observed for the scenario of 70 clusters of an average size of 70 and the scenario with a fixed number of clusters (figures not shown). In contrast, (see Figure 6.3), the distribution of the events followed the binomial distribution when both the number of events and ICC were small, but it became more positively skewed when either the number of events, the ICC or both increased.

6.3.1 Results

The results of the simulation studies are presented in this section.

Figures 6.4, 6.5 and 6.6 present the precision and bias of the predictive performance measures obtained using three types of predictions (cluster-specific (p_u), median (p_0) and marginal (p_m) predictions) by the number of events across ICCs for all scenarios.

As these figures show, the precision of the predictive performance measures was influenced by the number of events. The precision of the calibration slope, C statistic and Brier score increased as the number of events in the validation dataset increased.

The precision of the predictive performance measures was affected differently by the number of events depending on the type of predictions used. The precision of calibration slope obtained using cluster-specific predictions was slightly better than that obtained using other types of predictions. The precision of the C statistic and Brier score obtained using cluster-specific predictions was slightly worse than those obtained using other types of predictions.

Additionally, the estimated calibration slopes and C statistics were only biased when ICC was greater than 5%. This is probably because the random-intercept logistic models underestimate the variance of random-intercept even when the development sample is large, as confirmed in all previous studies (see, for example, Moineddin et al. (2007); Wynants et al. (2015)). Also, the estimated Brier scores in two sub-studies

(when the sample sizes were either 900 or 4900) were highly biased at all ICC levels (Figure 6.6 top two rows). This is due to different outcome prevalences in the validation and development datasets; as the differences in outcome prevalences in the validation and development data decreased, the biased decreased. Furthermore, the estimated Brier scores for the last sub-studies (fixed number of clusters, and fixed cluster sizes) were also slightly biased due to the fact that we used the predictions from models developed on full size development data, rather than those from the true model. The bias will be zero when increasing the number of simulations.

To conclude, according to our results of the simulation, there should be at least 200 number of events when validating a logistic model using clustered binary outcomes in order to precisely obtain predictive performance measures. In this number of events, the measures were less biased, and the width IQR for calibration slope, C statistic, Brier score were less than 20%, 5%, and 5%, respectively.

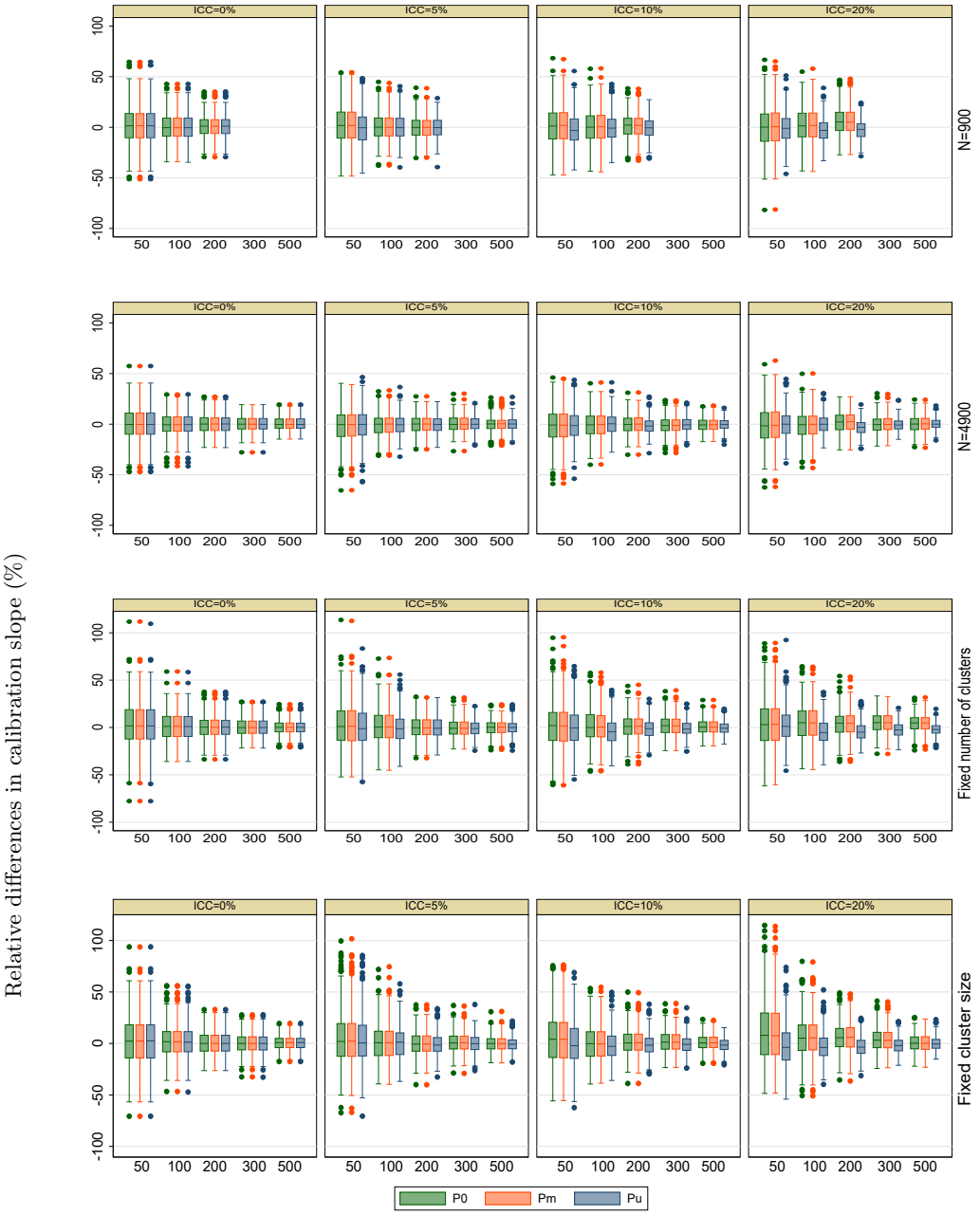


Figure 6.4: Percent relative differences in the calibration slope obtained using different types of predictions based on 500 simulations. The standard error range of simulation among all number of events, ICC and simulated scenarios for the estimated calibration slope obtained using cluster-specific, median and marginal predictions were (0.0007, 0.0101), (0.0007, 0.0011) and (0.0008, 0.0120), respectively.

