Validation measures for prognostic models for independent and correlated binary and survival outcomes

Mohammad Shafiqur Rahman

Department of Statistical Science

University College London

A thesis submitted for the degree of

Doctor of Philosophy

October 2012

In loving memory to my late father.

Declaration

I herewith declare 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 this thesis with an appropriate reference.

.....

Mohammad Shafiqur Rahman

Abstract

Prognostic models are developed to guide the clinical management of patients or to assess the performance of health institutions. It is essential that performances of these models are evaluated using appropriate validation measures. Despite the proposal of several validation measures for survival outcomes, it is still unclear which measures should be generally used in practice. In this thesis, a simulation study was performed to investigate a range of validation measures for survival outcomes in order to make practical recommendations regarding their use. Measures were evaluated with respect to their robustness to censoring and their sensitivity to the omission of important predictors. Based on the simulation results, from the discrimination measures, Gönen and Heller's K statistic can be recommended for validating a survival risk model developed using the Cox proportional hazards model, since it is both robust to censoring and reasonably sensitive to predictor omission. Royston and Sauerbrei's D statistic can be recommended provided that the distribution of the prognostic index is approximately normal. Harrell's C-index was affected by censoring and cannot be recommended for use with data with more than 30% censoring. The calibration slope can be recommended as a measure of calibration since it is not affected by censoring. The measures of predictive accuracy and explained variation (Graf *et al*'s integrated Brier Score and its R^2 version, and Schemper and Henderson's V) cannot be recommended due to their poor performance in the presence of censored data.

In multicentre studies patients are typically clustered within centres and are likely to be correlated. Typically, random effects logistic and frailty models are fitted to clustered binary and survival outcomes, respectively. However, limited work has been done to assess the predictive ability of these models. This research extended existing validation measures for independent data, such as the C-index, D statistic, calibration slope, Brier score, and the K statistic for use with random effects/frailty models. Two approaches: the 'overall' and 'pooled cluster-specific' are proposed. The 'overall' approach incorporates comparisons of subjects both within-and between-clusters.

The 'pooled cluster-specific' measures are obtained by pooling the cluster-specific estimates based on comparisons of subjects within each cluster; the pooling is achieved using a random effects summary statistics method. Each approach can produce three different values for the validation measures, depending on the type of predictions: conditional predictions using the estimates of the random effects or setting these as zero and marginal predictions by integrating out the random effects. Their performances were investigated using simulation studies. The 'overall' measures based on the conditional predictions including the random effects performed reasonably well in a range of scenarios and are recommended for validating models when using subjects from the same clusters as the development data. The measures based on the marginal predictions and the conditional predictions that set the random effects to be zero were biased when the intra-cluster correlation was moderate to high and can be used for subjects in new clusters when the intra-cluster correlation coefficient is less than 0.05. The 'pooled cluster-specific' measures performed well when the clusters had reasonable number of events. Generally, both the 'overall' and 'pooled' measures are recommended for use in practice.

In choosing a validation measure, the following characteristics of the validation data should be investigated: the level of censoring (for survival outcome), the distribution of the prognostic index, whether the clusters are the same or different to those in the development data, the level of clustering and the cluster size.

Thesis supervisor: Dr. Rumana Omar; Co-supervisor: Dr. Gareth Ambler

Acknowledgements

First of all, I would like to thank my supervisor Dr. Rumana Omar for her valuable guidance and supervision in the development of this thesis. She has been extremely patient and encouraging from beginning to the end of this work. I am also very much grateful to my co-supervisor Dr. Gareth Ambler for providing constructive ideas, comments, and suggestions throughout this research.

Besides, I would like to thank the authority of Overseas Research Student Award Scheme (ORSAS) and University College London (UCL) for providing my tuition fees. Also I wish to express my sincere gratitude to the Department of Statistical Sciences, UCL for their financial support during my stay at London. Special thanks to all staff, Postdocs, and PhD students in this department for helping me in many ways.

I would like to thank Drs Constantinos O'Mahony and Perry Elliott for allowing me to use their data for simulation purposes.

Many thanks to all of my colleagues at the Institute of Statistical Research and Training, University of Dhaka for their continuous support during my study leave from Dhaka University.

Finally, I am grateful to my wife and family for their patience and support during last three years.

Contents

List of Figures **xii**

List of Tables **xvi**

1 Introduction **1**

1.1 Context of this research 1

1.2 Objectives of this research 5

1.3 Organization of this research 6

2 Validating a prognostic model **7**

2.1 Introduction 7

2.2 Motivation for validating a prognostic model 7

2.3 Validation procedure 9

2.3.1 Design of a validation study 9

2.3.1.1 Internal validation 9

2.3.1.2 Temporal validation 10

2.3.1.3 External validation 10

2.3.2 Key aspects of a model that need to be validated 11

2.3.2.1 Calibration 11

2.3.2.2 Discrimination 11

2.3.2.3 Distinction between calibration and discrimination . . . 12

2.3.2.4 Overall performance 13

2.4 Measures that assess the predictive ability of a model 13

2.4.1 Measures of calibration 13

2.4.2 Measures of discrimination 16
2.4.2.1 Measures based on concordance probability 16
2.4.2.2 Measure based on prognostic separation 17
2.4.3 Overall performance measures: R^2 type measures 18
2.4.3.1 Measures of explained variation: based on loss function approach 19
2.5 Conclusion 20

3 Measures for independent survival outcomes 21
3.1 Introduction 21
3.2 Example data sets 22
3.2.1 Breast cancer data 22
3.2.2 Sudden cardiac death data 23
3.3 Validation measures for the Cox Proportional Hazards model 23
3.3.1 The Cox Proportional Hazards model 24
3.3.2 Measures of calibration 24
3.3.2.1 Calibration slope (CS) 25
3.3.3 Measures of discrimination 25
3.3.3.1 Harrell's C-index 26
3.3.3.2 Gönen and Heller's $K(\beta)$ 27
3.3.3.3 Royston and Sauerbrei's D 28
3.3.4 Measures of predictive accuracy and explained variation 29
3.3.4.1 Graf et al's Brier score 29
3.3.4.2 Graf et al's R^2_{IBS} 31
3.3.4.3 Schemper and Henderson's V 31
3.4 Evaluation of the measures 33
3.4.1 Criteria for evaluation 33
3.4.2 Simulation design 34
3.4.2.1 Simulation scenarios 34
3.4.2.2 Generating new survival and censoring times 35
3.4.2.3 Generating validation data with different risk profiles . 35
3.4.3 Assessing the effect of censoring 37
3.4.4 Assessing sensitivity to the exclusion of important predictors . . 38

3.5 Results and discussion . . . 39
3.5.1 Results . . . 39
3.5.1.1 Effect of censoring . . . 39
3.5.1.2 Sensitivity to the exclusion of important predictors . . . 44
3.5.1.3 Relationship between the validation measures . . . 47
3.5.2 Discussion and recommendations . . . 49
3.6 Conclusion . . . 51