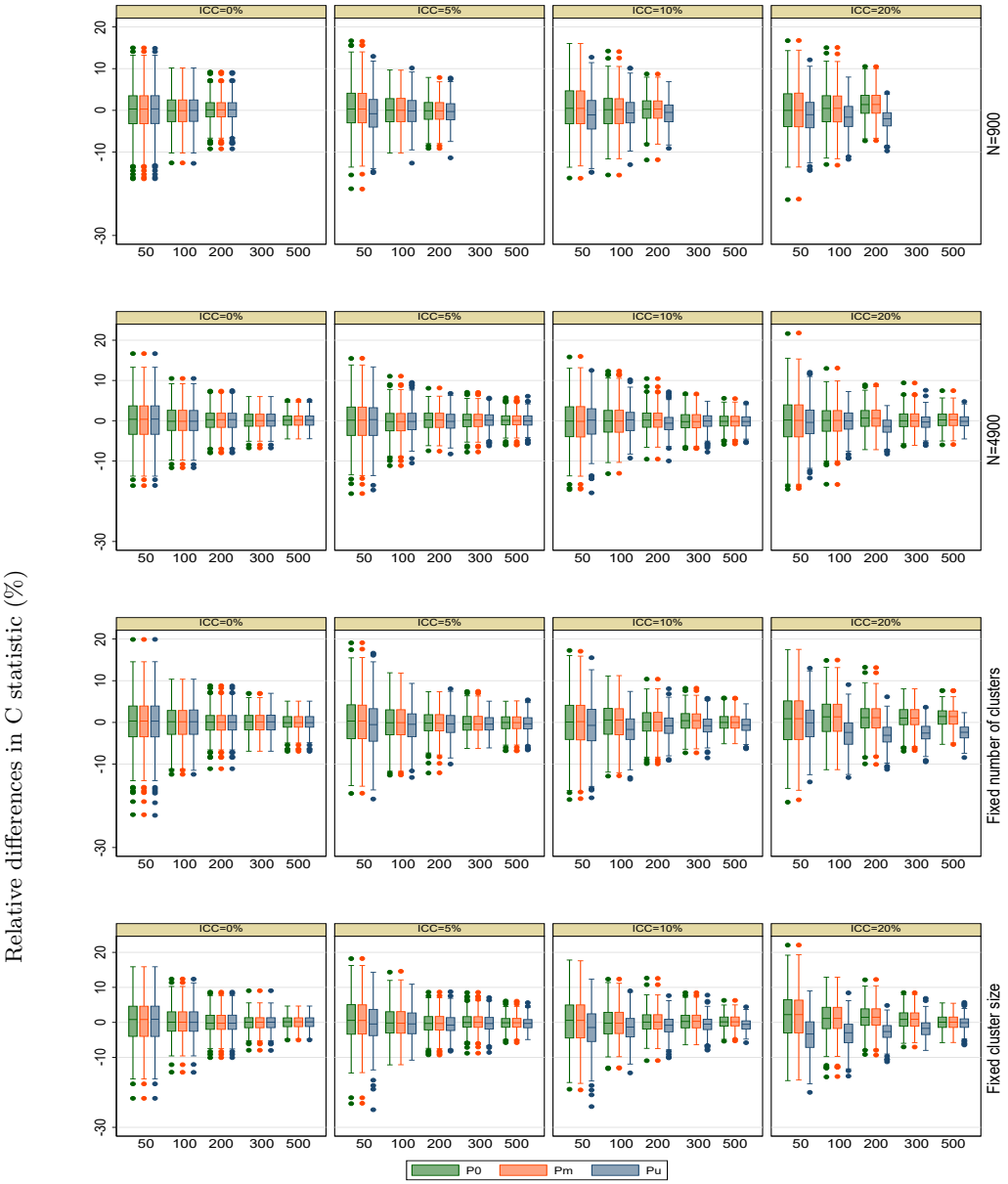


Figure 6.5: Percent relative differences in the C statistic obtained using obtained using different types of predictions based on 500 simulations.The standard error range of simulation among all number of events, ICC and simulated scenarios for the estimated C statistic obtained using all types of predictions was (0.0001, 0.0020).

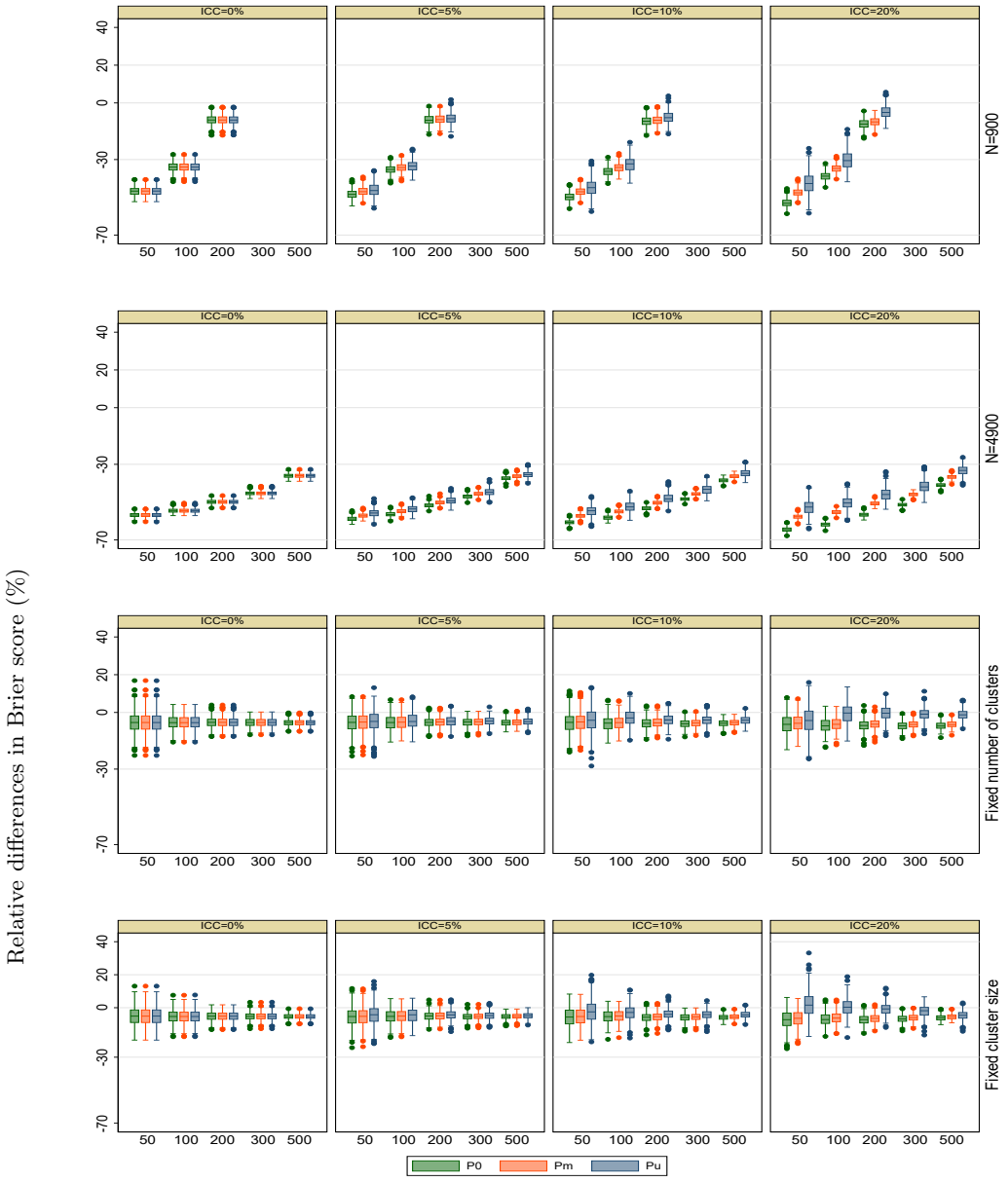


Figure 6.6: Percent relative differences in the Brier score obtained using different types of predictions based on 500 simulations. The standard error range of simulation among all number of events, ICC and simulated scenarios for the estimated Brier score obtained using cluster-specific, median and marginal predictions were (0.00004, 0.00049), (0.00004, 0.00045) and (0.00012, 0.00042), respectively.

6.4 Conclusion

The focus of this chapter was on the sample size requirements when validating a risk model using clustered binary outcomes. There are no guidelines on how many events are required when validating a hierarchical risk model using a clustered binary dataset.

This chapter opened with a case study using the heart valve surgery data to investigate whether validation performance measures computed using the predictions from a random-intercept or standard logistic models are influenced by the number of events.

With the use of three common performance measures of the calibration slope, C statistic and Brier score in the case study, it was observed that the precision of the performance measures decreased as the number of events in the validation data increased. In other words, the number of events had an effect on the performance of validation measures in the context of clustered data. This raised additional questions regarding whether the number of events can still affect the performance of validation measures if the number of clusters, their size and ICC changes. A simulation study was designed to address this question.

From the results of the simulations in this chapter, we observed that there should be at least 100 events in the validation dataset to obtain predictive performance measures with reasonable accuracy and precision for the random-intercept logistic risk models.

We recommend that the researchers use at least 100 events in their validation dataset.

Chapter 7

Discussion and Conclusion

7.1 Summary and discussion

Prognostic studies are often conducted to develop risk prediction models. These models are used to provide useful information with regard to a patient's health or to observe the performance of health institutes after considering differences in case-mix. It is essential for risk prediction models to produce accurate or acceptable predictions. Hence, the central prerequisite of developing and validating a prediction model is to use an adequate number of observations to ensure that the model performs satisfactorily in new data. This research was designed to investigate the sample size requirements when developing and/or validating a risk prediction model using independent or clustered binary outcomes.

The dissertation started with introductory remarks in Chapter one, followed by the description of key concepts of risk modelling in Chapter two. Sample size requirements to develop and validate risk prediction models using independent binary outcomes were investigated in Chapters three and four, and using clustered binary data were examined in Chapters five and six. The details of our contribution to the literature is discussed here.

In Chapter three, we investigated the sample size requirements when developing a risk model using binary outcomes. In this chapter, we found that EPV needs to be at

least five when developing risk models using independent binary data when there are a small number of strong predictors; at this EPV, the differences between the estimated and reference performance measures were low. In addition, we found that a large EPV is required when the prevalence of the outcome is expected to be high. For example, EPV needs to be at least 30 for outcome prevalence of 40% compared to at least 10 for outcome prevalence of 7%. We also found that the presence of noise variables, the presence of more continuous variables, or even the presence of further predictors in the model is not associated with the performance of a risk model if EPV is held constant. Moreover, a larger EPV (say, at least 15) is needed when developing a risk model using data with a large collinearity (say, greater than 46%).

However, in Chapter three, we also found that if it is possible to apply a linear postestimation shrinkage factor, the EPV requirements can be relaxed. For instance, with the possibility of using postestimation linear shrinkage, EPV can reduce to 10 (from 30) for an outcome prevalence of 40%.

In Chapter four, the required number of events to validate a risk model using independent binary outcomes was examined. We analytically showed that the precision of the performance measures is dependent on the number of events, examining the calibration slope, C statistic, and D statistic. Moreover, with the use of a simulation study, it was found that at least 75 events are needed to validate a risk model using independent binary outcomes regardless of the risk profile. This value was chosen as the corresponding performance measures had relatively high precision.

Based on the results of our simulations in chapter five, we suggest that , in general, EPV should exceed 20 when developing a random-intercept and/or a marginal risk prediction model. Moreover, we recommend adjusting EPV by including the parameters for the centre indicators when developing a fixed-effect model.

Finally, Chapter six was devoted to the sample size requirements to validate a random-intercept logistic model. In this chapter, we found that one needs at least 100

events in the validation data when validating a multilevel risk model. At this number of events, the corresponding performance measures had relatively high precision.

7.2 Possibilities of further research

A number of areas have been identified where further study is possible. They are described as follows.

The sample size requirement to develop and validate a standard logistic regression model was studied under a number of scenarios (Chapters three and four). Time-to-event data are common in medical setting. Further investigation could be conducted based on survival models, such as the Cox proportional hazard, lognormal, and accelerated failure time models, to assess whether the sample size requirements differ for these models.

In addition, further investigation may be needed to see how the *rule of the required number of EPV* to develop a survival risk model can be adjusted when one has the possibility of using a postestimation shrinkage factor.

The required number of events to develop a marginal, fixed-effect logistic regression, and to develop and validate a random-intercept logistic regression, was examined in Chapter five and six. Further investigation could be conducted to see how this rule could be adjusted when developing a random-slope logistic regression model, or a multilevel model with contextual variables.

The clustered survival outcomes are very common in practice. Random-effect frailty model which can take account of this clustering have been proposed for the analysis of clustered survival outcomes, It may also be of interest to study the required sample size to develop and validate a frailty model.

The multicenter data was used in this thesis. However, the measurements may be clustered within each subject. This type of data is also called longitudinal data. Longitudinal data is common in medical research. One could examine whether the rule

of number of events stands for a longitudinal data.

We only used split validation method in our study. There are other types of validation methods which are used in practice. It may be of interest to investigate the required number of events to validate a standard or multilevel risk prediction models using other validation methods such as cross-validation or bootstrap (see section 2.4).

One may also be interested to study the required number of events to validate a multilevel risk prediction models for various risk profile.

7.3 Conclusion

This research is written mainly from a methodological perspective, focussing on the required sample size to develop and/or validate a risk prediction model for a binary outcome.

We believe that the strengths of our research are as follows. We investigate the issue of sample size using both case studies and simulation. In the simulation studies, the investigation of sample size requirements considered a broad range of scenarios such as prognostic strength, outcome prevalence, types of predictors, collinearity, clustering, and size of clusters. We have suggested a range of practical guidelines on how many events are required to develop and validate risk models for binary outcomes; specially there are practical recommendations to use when developing and/or validating a logistic model using independent binary outcomes (Chapters three and four), and using clustered binary outcomes (Chapters five and six). Moreover, a model-maker could use the derived formulae (Chapter four) to obtain the sample size on the anticipated standard error of the calibration slope, D statistic, and C statistic.

One of the weakness of this study is that it was based on just one dataset, the heart valve surgery data, and so other datasets should be explored. Also, we have only explored binary outcome due to time restrictions, though our results can probably be generalised to the time-to-event setting. We also did not explore continuous outcome

setting for the same restriction. The results of our validation sections are applicable to both internal and external validation. For Cross-validation and particularly JackKnife validation, we consider that the overall number of events should be at least 75 for independent binary outcome, and at least 100 events for clustered binary outcomes. Finally, we were not able to run more simulations in some of the simulation studies due to the fact that it was often taking several weeks, and in some cases months, to run all the simulations. This was particularly the case in chapters 5 and 6 that considered clustered binary outcomes.