4 Measures for clustered binary outcomes 52
4.1 Introduction . . . 52
4.2 Validation measures for independent binary outcomes . . . 54
4.2.1 Logistic regression model . . . 54
4.2.2 The C-index: definition . . . 55
4.2.3 Non-parametric estimation of the C-index . . . 56
4.2.4 Parametric estimation of the C-index . . . 57
4.2.5 D statistic . . . 57
4.2.6 Relationship between the C-index and D statistic . . . 58
4.2.7 Calibration slope . . . 60
4.2.8 Brier score . . . 60
4.3 Extension of the validation measures for clustered data . . . 61
4.3.1 Random-intercept logistic model . . . 61
4.3.2 Predictions from the model . . . 62
4.3.3 Approaches for the calculation of the validation measures for clustered data . . . 64
4.3.4 Estimation: Overall measure . . . 66
4.3.4.1 The C-index for clustered data: definition . . . 66
4.3.4.2 Nonparametric estimation of the C-index . . . 67
4.3.4.3 Parametric estimation of the C-index . . . 71
4.3.4.4 D statistic . . . 74
4.3.4.5 Calibration slope . . . 74
4.3.4.6 Brier score . . . 75
4.3.5 Estimation: Pooled cluster-specific measure . . . 76
4.4 Application to clustered binary data . . . 78

4.4.1 Heart valve surgery data . . . 78
4.4.2 Analysis and results . . . 79
4.4.2.1 Model development . . . 79
4.4.2.2 Model validation . . . 79
4.5 Simulation study . . . 85
4.5.1 Simulation design . . . 86
4.5.1.1 True model . . . 86
4.5.1.2 Simulation scenarios . . . 87
4.5.2 Strategies for evaluating the measures . . . 87
4.5.2.1 Model fitting and calculation of the measures . . . 87
4.5.2.2 Assessing the properties . . . 88
4.5.3 Results . . . 89
4.5.3.1 The Overall validation measures . . . 89
4.5.3.2 The Pooled cluster-specific validation measures . . . 97
4.6 Conclusion . . . 102

5 Measures for clustered survival outcomes 105
5.1 Introduction . . . 105
5.2 Extension of the validation measures for use with clustered survival data 106
5.2.1 The Proportional Hazards frailty model . . . 106
5.2.2 Predictions from the frailty model . . . 108
5.2.3 Approaches for the calculation of the validation measures . . . 108
5.2.4 Estimation: Overall measures . . . 109
5.2.4.1 Harrell's C-index . . . 109
5.2.4.2 Gonen and Heller's $K(\beta)$ . . . 113
5.2.4.3 Royston and Sauerbrei's D . . . 116
5.2.4.4 Calibration slope . . . 117
5.2.4.5 Brier score . . . 117
5.2.5 Estimation: Pooled cluster-specific measures . . . 118
5.3 Application to child mortality data . . . 118
5.3.1 Child mortality data . . . 119
5.3.2 Analysis and results . . . 122
5.3.2.1 Model development . . . 122

5.3.2.2 Model Validation 123
5.4 Simulation study 129
5.4.1 Simulation design 129
5.4.1.1 True model 129
5.4.1.2 Simulation scenarios 130
5.4.2 Strategies for evaluating the measures 131
5.4.2.1 Model fitting and calculation of the validation measures 131
5.4.2.2 Assessing the properties of the measures 131
5.4.3 Results 132
5.4.3.1 The Overall validation measures 132
5.4.3.2 Pooled cluster-specific validation measures 139
5.5 Conclusion 142

6 Summary and Conclusions **146**
6.1 Summary of the research 146
6.2 Summary of the methods and results 147
6.2.1 Validation measures for independent survival outcomes 147
6.2.2 Validation measures for clustered data 149
6.3 Conclusions 154
6.4 Possibilities for future research 154

A Additional Results for Chapter 3 **157**

B Stata code for validation measures **160**

Bibliography **174**

List of Figures

2.1 Plots to show the distinction between calibration and discrimination. Plots are for two hypothetical models (M1 and M2) with equal (perfect) calibration but different discriminatory abilities. 12

2.2 Theoretical calibration plots to assess the agreement between the observed proportion and predicted probability with a dot line through all outcome value (0 and 1): (a) $\hat{\alpha} = 0$, $\hat{\beta} = 1$; (b) $\hat{\alpha} = 0$, $\hat{\beta} = 0.74$; (c) $\hat{\alpha} = -0.65$, $\hat{\beta} = 1$; (d) $\hat{\alpha} = -0.65$, $\hat{\beta} = 0.74$. 14

3.1 Empirical distribution of the validation measures by degree of censoring was summarised using box plots. The results are from the medium risk breast cancer simulations under the random censoring mechanism. The horizontal dashed line indicates the true/reference value of the respective measure. 40

3.2 Relative bias (%) with 95% confidence intervals for the C-index, $K(\beta)$, D statistic, Calibration slope, and Integrated Brier score (IBS). The first and second rows show the results for the breast cancer and sudden cardiac death simulations with different risks profile (low, medium, and high), respectively. All simulations were under the random censoring mechanism. 41

3.3 Relative bias (%) with 95% confidence intervals for R^2_{IBS} and V. The first and second rows show the results for the breast cancer and sudden cardiac death simulations with different risks profile (low, medium, and high), respectively. All simulations were under the random censoring mechanism. 42

3.4 Distribution of the prognostic index derived from the true model for the breast cancer and sudden cardiac death patients with different risk profiles. 43
3.5 Sensitivity of the measures to the exclusion of important predictors, described as the percentage of reduction in a measure's value relative to that for the full model. The results are from the breast cancer simulations with different risk profiles. No censoring was considered. 46
3.6 Empirical agreement between the measures by degrees of censoring. The results are from the medium risk breast cancer simulation under the random censoring mechanism. 48

4.1 Distribution of the predicted probability, $\Pr(Y = 1)$, by types of prediction such as $\pi_{ij}(u)$, $\pi_{ij}(0)$, and $\pi_{ij}(\text{pa})$. 80
4.2 Distribution of the predicted prognostic index (PI) or log odds for the population who survived and those who died by types of prediction: (a) $\pi_{ij}(0)$, (b) $\pi_{ij}(u)$, and (c) $\pi_{ij}(\text{pa})$. 81
4.3 Cluster (hospital)-specific estimates against their rank order: the C-index (non-parametric), D statistic, calibration slope, and Brier score. Each horizontal solid line indicates the 'pooled estimate' of the respective measures. 84
4.4 Relative bias (%) in the 'overall' estimates of the validation measures for different ICC values. The results are from the different simulation scenarios based on the number of clusters and their size (clusters×size). Each column represents plots of bias for the different estimates of a validation measure based on the model prediction $\hat{\pi}_{ij}(u)$, $\hat{\pi}_{ij}(0)$, and $\hat{\pi}_{ij}(\text{pa})$. 90
4.5 Agreement between the estimated ($\hat{u}$) and the true random effects u in the validation data. The results are from the different simulation scenarios under ICC=20%: number of clusters (a) 10 of size 10, (b) 10 of size 300, (c) 100 of size 10, and (d) 100 of size 100. Figure 2b shows nine points, because two points amongst the ten correspond to the same values and hence represents one point. 91