Appendix A

Relationship between accuracy of estimated regression coefficients and EPV

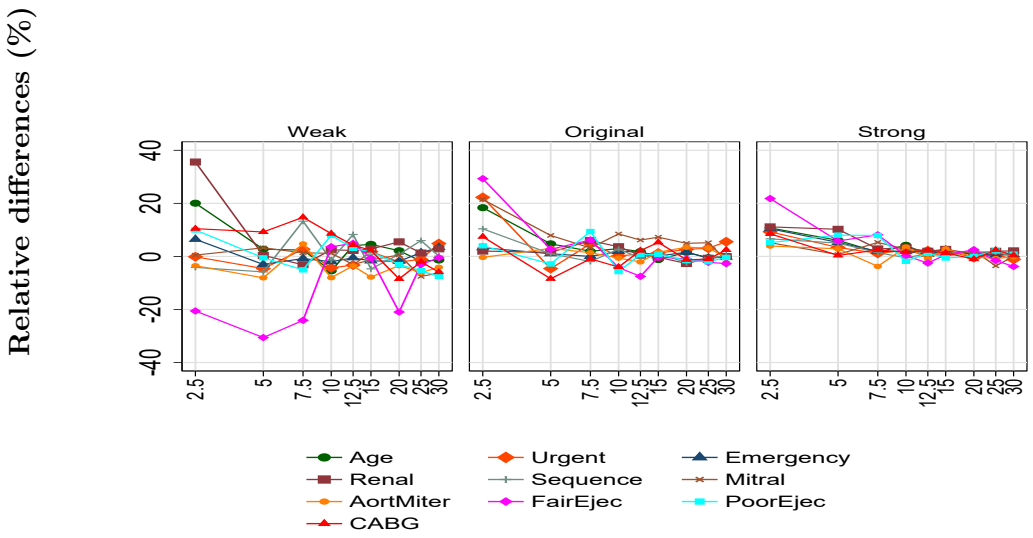


Figure A.1: Relative differences in the estimated regression coefficients over EPVs by strength of risk model based on 200 simulated data.

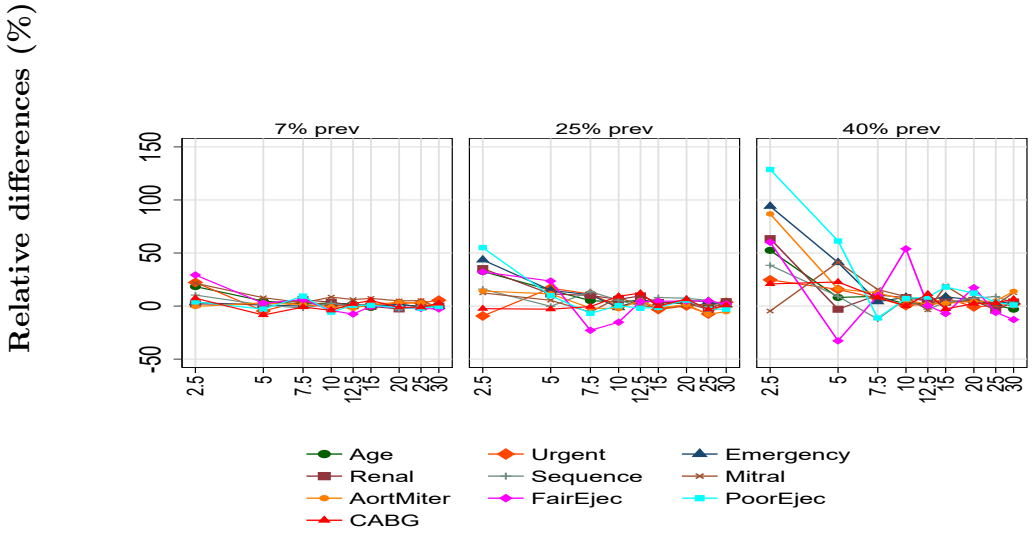


Figure A.2: Relative differences in the estimated regression coefficients over EPVs by outcome prevalence based on 200 simulated data.

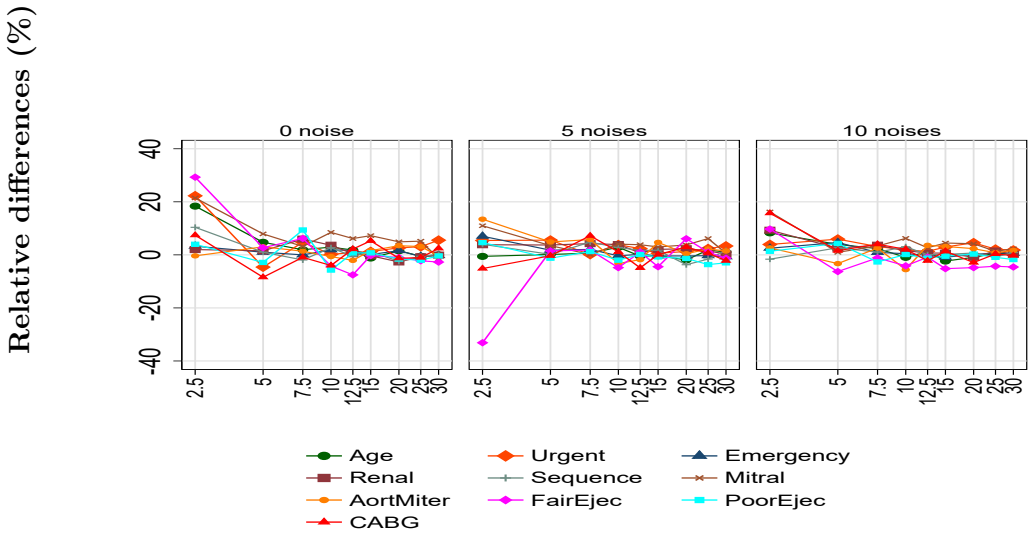


Figure A.3: Relative differences in the estimated regression coefficients over EPVs across the number of noise variables based on 200 simulated data.

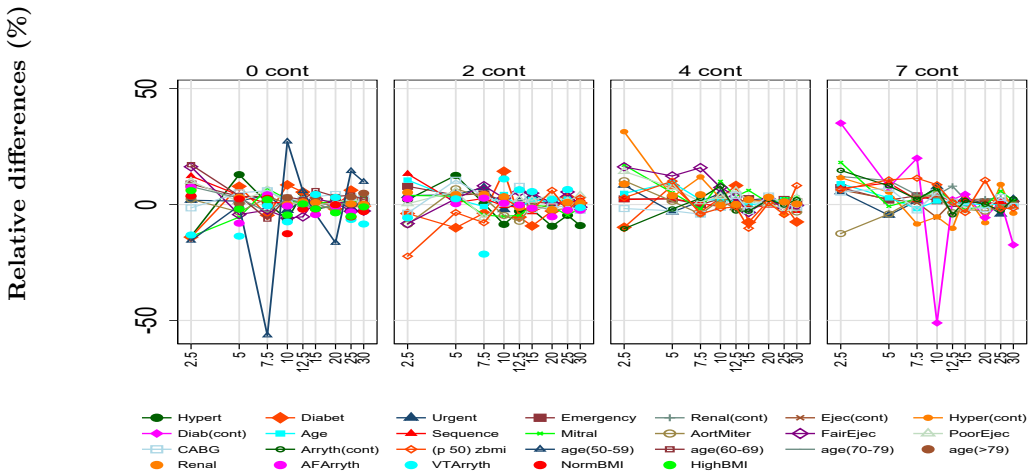


Figure A.4: Relative differences in the estimated regression coefficients over EPVs across the number of continuous variables in the model based on 200 simulations. Note the scale of Y axis.

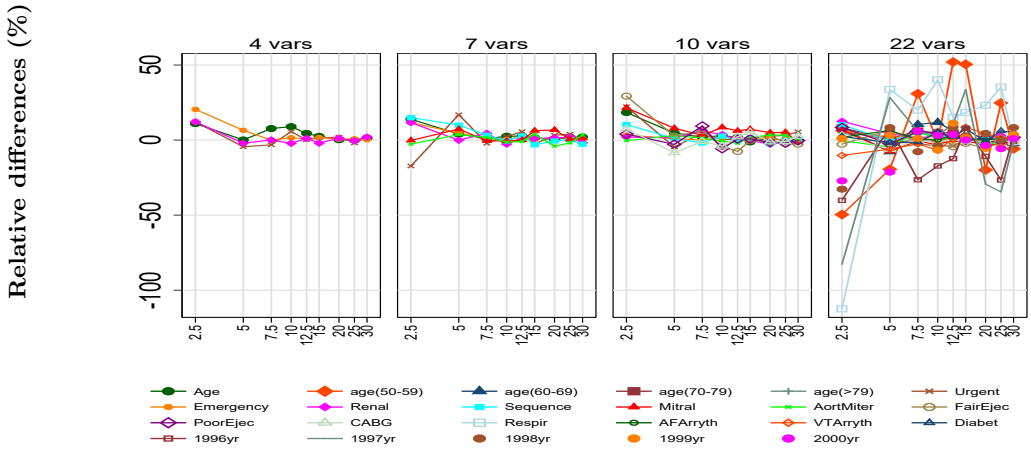


Figure A.5: Relative differences in the estimated regression coefficients over EPVs across the number of variables based on 200 simulations.

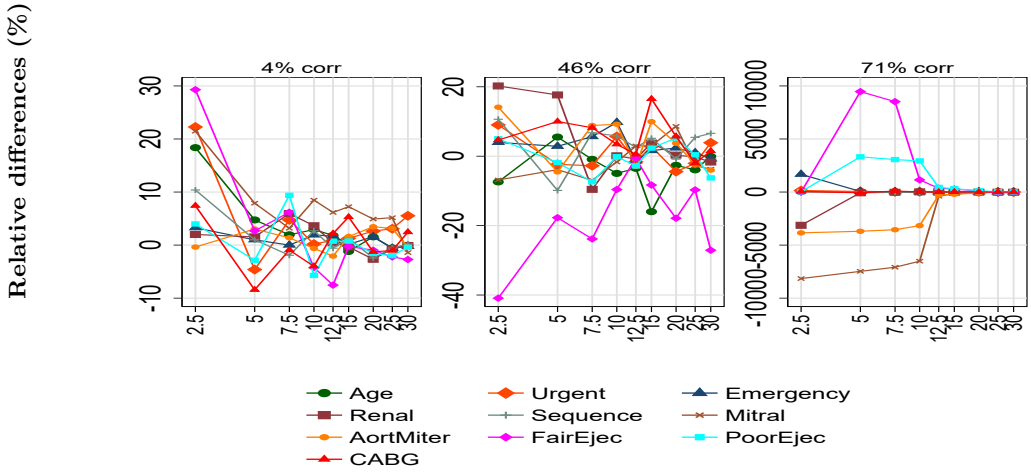


Figure A.6: Relative differences in the estimated regression coefficients over EPVs across the amount of collinearity based on 200 simulations. Note different scale of y axis.

References

Adams, G., M. Gulliford, O. Ukoumunne, S. Eldridge, S. Chinn, and M. Campbell (2004). Patterns of intra-cluster correlation from primary care research to inform study design and analysis. *Journal of Clinical Epidemiology 57*(8), 785–794. 126

Akins, C. W., D. C. Miller, M. I. Turina, N. T. Kouchoukos, E. H. Blackstone, G. L. Grunkemeier, J. J. M. Takkenberg, T. E. David, E. G. Butchart, D. H. Adams, D. M. Shahian, S. Hagl, J. E. Mayer, and B. W. Lytle (2008). Guidelines for reporting mortality and morbidity after cardiac valve interventions. *European Journal of Cardio-Thoracic Surgery 33*(4), 523–528. 42, 82

Albert, A. and J. Anderson (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika 71*(1), 1–10. 51

Altman, D. and P. Royston (2000). What do you mean by validating a prognostic model? *Statistics in Medicine 19*(4), 453–473. 27, 28, 34, 59, 64, 87

Altman, D., Y. Vergouwe, P. Royston, and K. Moons (2009). Prognosis and prognostic research: validating a prognostic model. *BMJ 338*, 1432–1435. 31, 32

Ambler, G., A. Brady, and P. Royston (2002). Simplifying a prognostic model: a simulation study based on clinical data. *Statistics in Medicine 21*(24), 3803–3822. 49

Ambler, G., R. Omar, P. Royston, R. Kinsman, B. Keogh, and K. Taylor (2005).

Generic, simple risk stratification model for heart valve surgery. *American Heart Association 112*, 224–231. 51, 97, 107

Ambler, G., S. Seaman, and R. Omar (2011). An evaluation of penalised survival methods for developing prognostic models with rare events. *Statistics in Medicine 31*, 1150–1161. 29, 48, 49, 50, 64, 84

Anderson, T. (1958). *An Introduction to Multivariate Statistical Analysis*. New York, NY:John Wiley & Sons. 94

Andreas, L., S. Nandita, and G. lan (1997). Clinical prediction rules: A review and suggested modifications of methodological standards. *JAMA 277*, 488–494. 27

Armitage, P., G. Berry, and J. , Matthews (2001). *Statistical Methods in Medical Research* (4 ed.). Blackwell Science. 89

Ash, A. and M. Shwartz (2003). *Risk adjustsments for measuring healthcare outcomes*. Health Administration Office Press. 36

Beitler, P. and J. Landis (1985). A mixed-effects model for categorical data. *Biometrics 41*, 991–1000. 15, 107

Bell, B., J. Ferron, and J. Kromrey (2008). Cluster size in multilevel models: The impact of sparse data structures on point and interval estimates in two-level models. *JSM*, 1122–1129. 109

Bell, B., G. Morgan, J. Kromrey, and J. Ferron (2010). The impact of small cluster size on multilevel models: A monte carlo examination of two-level models with binary and continuous predictors. *JSM*, 4057–4067. 109

Bleekera, S., H. Molla, E. Steyerbergd, A. Donders, G. Derksen-Lubsenc, D. Grobbee, and K. Moons (2003). External validation is necessary in prediction research: A clinical example. *Journal of Clinical Epidemiology 56*, 826–832. 27

Bouwmeester, W., J. Twisk, T. Kappen, W. Klei, K. Moons, and Y. Vergouwe (2013). Prediction models for clustered data: comparison of a random intercept and standard regression model. *BMC Medical Research Methodology 13*(1), 13–19. 111, 112, 117