4.6 Relative bias (%) in the 'overall' estimates of the validation measures for $\pi_{ij}(u)$ when they were calculated using the true values of the random effects u, rather than the estimates. The results are from the different simulation scenarios (clusters×size). 92

4.7 Relative rMSE (%) of the 'overall' estimates of the validation measures for different ICC values. The results are from the different simulations scenarios (clusters×size). Each column represents plots of rMSE for different estimates of a validation measure based on $\hat{\pi}_{ij}(u)$, $\hat{\pi}_{ij}(0)$, and $\hat{\pi}_{ij}(\text{pa})$. 94

4.8 Relative bias (%) in the 'overall' estimates of the validation measures for different ICC values. The results are from the different simulation scenarios based on unequal cluster sizes: 30 clusters with median sizes 50 or 145. Each column represents plots of bias for the different estimates of a validation measure based on the model prediction $\hat{\pi}_{ij}(u)$, $\hat{\pi}_{ij}(0)$, and $\hat{\pi}_{ij}(\text{pa})$. 96

4.9 Relative bias (%) in the 'pooled' estimate of the validation measures for different ICC values. The results are from the different simulations scenarios (clusters×size). 97

4.10 Relative bias (%) in the 'pooled' estimates of the C-index and D statistic when calculating bias against the 'overall' true values. The results are from the different simulations scenarios (clusters×size). 98

4.11 Relative rMSE (%) of the 'pooled' estimates of the validation measures for different ICC values. The results are from the different simulations scenarios (clusters×size). 100

4.12 Relative bias (%) in the 'pooled' estimate of the validation measures for different ICC values. The results are from the simulations based on unequal cluster sizes. 101

5.1 Map of Bangladesh indicating the distribution of the urban and rural sampling points (a total of 361 clusters), visited in the 2007 BDHS survey. Source: 2007 BDHS report. 120

5.2 Distribution of clusters in both the development and validation data by the number of births per cluster (a,c) and by the number of deaths per cluster (b,d). 121
5.3 (a) Distribution of the predicted prognostic index $\hat{\eta}_{re} = \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_{ij} + \hat{\varpi}_j$ (b) Kaplan-Meier survival function at the tertiles of the predicted prognostic index. 124
5.4 Brier scores over the entire follow-up period. The results are obtained for the predictions from the model with all fixed predictors and the frailties, the model with all fixed predictors only, and the null model. 126
5.5 Cluster-specific estimates of the validation measures with their level of uncertainty against the rank order of the clusters. The horizontal solid line indicates the pooled estimate of the measure. 128
5.6 Empirical (sampling) distribution of the validation measures by degree of censoring, summarised using box plots. The results are from the simulations with 10 clusters of size 100 under $\theta = 0.98$. The horizontal (dashed) lines indicate the true values of the measures. 133
5.7 Relative bias (%) in the 'overall' estimate of the validation measures for degrees of censoring. The results are from the different simulation scenarios based on the number of clusters, their sizes, and the frailty parameter θ. 134
5.8 Relative rMSE (%) of the 'overall' estimates of the validation measures for different degrees of censoring. The results are from the different simulation scenarios. 136
5.9 Agreement between the validation measures for different degrees of censoring. The results are from the simulations with 50 clusters of size 30 under $\theta = 0.98$. The r values indicate the estimated Pearson correlation coefficients between the measures. 138
5.10 Relative bias (%) in the 'pooled estimates' of cluster-specific validation measures for different degrees of censoring. The results are from the different simulation scenarios. 140
5.11 Relative rMSE (%) of the 'pooled estimates' of the cluster-specific validation measures for different degrees of censoring. The results are from the different simulation scenarios. 141

List of Tables

3.1 Risk profile of patients in validation scenarios, described by the failure probability ($1 - \hat{S}(t^*|x)$) estimated at a single time point t^*: the overall probability and tertiles (mean over 500 simulations of uncensored data, maximum Monte Carlo standard error=0.0007). The distribution of the true prognostic index is also discussed. 37

3.2 Models with different predictive abilities, relative to the full model, are summarised in terms of R^2 values. The results are from the breast cancer simulations with different risk profiles. No censoring was considered. . . 45

4.1 Estimates of the validation measures for the model predicting in-hospital mortality following heart valve surgery in the validation sample. 82

4.2 Coverage (%) of nominal 90% confidence intervals (CIs) of the 'overall' validation measures. The confidence interval are based on both analytical and bootstrap standard errors. Maximum Monte Carlo Standard Error=2.25%. 95

4.3 Distribution of the number of clusters dropped when calculating validation measures within a cluster. The results are presented by the number of events required to calculate a measure. Each figure is the average over 500 simulations. 99

4.4 Coverage (%) of nominal 90% confidence intervals (CIs) for the 'pooled' estimates of the cluster-specific measures. The CIs are based on analytical standard errors of the measure. Maximum Monte Carlo Standard Error = 2.37%. 101

5.1 Estimates of the PH frailty model in the development data 123
5.2 Estimates of the validation measures based on the validation data . . . 126
5.3 Estimated coverage of nominal 90% confidence intervals for the 'overall' measures. The confidence intervals were calculated based on bootstrap standard errors. The results are from all simulation scenarios where the level of clustering was high ($\theta = 0.98$). Maximum Monte Carlo Standard Error=2.4%. 137
5.4 Coverage of 90% nominal confidence intervals for the 'pooled cluster-specific' measures. The confidence intervals were calculated based on analytical standard errors. The results are from the different simulation scenarios under $\theta = 0.98$. Maximum Monte Carlo Standard Error=2.5%. 142

A.1 Relative bias (%) and 95% CIs are given by censoring proportions. The results are from the (a) breast cancer simulations (maximum Monte Carlo standard error (%)=0.88) and (b) sudden cardiac death simulations (maximum Monte Carlo standard error (%)=0.82), with different risks profile (low, medium, and high) and under administrative censoring mechanism. 158
A.2 The Cox model estimates for the breast cancer data 159
A.3 The Cox model estimates for the sudden cardiac death data 159

Chapter 1

Introduction

1.1 Context of this research

In medicine, prognosis literally means forecasting, predicting or estimating the probability or risk of an individual's future health outcomes, such as illness, or complication, or death. For example, in oncology, it may be important to predict the probability of survival beyond a specific time point for cancer patients, and, in cardiology, to predict the risk of developing a cardiovascular disease or death from a cardiovascular disease. Prognostic studies are usually carried out to predict patients' future health status as accurately as possible using their clinical and demographic characteristics. For example, the study carried out by Ambler et al. [1] focuses on predicting the risk of in-hospital mortality for patients following heart valve surgery. Similarly, the Nottingham prognostic index derived by Galea et al. [2] is used to estimate the risk of cancer recurrence or death in breast cancer patients.