Breiman, L. (1995). Better subset selection using nonnegative garotte. *Technometrics 37*, 373–384. 49

Brusselaers, N., I. Juhasz, I. Erdei, S. Monstrey, and S. Blot (2009). Evaluation of mortality following severe burns injury in hungary: External validation of a prediction model developed on belgian burn data. *Burns 35*(7), 1009–1014. 85

Bui, A., T. Horwich, and G. Fonarow (2011). Epidemiology and risk profile of heart failure. *Nature Reviews Cardiology 8*, 30–41. 102

Chakraborty, H., J. Moore, W. Carlo, T. Hartwell, and L. Wright (2009). A simulation based technique to estimate intracluster correlation for a binary variable. *Contemporary Clinical Trials 30*(1), 71 – 80. 115

Chamogeorgakis, T., I. Toumpoulis, P. Tomos, C. Ieromonachos, D. Angouras, E. Georgiannakis, P. Michail, and C. Rokkas (2009). External validation of the modified thoracoscore in a new thoracic surgery program: prediction of in-hospital mortality. *Interactive CardioVascular and Thoracic Surgery 9*(3), 463–466. 85

Choodari-Oskooei, B., P. Royston, and M. Parmar (2012a). A simulation study of predictive ability measures in a survival model i: Explained variation measures. *Statistics in Medicine 31*(23), 2627–2643. 89

Choodari-Oskooei, B., P. Royston, and M. Parmar (2012b). A simulation study of predictive ability measures in a survival model ii: explained randomness and predictive accuracy. *Statistics in Medicine 31*(23), 2644–2659. 89

Choodari-Oskooei, B., P. Royston, and M. K. B. Parmar (2012c). A simulation study

of predictive ability measures in a survival model i: Explained variation measures. *Statistics in Medicine 31*(23), 2627–2643. 92

Cochran, W. (1977). *Sampling Techniques*. John Wiley & Sons:New York. 41

Collette, S., F. Bonnetain, X. Paoletti, M. Doffoel, O. Bouch, J. Raoul, P. Rougier, F. Masskouri, L. Bedenne, and J. Barbare (2008). Prognosis of advanced hepatocellular carcinoma: comparison of three staging systems in two french clinical trials. *Annals of Oncology 19*(6), 1117–1126. 90

Collins, G., J. de Groot, S. Dutton1, O. Omar, M. Shanyinde, A. Tajar, M. Voysey, R. Wharton, L. Yu, K. Moons, and D. Altman (2014). External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMJ 14*, 40. 15, 85, 86, 91, 106

Collins, G., E. Ogundimu, and D. Altman (2015). Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. *Statistics in Medicine*, n/a–n/a. SIM-15-0170.R2. 15, 85, 88, 89, 91

Concato, J., P. Peduzzi, T. Holford, and A. Feinstein (1995). Importance of events per independent variable in proportional hazards analysis. i.background, goals, and general strategy. *Journal of Clinical Epidemiology 12*(48), 1495–1510. 41, 44

Copas, J. (1983). Regression, prediction and shrinkage. *Journal of the Royal Statistical Society 45*(3), 311–354. 14, 29, 33, 49

Costantino, J., M. Gail, D. Pee, S. Anderson, C. Redmond, J. Benichou, and H. Wieand (1999). Validation studies for models projecting the risk of invasive and total breast cancer incidence. *JNCI: Journal of the National Cancer Institute 91*(18), 1541. 21

Courvoisier, D., C. Combescure, T. Agoritsas, A. Gayet-Ageron, and T. Perneger (2011a). Performance of logistic regression modeling: beyond the number of events

per variable, the role of data structure. *Journal of Clinical Epidemiology 64*(9), 993–1000. 14, 45, 46, 48, 50

Courvoisier, D., C. Combescure, T. Agoritsas, A. Gayet-Ageron, and T. Perneger (2011b). Performance of logistic regression modeling: beyond the number of events per variable, the role of data structure (letter to the editor). *Journal of Clinical Epidemiology 64*, 1465. 46

D'Agostino, R., S. Grundy, L. Sullivan, P. Wilson, and for the CHD Risk Prediction Group (2001). Validation of the framingham coronary heart disease prediction scores: Results of a multiple ethnic groups investigation. *JAMA 286*(2), 180–187. 21

D'Agostino, R., R. Vasan, M. Pencina, P. Wolf, M. Cobain, J. Massaro, and W. Kannel (2008). General cardiovascular risk profile for use in primary care. *Circulation 117*, 743–753. 21

Dorfman, D. and E. Alf (1969). Maximum likelihood estimation of parameters of signal detection theory and determination of condence intervals-rating-method data. *Journal of Mathematical Psychology 6*, 487–496. 34

Efron, B. (1975). The efficiency of logistic regression compared to normal linear discriminant analysis. *Journal of the American Statistical Association 70*, 892–898. 94

Flack, V. and P. Chang (1987). Frequency of selecting noise variables in subset regression analysis: A simulation study. *The American Statistician 41*(1), 84–86. 67

Gail, M., L. Brinton, D. Byar, D. Corle, S. Green, C. Schairer, and J. Mulvihill (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *JNCI: Journal of the National Cancer Institute 81*(24), 1879. 21

Gelman, A. and J. Hill (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, New York. 24

Glenn, W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review 78*(1), 1–3. 36

Goldstein, H. (2003). *Multilevel Statistical Models*. Edward Arnold, London. 15, 16, 24, 25

Guo, G. and H. Zhao (2000). Multilevel modeling for binary data. *Annual Review of Sociology 26*, 441–462. 15

Harbarth, S., N. Liassine, S. Dharan, P. Herrault, R. Auckenthaler, and D. Pittet (2000). Risk factors for persistent carriage of methicillin-resistant staphylococcus aureus. *Clinical Infectious Diseases 31*(6), 1380–1385. 42, 82

Harrell, F. (2001). *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer- Verlag. 14, 15, 20, 22, 27, 28, 29, 30, 31, 33, 34, 36, 40, 46, 48, 49, 50, 58, 70, 85, 106

Harrell, F., K. Lee, R. Calif, D. Pror, and R. Rosati (1984). regression modeling strategies for improved prognostic prediction. *Statistcs in medicine 3*, 143–152. 33, 34, 41, 42, 43, 82, 83, 113

Harrell, F., K. Lee, D. Matchar, and T. Reichert (1985). Regression models for prognostic prediction: advantages, problems, and suggested solutions. *Cancer Treat Rep. 69*(10), 1071–1077. 41, 42, 43

Harrell, F. E., K. Lee, and D. Mark (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine 15*(4), 361–387. 27, 29, 32, 42, 86, 91, 106

Hedeker, D., R. Gibbons, and J. Davis (1991). Random regression models for multicenter clinical trials data. *Psychopharmacol Bull 27*, 73–77. 107

Heinze, G. and M. Schemper (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine 21*, 21 2409–2419. 65

Hippisley-Cox, J., C. Coupland, J. Robson, A. Sheikh, and P. Brindle (2009). Predicting risk of type 2 diabetes in england and wales: prospective derivation and validation of qdscore. *BMJ 338*. 88

Hippisley-Cox, J., C. Coupland, Y. Vinogradova, J. Robson, M. May, and P. Brindle (2007). Derivation and validation of qrisk, a new cardiovascular disease risk score for the united kingdom: prospective open cohort study. *BMJ 335*(7611), 136. 21

Hippisley-Cox, J., C. Coupland, Y. Vinogradova, J. Robson, R. Minhas, A. Sheikh, and P. Brindle (2008). Predicting cardiovascular risk in england and wales: prospective derivation and validation of qrisk2. *BMJ 336*(7659), 1475–1482. 21, 88

Hosmer, D. and S. Lemeshow (2001). *Applied logistic regression*. A Wiley-InterScience Publication. 31, 32

Jinks, R., P. Royston, and M. Parmar (2015). Discrimination-based sample size calculations for multivariable prognostic models for time-to-event data. *BMC Medical Research Methodology 15*(1), 82. 89, 90

Jinks, R. C. (2012). Sample size for multivariable prognostic models. 14, 47, 48, 50

Jonas, F. and L. Johnny (2005). Parental smoking and risk of coeliac disease in offspring. *Scandinavian Journal of Gastroenterology 40*(3), 336–342. 42, 82

Judith, A. L., R. I. Jeannette, M. G. Thomas, and I. H. Ralph (2002). Social class and mortality in older women. *Journal of Clinical Epidemiology 55*(10), 952 – 958. 42, 82

Justice, A., K. Covinsky, and J. Berlin (1999). Assessing the generalizability of prognostic information. *Annals of Internal Medicine 130*(6), 515–524. 22, 32

Kahan, B. (2014). Accounting for centre-effects in multicentre trials with a binary outcome - when, why, and how? *BMC Medical Research Methodology 14*(20), 1471–2288. 24, 25

Keogh, B. and R. Kinsman (2003). Fifth national adult cardiac surgical database report. *s.l.:Society of Cardiothoracic Surgeons of Great Britain and Ireland*. 51

Kish, J. (1965). *survey sampling*. Health Administration Office Press. 108

Knaus, W., E. Draper, D. Wagner, and J. Zimmerman (1985). Apache ii: A severity of disease classification system. *Critical Care Medicine 13*, 818–829. 88

Konig, I., J. Malley, C. Weimar, H. Diener, and A. Ziegler (2007). Practical experiences on the necessity of external validation. *Statistics in Medicine 26*(30), 5499–5511. 31, 32

Kreft, I. and J. Leeuw (1998). *Introducing Multilevel Modeling*. Sage. 15, 107

Lassnigg, A., D. Schmidlin, M. Mouhieddine, L. M. Bachmann, W. Druml, P. Bauer, and M. Hiesmayr (2004). Minimal changes of serum creatinine predict prognosis in patients after cardiothoracic surgery: A prospective cohort study. *Journal of the American Society of Nephrology 15*(6), 1597–1605. 42, 82

Le Gall, J., S. Lemeshow, and F. Saulnier (1993). A new simplified acute physiology score (saps ii) based on a european/north american multicenter study. *JAMA 270*(24), 2957–2963. 88

Lee, Y. and J. Nelder (2004). Conditional and marginal models: another view. *Statistical Science 19*(2), 219–238. 25

Lemeshow, S., D. Teres, J. Klar, J. Avrunin, S. Gehlbach, and J. Rapoport (1993). Mortality probability models (mpm ii) based on an international cohort of intensive care unit patients. *JAMA 270*(20), 2478–2486. 88

Localio, A. R., J. A. Berlin, T. R. Ten Have, and S. E. Kimmel (2001). Adjustments for center in multicenter studies: An overview. *Annals of Internal Medicine 135*(2), 112–123. 16

Maas, K. and J. Hox (2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica 58*, 127–137. 16, 109, 111, 113

Maas, K. and J. Hox (2005). Sufficient sample sizes for multilevel modeling. *Methodology 58*, 86–92. 109, 110, 111, 113

Metz, C. (1978). Basic principles of roc analysis. *Seminars in Nuclear Medicine 8*, 283–298. 34

Miller, M., S. Hui, and W. Tierney (1991). Validation techniques for logistic regression models. *Statistcs in medicine 10*(8), 1213–1226. 29, 32, 33

Moineddin, R., F. Matheson, and R. Glazier (2007). A simulation study of sample size for multilevel logistic regression models. *BMC Medical Research Methodology 7*(1), 34. 110, 111, 113, 137, 153

Moons, K., D. Altman, Y. Vergouwe, and P. Royston (2009, May). Prognosis and prognostic research: Application and impact of prognostic models in clinical practice. *BMJ 338*, 1488–1490. 22, 27, 30, 32

Moons, K., P. Royston, Y. Vergouwe, D. Grobbee, and D. Altman (2009). Prognosis and prognostic research: what, why, and how? *BMJ 338*, 1317–1320. 13, 21, 32

Nagelkerke, N. (1991). A note on a general definition of the coefficient of determination. *78*(3), 691–692. 36

Neuhaus, J., J. Kalbfleisch, and W. Hauck (1991). A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review 59*, 25–36. 24

Ogundimu, E., D. Altman, and G. Collins (2016). Adequate sample size for developing prediction models is not simply related to events per variable. *Journal of Clinical Epidemiology 76*, 175 – 182. 46, 47, 48

Omar, R., G. Ambler, P. Royston, J. Eliahoo, and K. Taylor (2004). Cardiac surgery risk modeling for mortality: a review of current practice and suggestions for improvement. *The Annals of Thoracic Surgery 77*, 22327. 46

Paccagnella, O. (2011). Sample size and accuracy of estimates in multilevel models. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences 7*(3), 111–120. 110, 111

Peduzzi, P., J. Concato, A. R. Feinstein, and T. R. Holford (1995). Importance of events per independent variable in proportional hazards regression analysis ii. accuracy and precision of regression estimates. *Journal of clinical epidemiology 48*(2), 1503–1510. 14, 29, 42, 44, 50

Peduzzi, P., J. Concato, E. Kemper, T. Holford, and A. Feinstein (1996). A simulation study of the number of events per variable in logistic regression analysis. *Clinical Epidemiology 49*(12), 1373–1379. 41, 42, 44, 50

Peek, N., D. Arts, R. Bosman, P. van der Voort, and N. de Keizer (2007). External validation of prognostic models for critically ill patients required substantial sample sizes. *Journal of Clinical Epidemiology 60*(5), 491.e1 – 491.e13. 15, 85, 88, 91, 106

Rabe-Hesketh, S. and A. Skrondal (2005). *Multilevel and Longitudinal Modeling Using Stata*. Stata Press. 16, 24, 27

Rabe-Hesketh, S., A. Skrondal, and A. Pickles (2002). Reliable estimation of generalized linear mixed models using adaptivem quadrature. *The Stata Journal 1*, 1–21. 27

Robertson, N., C. Moore, G. Ambler, S. Bott, A. Freeman, G. Gambarota, C. Jameson, A. Mitra, B. Whitcher, M. Winkler, A. Kirkham, C. Allen, and M. Emberton (2013). Mapped study design: A 6 month randomised controlled study to evaluate the effect of dutasteride on prostate cancer volume using magnetic resonance imaging. *Contemp Clin Trials 34*, 80–89. 16, 107