Prognostic studies are similar to aetiological studies in terms of design and analysis, but have different purposes: the former focuses on predicting health outcome of interest while the latter on explaining their causes [3]. In particular, aetiological studies investigate the association between risk factors and an outcome of interest, with possible adjustment for other factors (confounders), typically using a multivariable statistical model. Prognostic studies also use a multivariable statistical model to identify

all important predictors that are potentially associated with the outcome and, using a combination of these, provide prediction algorithms or rules to predict the risk of future outcome. These algorithms are commonly known, in the literature, as prognostic models or prediction models [4–10].

Prognostic models are increasingly being used in various settings of clinical research such as cardiology, intensive care medicine, and oncology to estimate individual patients' prognosis and/or to classify patients into clinical risk groups with different prognoses, for example, low, medium, and high. The clinical use of these models mainly consists in providing information for patients about the future course of their illness (or their risk of developing illness) and in guiding doctors on joint decisions with patients to plan for possible treatment.

Prognostic models may be useful in cost effectiveness programs or to select appropriate tests or therapies in patient management including decisions on withholding or withdrawing therapy. For example, models may be used to classify patients with good prognosis for whom adjuvant therapy would not be (cost-)effective, or a group of patients with a poor prognosis for whom more aggressive adjuvant therapy would not be justified [9, 11]. These models may also be used to select homogeneous groups of patients for clinical trials, for example, to select patients with a low risk of cancer recurrence for a randomised trial on the efficacy of radiotherapy after breast conserving resection. Finally, prognostic models may be used to assess the performance of clinicians or hospitals and to conduct comparisons between them after adjusting for the case-mix of patients. For example, the clinical risk index for babies (CRIB) [12] is used to predict the risk of mortality for newborn babies and to assist comparative assessment across the neonatal intensive care units by case-mix-adjusted risk predictions.

Although the use of prognostic models in clinical management is promising, clinicians will be reluctant to use these models unless they can trust their predictions [13]. Therefore, the prime goal of prognostic studies should be to develop such a model which is statistically valid and clinically useful. To facilitate such clinical prognostication successfully, researchers [4–10, 14, 15] have paid attention to the methodological aspects of

prognostic studies and models, particularly focusing on the validation of models' predictive performance. The general idea of validating a prognostic model is to establish that it performs well for patients other than those used to develop the model [7, 14–16].

Prognostic models are usually developed using multivariable regression models. For example, logistic regression is commonly used for binary outcomes while Cox proportional hazards regression is used for survival outcomes. Often an index is developed from a prognostic model based on weighted sum of the predictors in the model, where the weights are the estimated regression coefficients. This is known as the 'prognostic index' and can be used to classify patients into different risk groups, for example, low, medium, and high. Building a prognostic model from a set of candidate predictors is a complex process [17–19], and there is no widely agreed approach to this. However, the importance of carefully dealing with some of the statistical and clinical aspects of developing a prognostic model, such as choosing clinically relevant patient sample, selecting important predictors, modeling continuous predictors, having adequate sample size, and handling missing data, if any, is widely accepted. These aspects are discussed in details in some recent studies; see, for example, Royston et al. [4], Altman [9], and Omar et al. [10].

Compared to the methodology published in the literature on the development of prognostic models, the methodology for validating their predictive performance is not well developed [7, 20]. However, validation of the predictive performance of a newly developed model or a model updated from an existing one plays a key role in prognostic studies. This research focuses on the methodological aspects of validating a prognostic model. Validating a prognostic model implies gaining evidence that it performs well for new patients different from those used to develop the model. This idea is motivated by the fact that the predictive ability of a model is likely to be overestimated in the sample of patients used to develop the model (training/development data), compared to the predictive ability of the model in other patient samples (test/validation data), even if both samples are derived from the same population [7, 8, 15, 21].

Different types of validation process have been discussed in the literature [7, 16].

The most commonly used processes include (i) splitting a single dataset (randomly or based on time) into two parts, one of which is used to develop the model and the other used for validation, (internal or temporal validation) and (ii) validating on an independent dataset collected by different centres or investigators (external validation). Apart from the type of validation process, there are several aspects of a model that are usually assessed on new data. These include (i) the agreement between the observed and predicted outcome of interest for a group of patients (calibration) or individual patients (accuracy scores), (ii) the ability of the model to distinguish between patients who do or do not experience the outcome of interest (discrimination) [7, 22]. Another aspect that is sometimes used to assess the model's predictive performance is the concept of 'explained variation', which refers to the proportion of variation in the outcome that can be explained by the predictors in the model [23]. Intuitively, high explained variation depends on making a wide range of accurate predictions. This aspect captures both the calibration and discrimination of the model.

The methodology for validating prognostic models with independent binary outcomes are reasonably well developed; for example, see Omar et al. [10], Steyerberg et al. [24], and Royston and Altman [25]. Although a number of validation measures have been proposed for survival outcomes, it is still unclear which measures should be used in practice. This research evaluates some of the proposed measures in order to make practical recommendations. Furthermore, patients' health outcomes may be clustered within large units. For example, in a multi-centre study, patients within the same hospitals are likely to be more similar compared to patients across hospitals. This correlation between patients within a hospital is known as clustering. Random effects logistic and frailty models which can take account of this clustering have been proposed for the analysis of clustered binary and survival outcomes, respectively. In risk prediction research, this clustering is often ignored both in the process of model development and the validation of the models predictive performance. Limited work has been done to assess the predictive ability of models developed with clustered outcomes, regardless of the types of outcomes (binary or survival). This research also focuses on the use of these models for risk prediction for clustered data and the validation methods for assessing the predictive performance of the models.

1.2 Objectives of this research

The primary objective of this research is to consider the methodological aspects of validating a prognostic model, particularly focusing on the validation measures that could be useful in assessing the predictive performance of the model.

Several validation measures have been proposed for models with independent survival outcomes, but it is still not clear what measures should be generally used. One of the objectives of this research is to review some of the proposed measures and evaluate these by a simulation study based on two real datasets in order to make recommendations for their use in practice.

Furthermore, limited work has been done for validation measures for models developed with clustered outcomes. This research discusses possible extensions of some of the standard validation measures that have been used for independent binary or survival outcomes for use with models for clustered outcomes. An application of these measures is illustrated using data on patients who underwent heart valve surgery (binary outcome: in-hospital mortality) and child mortality data (survival outcome: time to event, died/alive, by the 5th birthday). A simulation study is further conducted to investigate the properties of the new measures under various simulation scenarios formulated by varying the number of clusters and their size, varying the intra-cluster correlation between subjects within a cluster, and for survival outcomes, varying the degree of censoring.

The real data presented in this thesis are mainly used to illustrate and evaluate validation measures. The clinical motivation for developing the risk models is not the main focus here; it is assumed that the model development has been carried out appropriately.

1.3 Organization of this research

This research is organised as follows. Chapter 2 discusses the motivation and general procedures for validating a prognostic model. This chapter also discusses a literature review of validation measures that have been proposed for models with binary and survival outcomes.

Chapter 3 evaluates some of the validation measures for models with independent survival data. In particular, this chapter includes a motivation for choosing the measures to be evaluated, their calculation for the Cox proportional hazards model, and a simulation study based on two clinical datasets.

In Chapter 4, some of the standard validation measures that have been used for independent binary outcomes are extended for use with models for clustered binary outcomes. This chapter particularly discusses the detailed calculation of these measures for random intercept logistic model, starting with a description of the model and its possible approaches to prediction. An illustration of these methods using real data and a simulation study to assess their performance are also discussed.

Chapter 5 discusses possible extensions of some of the standard validation measures for use with models for clustered survival data. Specifically, this chapter discusses a frailty model along with its different approaches to prediction, and the detailed calculation of the validation measures for the frailty model. This chapter also includes an illustration of the new methods using child mortality data and a simulation study to assess the properties of the methods.

Chapter 6 starts by summarising the research and the findings, then discusses some recommendations for use in practice, and ends by discussing the possibilities for future research.

Chapter 2

Validating a prognostic model

2.1 Introduction

A vital aspect of the prognostic modelling process is to consider whether a model developed using a patient-sample is transportable to other patients from a relevant but different population, who are different in terms of patient characteristics. This concept is generally referred to as validity (or generalisability), and a model that is found to have such quality is said to have been validated [26]. A validated model may be recommended for use as a clinical decision making tool. Different types of model validity have been discussed in the literature. These include 'internal', 'temporal', and 'external' validity. When conducting a validation study, there are some key aspects of a model that need to be evaluated. These include 'calibration', 'discrimination' and 'explained variation'. This chapter discusses all these aspects of validating a prognostic model and includes a brief literature review of existing validation measures, starting with a motivation for validating a model.

2.2 Motivation for validating a prognostic model

There are several reasons why we need to validate the performance of a prognostic model. One of the main reasons is that we need evidence of the accuracy of predictions made by the model before using it in clinical practice. Furthermore, a prognostic model that accurately predicts for patients in the development data may not perform

similarly for new patients from a different but relevant population [7, 26]. Therefore, to establish that the model is useful to clinicians who would use it, it is essential to evaluate (validate) its predictive performance particularly on new data different from which the model was derived. There are several statistical and clinical reasons discussed in the literature [7, 14, 16, 20, 27] explaining why a prognostic model may perform poorly on new data. The reasons are discussed below:

(i) *Inadequate design of prognostic studies:*
The inadequate design of prognostic factor studies may lead to overoptimistic results [7, 18]. The design may be inadequate if there is no standard inclusion and exclusion criteria for selecting patients (many patients may be excluded because of missing data), no justification for the choice of treatments, an inadequate sample size, and an inadequate number of events of interest per predictor. For more details, see Altman and Royston [7], Harrell et al. [28], and Peduzzi et al. [29, 30].

(ii) *Lack of a standard approach to developing the model:*
A prognostic model may not perform well for new patients if the model was inadequately developed in the original sample. The model may be developed, for example, using 'stepwise variable selection algorithm', by which one selects the best model from many alternative models, but these methods usually have data-dependent aspects. These aspects are likely to lead to an overoptimistic assessment of the predictive performance. For more details, see Altman and Royston [7], Altman et al. [14].

(iii) *Differences between patients' characteristics in the development and validation data*:
Even if the model is adequately developed, it may not perform well for new patients. This may be because there are differences between the characteristics of the patients in the development and validation data. This is known as a difference in 'case-mix' [7, 15, 20]. Both the discrimination and calibration of a model could be affected by the difference in 'case-mix'. For example, if age is one of the predictors included in the model and ranges from 60 to 80 years in the validation sample and 20 to 80 years in the development sample, then the discrimination (between patients who experienced the outcome and who did not) in the more

homogeneous validation population would be expected to be worse than in the more heterogenous development population [20]. Another example would be if the validation sample contains relatively more patients with hypertension than the development sample, and presence of hypertension increases the probability of the outcome but hypertension was not included in the model (missed predictor), then the predicted probability derived from the model may be underestimated in the validation population [15, 20].

2.3 Validation procedure

This section discusses a procedure for validating a prognostic model, following the validation strategies discussed in the literature [7, 14]. Typically, a validation procedure involves (i) designing a validation study, which describes the development and validation data and what type of validation process one should choose, and (ii) identifying the aspects of the model, for example, calibration and discrimination that need to be validated.

2.3.1 Design of a validation study

The main validation processes discussed in the literature are internal validation, temporal validation, and external validation study [7, 14, 16, 20, 27]. These are now discussed in the following subsections.

2.3.1.1 Internal validation

The key feature of internal validation process is that only one dataset (the primary data) is used. The most common approach is to randomly split the data into two parts (often 2:1) before model development begins [7, 14, 20]. The first part, which is usually called the 'development' set, is used to develop the model and the second part, called the 'validation' set, is used to evaluate the model's predictive performance. This data-splitting process has some limitations. For example, this process is likely to provide overoptimistic results on the model's performance. In addition, the estimates of the predictive performance from this procedure may be unbiased, but they tend to be imprecise [31]. A possible reason is that both datasets are very similar as they

were extracted from the same underlying population. Furthermore, one issue that commonly arises in data-splitting process is how to split the data; there is no guideline on what proportion of patients should be in the development and validation sets [7, 32]. Some alternative, but better, approaches are to use a resampling technique such as 'bootstrapping' and 'cross-validation'. These resampling techniques are commonly used to overcome overoptimism [7, 33–35].

Briefly, bootstrapping [36] involves taking a large number of samples with replacement from the original sample, of the same size as the original data set. Then models may be developed in the bootstrap samples and validated in the original sample. In cross-validation, for example, k-fold cross-validation, the original sample is partitioned into k subsets, one of which is used to validate the model and the remaining $k-1$ subsets are used to develop the model. This procedure is repeated k times. To improve the efficiency of the cross-validation, the whole procedure can be repeated several times taking new random subsamples [35]. The most extreme cross-validation technique is to leave one subject out at a time, which is equivalent to the jack-knife technique [36].

2.3.1.2 Temporal validation

In principle, temporal validation approach is similar to internal validation using data-splitting. In this procedure, a single dataset is partitioned into two cohorts observed at different time points. The model is usually developed with data from one cohort of patients collected at a particular time point and is evaluated on a subsequent cohort from the same centre(s). Temporal validation is a prospective evaluation of a model, independent of the original data and the development process [7]. In addition, this approach can be considered as external validation with respect to time.

2.3.1.3 External validation

In this procedure, the performance of the model is evaluated on new data collected from a relevant patient population in a different centre. The second dataset, the validation set, must have information available on all the predictors in the model. The acceptable degree of similarities (or dissimilarities) between development and validation populations from which the samples were drawn is a matter of debate. However, it would

not be reasonable to expect that a model developed on a sample of older patients to perform well in younger patients. Of the three validation processes discussed above, only the external validation process appears to serve the purpose that a prognostic model should be transportable (or generalisable) to new patients [7, 14].