Rosner, B. (2006). *Fundamentals of Biostatistics*. Belmont, Calif London. 41

Royston, P. and D. Altman (2013). External validation of a cox prognostic model: principles and methods. *BMJ 13*. 86

Royston, P., K. Moons, D. Altman, and Y. Vergouwe (2009). Prognosis and prognostic research: Developing a prognostic model. *BMJ 338*, 1373–1377. 14, 22, 27, 28

Royston, P. and W. Sauerbrei (2004). A new measure of prognostic separation in survival data. *Statistics in Medicine 23*(5), 723–748. 35

Royston, P. and W. Sauerbrei (2008). *Multivariable Model - Building: A Pragmatic Approach to Regression Analysis based on Fractional Polynomials for Modelling Continuous Variables* (1 ed.). Wiley. 48

Sauerbrei, W. and P. Royston (1999). 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) 162*(1), 71–94. 48

Simpson, J., N. Klar, and A. Donnor (1995). Accounting for cluster randomization: A review of primary prevention trials, 1990 through 1993. *American Journal of Public Health 85*(10), 13781383. 108

Skrondal, A. and S. Rabe-Hesketh (1984). Prediction in multilevel generalized linear models. *Journal of the Royal Statistical Society, Series A 172*(3), 659–687. 24, 25, 27

Spiegelhalter, D. (1986). Probabilistic prediction in patient management and clinical trials. *Statistcs in medicine 5*, 421–433. 32

Steele, F. (2017). *Module 7: Multilevel Models for Binary Responses*. https://www.cmm.bris.ac.uk/lemma/login/index.php: LEMMA VLE, University of Bristol, Centre for Multilevel Modelling. 24

Steyerberg, E., S. Bleeker, H. Moll, D. Grobbee, and K. Moons (2003). Internal and external validation of predictive models: a simulation study of bias and precision in small samples. *Journal of Clinical Epidemiology 56*(5), 441–447. 33

Steyerberg, E., M. Eijkemans, and J. Habbema (2001). Application of shrinkage techniques in logistic regression analysis: a case study. *Statistica Neerlandica 55*(1), 76–88. 29, 64

Steyerberg, E., M. Eijkemans, J. F. Harrell, and J. Habbema (2000). Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. *Statistics in Medicine 19*(8), 1059–1079. 14

Steyerberg, E., J. F. Harrell, G. Borsboom, M. Eijkemans, Y. Vergouwe, and J. Habbema (2001). Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *Journal of Clinical Epidemiology 54*, 774–781. 29

Steyerberg, E., M. Schemper, and F. Harrell (2011). Logistic regression modeling and the number of events per variable: selection bias dominates. *Journal of Clinical Epidemiology 64*, 1464–1465. 46

Steyerberg, E., A. Vickers, N. Cook, T. Gerds, M. Gonen, N. Obuchowski, M. Pencina, and M. Kattan (2010). Assessing the performance on prediction models: a framework for tradional and novel measures. *Epidemiology 1*, 128–138. 21

Steyerberg, E. W. (2009). *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Spriner. 14, 15, 20, 28, 30, 31, 33, 46, 48, 49, 73, 85, 106, 122

Steyerberg, E. W., G. J. J. M. Borsboom, H. C. van Houwelingen, M. J. C. Eijkemans, and J. D. F. Habbema (2004). Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Statistics in Medicine 23*(16), 2567–2586. 87

Stone, G. W., A. Maehara, A. J. Lansky, B. de Bruyne, E. Cristea, G. S. Mintz, R. Mehran, J. McPherson, N. Farhat, S. P. Marso, H. Parise, B. Templin, R. White, Z. Zhang, and P. W. Serruys (2011). A prospective natural-history study of coronary atherosclerosis. *New England Journal of Medicine 364*(3), 226–235. 42, 82

Thayyil, S., M. Chandrasekharan, A. Bainbridge, R. Omar, S. Murad, E. Cady, and N. Robertson (2010). Quantitative magnetic resonance biomarkers for prediction of long term outcome following neonatal encephalopathy: a meta-analysis. *Pediatrics 125*, 382–395. 107

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B 58*, 267–288. 49

Twisk, J. (2006). *Applied Multilevel Analysis: A Practical Guide*. Cambridge University Press. 15, 16, 24, 26, 107, 131

Van Houwelingen, J. and S. Le Cessie (1990). Predictive value of statistical models. *Statistcs in medicine 9(11)*, 1303–1325. 29, 33, 49

van Smeden, M., J. de Groot, K. Moons, G. Collins, D. Altman, M. Eijkemans, and J. Reitsma (2016). No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC Medical Research Methodology 16*(1), 163. 47

Vergouwe, Y., E. Steyerberg, M. Eijkemans, and J. Habbema (2002). Validity of prognostic models: when is a model clinically useful? *Seminars in Urologic Oncology 20*, 96–107. 85, 106

Vergouwe, Y., E. Steyerberg, M. Eijkemans, and J. Habbema (2005). Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *Journal of Clinical Epidemiology 58*(5), 475–483. 15, 28, 30, 32, 34, 85, 86, 87, 91, 106

Verweij, P. and H. van Houwelingen (1994). Penalized likelihood in cox regression. *Statistics in Medicine 13*, 2427–2436. 49

Vittinghoff, E. and C. McCulloch (2007). Relaxing the rule of ten events per variable in logistic and cox regression. *American Journal of Epidemiology 165*(6). 14, 44, 45, 50

Voerman, G. E., L. Sandsj, M. M. R. Vollenbroek-Hutten, P. Larsman, R. Kadefors, and H. J. Hermens (2007). Changes in cognitive-behavioral factors and muscle activation patterns after interventions for work-related neck-shoulder complaints: Relations with discomfort and disability. *Journal of Occupational Rehabilitation 17*(4), 593–609. 42, 82

William, R. L., B. S. Jane, S. C. Matthew, H. A. Jessie, C. Victoria, and J. B. Edward (2013). Diabetic foot ulcer incidence in relation to plantar pressure magnitude and measurement location. *Journal of Diabetes and its Complications 27*(6), 621 – 626. 42, 82

Wilson, P., R. D'Agostino, D. Levy, A. Belanger, H. Silbershatz, and W. Kannel (1998). Prediction of coronary heart disease using risk factor categories. *Circulation 97(97)*, 1837–1847. 21

Wynants, L., W. Bouwmeester, M. Moonsc, KGM Moerbeeke, D. Timmerman, S. Van Huffel, B. Van Calsterf, and Y. Vergouwe (2015). A simulation study of sample size demonstrated the importance of the number of events per variable to develop prediction models in clustered data. *Journal of Clinical Epidemiology*. 16, 18, 37, 112, 113, 117, 137, 153

Wynants, L., Y. Vergouwe, S. Van Huffel, D. Timmerman, and B. Van Calster (2016). Does ignoring clustering in multicenter data influence the performance of prediction models? a simulation study. *Statistical Methods in Medical Research*. 25

Zeger, S., K. Liang, and P. Albert (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrika 44*, 1049–1060. 23