2.3.2 Key aspects of a model that need to be validated

This section discusses the key aspects of a model that need to be validated. These include (i) the agreement between the observed and predicted outcome of interest for a group of patients (calibration) or for an individual patient (accuracy score), and (ii) the ability of the model to distinguish high risk patients from those with low risk (discrimination) [7, 15, 22]. Another aspect that is used to assess the overall predictive performance of the model (both the calibration and discrimination simultaneously) is the concept of 'explained variation' [23, 35, 37]. The more the variability in the outcome explained, the better the predictive ability of the model. All these aspects are discussed in detail below.

2.3.2.1 Calibration

Calibration is an important aspect of a prognostic model that considers the answer to the question 'Are the predictions made by the model reliable?' More specifically, the calibration aspect of the model refers to the agreement between the predicted outcome of interest and the observed outcome. For example, for a group of 100 patients, if the probability that the event of interest will occur is predicted by the model to be 10%, then the model would be well calibrated if it actually does occurs for approximately in every 10 out of 100 patients. This suggests that the model has good predictive ability. When such agreement is quantified for an individual prediction by means of a loss function, for example, squared error loss, then it is called an 'accuracy score'.

2.3.2.2 Discrimination

Discrimination is the ability of the model to distinguish between patients with high risk for a given event and patients with low risk for that event. A model with reasonably good discriminatory ability shows a wide spread in the distribution of predicted

probabilities. For example, such a model might predict probabilities close to 100% for patients who had experienced the event of interest and probabilities close to 0% for patients who did not experience the event.

2.3.2.3 Distinction between calibration and discrimination

There is a conceptual difference between the discrimination and calibration aspects of a model. Good calibration does not necessarily lead to good discrimination in the model. In fact, an well calibrated model can exhibit poor even no discrimination. This phenomenon is illustrated by a hypothetical example of two models. Figure 2.1

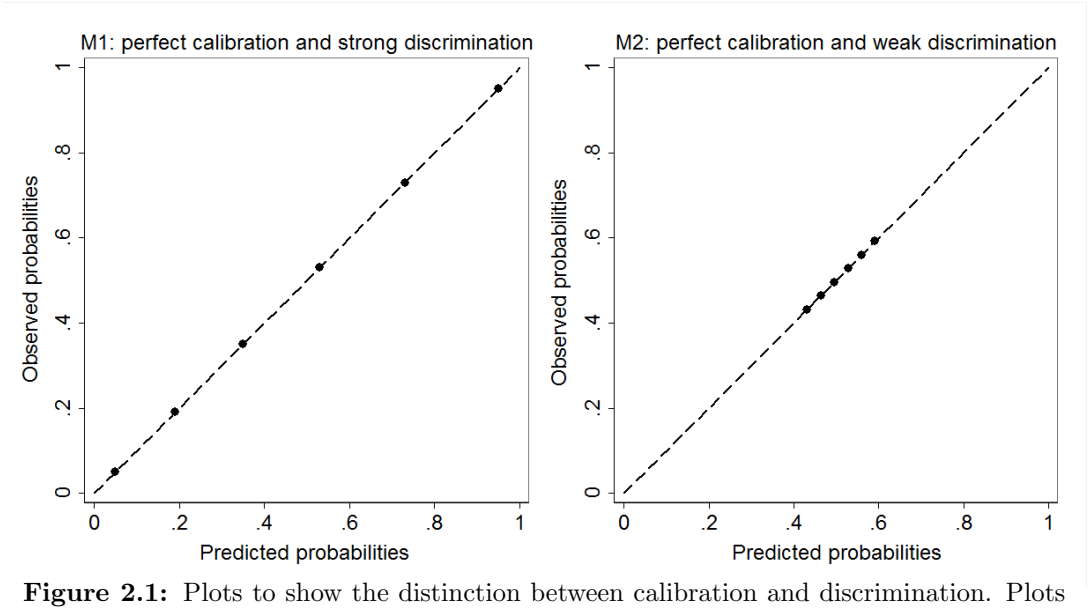

Figure 2.1: Plots to show the distinction between calibration and discrimination. Plots are for two hypothetical models (M1 and M2) with equal (perfect) calibration but different discriminatory abilities.

demonstrates the assessments of two well calibrated models, say M1 and M2, which have different discriminatory abilities. In case of both models, the predicted and observed probabilities for each of the six groups agree with each other, indicating perfect calibration. However, the predicted probabilities made by the model M1 ranges between 10% and 95% indicating strong discriminatory ability, whereas those of model M2 ranges between 40% and 55% indicating weak discriminatory ability. Generally for clinical purposes, one would like to have both good calibration and discrimination in a model. However, of these two aspects, the primary focus should be on good discrimi-

nation [8]. This is because if there is mis-calibration, re-calibrated is possible, but poor discrimination can not be fixed to a good discrimination.

2.3.2.4 Overall performance

This aspect of a model quantifies the accuracy of predictions for each patient or for a group of patients (calibration) and also quantifies the spread in predictions (discrimination). Therefore, by assessing this aspect one validates both the calibration and discriminatory ability of the model, often referred to as 'overall' predictive performance. This aspect is also known as 'explained variation'. A value of explained variation is interpreted as the proportion of variation in the outcome that can be explained by the predictors in the model. Intuitively, good calibration and strong discrimination implies a high value of explained variation [23, 38]. In the previous hypothetical example illustrated by Figure 2.1, the model M1 exhibits more explained variation than M2.

2.4 Measures that assess the predictive ability of a model

Measures that assess the predictive ability (calibration or discrimination or both) of a model can be considered as estimates of underlying parameters, which summarise the intrinsic ability of the model to predict accurately and to discriminate well between patients in the target population. Such measures are usually known as validation measures. This section briefly discusses some popular validation measures that have been proposed in the prognostic modelling literature.

2.4.1 Measures of calibration

The calibration of a model can be assessed graphically, with predictions made by the model on the x-axis and the observed outcome on the y-axis. This plot is known as calibration plot [39]. If the model's predictions agree with the observed outcomes over the entire range of predictions, the plot will show a 45-degree line. For an outcome with normal distribution, the calibration plot can be achieved from a scatter plot of observations against predictions. For a binary outcome y , the y-axis of the plot contains only 0 and 1 values. To estimate the observed proportions of the outcome for each patient in relation to the predicted probabilities, a smoothing technique, such as the

loess algorithm [40], can be applied [24]. The plot can also be obtained by grouping patients with similar predicted probabilities and then by comparing the mean observed outcome with the mean predicted probability obtained for each group of patients. For example, one can plot the observed outcome by decile of the predicted probabilities, which is essentially a graphical illustration of the Hosmer-Lemeshow test [41].

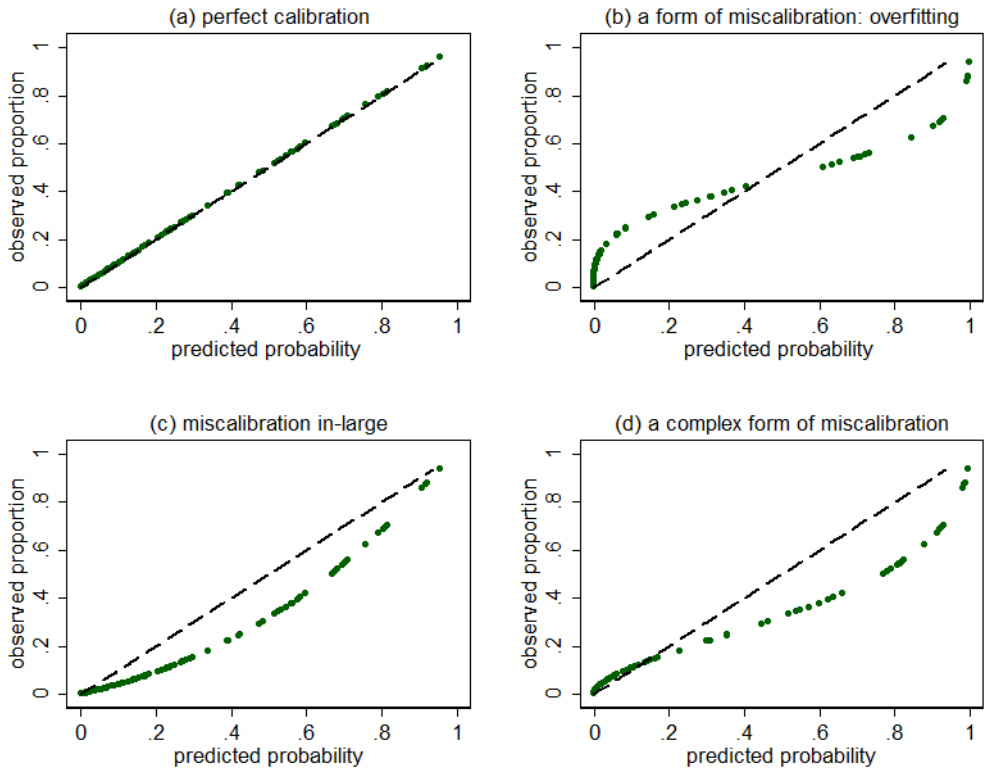

Figure 2.2: Theoretical calibration plots to assess the agreement between the observed proportion and predicted probability with a dot line through all outcome value (0 and 1): (a) $\hat{\alpha} = 0$, $\hat{\beta} = 1$; (b) $\hat{\alpha} = 0$, $\hat{\beta} = 0.74$; (c) $\hat{\alpha} = -0.65$, $\hat{\beta} = 1$; (d) $\hat{\alpha} = -0.65$, $\hat{\beta} = 0.74$.

The calibration plot can be summarised by fitting a (regression) line of the observed proportions on the predicted probabilities, which results in an intercept $\hat{\alpha}$ and slope $\hat{\beta}$. For a plot showing a 45-degree line, $\hat{\alpha} = 0$ and $\hat{\beta} = 1$ (Figure 2.2a). This approach to summarise the calibration aspect of a model was originally proposed by Cox [42] and was further considered by Miller et al. [39] for a model with binary outcomes. The intercept and slope of the calibration line can be estimated using a logistic regression model with the predicted 'prognostic index' derived from the model for the validation

sample as the only predictor. The 'prognostic index' is the linear combination of the predictors in the model weighted by the estimated regression coefficients.

If the estimated slope $\hat{\beta}$ is much smaller than 1, it indicates optimism (overfitting). This implies that the predictions are too extreme: the predictions are too low for low risk subjects and too high for high risk subjects (Figure 2.2b). If the opposite occurs, that is, the estimated slope is much larger than 1, it indicates that the predictions are too high for low risk subjects and too low for high risk subjects [38, 40]. The estimated intercept $\hat{\alpha}$ assesses the overall agreement between the observed and the predicted outcomes, that is, the agreement between the sum of all predicted probabilities and total number of observed outcomes. This is referred to as 'calibration-in-the-large'. If the intercept $\hat{\alpha}$ is much different from 0, it may indicate that the predicted probabilities are systematically too high ($\hat{\alpha} \ll 0$, Figure 2.2c) or too low ($\hat{\alpha} \gg 0$). If both the slope and intercept are far away from 1 and 0, respectively (Figure 2.2d), the interpretation of miscalibration is difficult, because the values of both slope and intercept are highly correlated [38].

For survival outcomes, a calibration plot could be a plot of Kaplan-Meier (K-M) estimates of survival probabilities at a selected time point, say t^*, against the survival probabilities predicted by the model at t^*. In a similar manner to that for binary outcomes, the plot can be achieved by grouping patients with similar predicted probabilities, which can be determined by the decile of predicted probabilities at t^*, and then by comparing the mean K-M probabilities with the mean predicted probabilities for each group of patients [38].

Furthermore, the slope of the calibration line can be estimated from a regression analysis using, for example, the Cox proportional hazards (PH) model with the 'prognostic index' as the only predictor [43, 44]. The intercept of the calibration model cannot be calculated directly using the Cox PH model, as it does not estimate an intercept, but it includes the intercept within the baseline hazard. If the intention is to validate the whole model, that is, both the slope and intercept, the time-axis needs to be transformed into the cumulative baseline hazard obtained from the model. Then a Weibull

model in an accelerated failure time (AFT) formulation of this transformed time scale with the 'prognostic index' as a single predictor can be used to assess the whole model [43, 44]. The resulting model takes the form: $\ln(t) = \alpha + \beta \times$ prognostic index $+ \gamma e$, where e is distributed as the logarithm of a negative exponential. The whole model is strictly well calibrated if $\hat{\alpha}$ is close to 0, $\hat{\beta}$ is close to -1, and $\hat{\gamma}$ is close to 1. Note that the Weibull model is also a proportional hazards model with regression coefficient $\beta^* = -\beta/\gamma$. For more details, see van Houwelingen and Thorogood [43] and van Houwelingen [44].

2.4.2 Measures of discrimination

A number of measures have been published in the literature to assess the discriminatory ability of prognostic models. The most commonly used measure is the concordance probability, which is also called the concordance statistic or C-index. Another measure is based on prognostic separation that quantifies the spread of the observed risks across the range of predicted risks. This measure is known as a measure of prognostic separation or a separation statistic.

2.4.2.1 Measures based on concordance probability

Concordance probability (C-index) quantifies the concordance between the ranking of the predicted and observed outcomes. For binary outcomes, concordance is measured between the predicted and observed event of interest. Whereas for survival outcomes, concordance is measured between the observed and predicted orders of failure. For binary outcomes, the concordance statistic (or C-index) is identical to the area under the receiver operating characteristic curve (AUC) [45]. The receiver operating characteristic (ROC) curve is the graph of sensitivity (true-positive rate) versus one minus specificity (true-negative rate) evaluated at consecutive threshold values of the predicted probability. Briefly, the sensitivity refers to the percentage of patients with an event who are correctly identified as having the condition, and the specificity refers to the percentage of patients without an event who are correctly identified as not having the condition.

The AUC measures the 'concordance' of ranking between the predicted probabilities of having the event for a random pair of subjects who had the event and who did not. This represents the probability that the event subject has higher predicted probability than the non-event subject. For a model with perfect discriminatory ability, the ROC curve passes through the coordinate (0,1) of the ROC space, which corresponds to sensitivity = 100% and specificity = 100% for each threshold value and AUC = 1. A straight line from the bottom left (0,0) to the top right (1,1) corners corresponds to the AUC = 0.5, which indicates a model with no discriminatory ability.

Applying ROC methodology to survival data is not straightforward. The AUC assesses the discriminatory ability of the model at an arbitrary time point rather than the entire time period, and it does not take into account the censoring pattern of the subjects. To overcome these drawbacks, an extension of the C-index or AUC proposed by Harrell et al. [8, 28, 46] for use with right censored survival data is commonly known as Harrell's C-index. This is a rank-order statistic motivated by Kendall's τ [47] that measures the association between the ranked predicted and observed survival times. Specifically the C-index is based on the idea that, for a randomly selected pair of subjects, the subject who fails first has shorter predicted survival time. This is described in detail in Chapter 3.

Gonen and Heller [48] discussed the possibility of bias in Harrell's C-index, induced by censoring, and proposed a new measure of concordance probability, $K(\beta)$, for use with the Cox proportional hazards model. $K(\beta)$ is a model based estimator and is a function of model parameters and the covariate distribution. This is explained in more details in Chapter 3.

2.4.2.2 Measure based on prognostic separation

This measure quantifies the separation between the observed risks across the range of predicted risks. In survival analysis, the standard approach often used in the literature is to generate a prognostic classification scheme comprising of two or more risk groups and to plot the Kaplan-Meier survival curves for each group, which leads to the idea of separation of survival curves as a measures of prognostic information. Based on this

idea, Royston and Sauerbrei [49] proposed a measure of prognostic separation, called separation statistic, D statistic. The D statistic can be calculated by first transforming the prognostic index derived from the model using the Blom's approximation [50] to give a standard normal order rank statistic z; D is then the coefficient of z in a model fitted with z as the only predictor. The D statistic can be interpreted as the log hazard ratio (for survival outcomes) or log odds ratio (for binary outcomes) between low-and high-risk patient-groups obtained by dichotomising the predicted prognostic index at their median value.

2.4.3 Overall performance measures: R^2 type measures

Measures in this category are equivalent to the R^2 measures generally used in normal linear regression and are also used to quantify the prognostic ability of the predictors in the model. The main reason for the popularity of R^2 in normal linear regression is its interpretation as the proportion of variation in the outcome that is explained by the predictors in the regression model. R^2 measures in prediction research aim at quantifying the increase in the amount of explained variation in the observed outcome resulting from the addition of the predictor to the model. The value of R^2 measures range between 0 and 1 (or 0 - 100%). A maximum value of 1 indicates that the predictors fully explain the variation in the outcome, whereas the minimum value 0 indicates that the predictors have failed to explain any of the outcome.

Extending the definition of R^2 for linear regression, several measures have been proposed for both binary and survival outcomes, exclusively for logistic and Cox models. Such measures have been reviewed and compared by Mittlbock and Schemper [51] for models with binary outcomes and by Choodari-Oskooei et al. [52] and Schemper and Stare [53] for models with survival outcomes. As discussed by Mittlbock and Schemper [51] and Choodari-Oskooei et al. [52], R^2 measures are mainly defined based on either a loss function (for example, squared error loss), or the model's log-likelihood function, or the Kullback-Leibler distance [54].

Commonly used R^2 type measures based on the loss function approach include those proposed by Schemper and Henderson [23], Graf et al. [55], Schemper [56], Margolin

and Light [57], Haberman [58], van Houwelingen and Le Cessie [59], and Schemper [60]. Measures based on the model's likelihood and the Kullback-Leibler distance include those proposed by Kent and O'Quigley [61], Cox and Snell [62], Korn and Simon [63], Magee [64], Nagelkerke [65] O'Quigley et al. [66], and Royston [67]. Among all types of R^2 measures, the measures based on the loss function approach have an interpretation closely related to the measures of discrimination and calibration, and have also been used in practice [37, 68]. In the following subsection, a brief discussion on the measures based on the loss function approach are given.

2.4.3.1 Measures of explained variation: based on loss function approach

Measures in this category quantify the relative gain in predictive accuracy resulting from the addition of predictors to the model. The predictive accuracy is usually obtained by quantifying the distance between the observed and predicted outcome. For continuous outcomes, the distance is usually $Y - \hat{Y}$, where $\hat{Y}$ is the predicted value of the outcome Y. For binary outcomes, $\hat{Y}$ is equal to the predicted probability of the event occurring, and for survival outcomes, it is the predicted survival probability at a given time or as a function of time. A measure of predictive accuracy is then defined by applying a loss function to the distance $Y - \hat{Y}$. The most commonly used loss functions include the squared error loss, for example, $(Y - \hat{Y})^2$, and absolute error loss, for example, $|Y - \hat{Y}|$). A wide variety of loss functions, most of which are adapted from these two, for binary and survival outcomes have been discussed in the paper of Mittlbock and Schemper [51] and Korn and Simon [63], respectively.

The most commonly used measure of predictive accuracy for both binary and survival outcomes is the Brier score [55, 69], which was originally developed by Brier [70] for assessing the inaccuracy of probabilistic weather forecasts. The Brier score is based on the squared error loss function and assesses the predictive accuracy of individual predictions. Schemper and Henderson [23] proposed another measure of predictive accuracy, D_x, based on the absolute error loss function. The calculation of both the Brier score and D_x for survival data require an additional weight factors to adjust for the effects of censoring.

The relative gain in predictive accuracy, which gives an R^2 value, can be obtained by comparing the prediction error PE_x (for example, the Brier score) obtained for the model with predictors X and the prediction error PE_0 obtained for the null model. Then a general measure of explained variation based on the loss function approach is defined as

$$R^2_{PA} = 1 - \frac{PE_x}{PE_0}.$$

R^2_{PA} ranges between 0 and 1; a maximum value of 1 indicates that the outcome is fully explained by the predictors in the model while the minimum value 0 indicates that the predictors have failed to explain the outcome at all.

2.5 Conclusion

This chapter has discussed motivation and a general procedure for validating a prognostic model, particularly focusing on models with binary and survival outcomes. Validation of a prognostic model is essential before using it in clinical practice. Generally, validating a prognostic model implies achieving evidence regarding the accuracy of predictions for new patients different from those used to developed the model. This chapter has discussed the design of the validation process and key aspects of the model that are evaluated when conducting a validation study. This chapter has also provided a brief literature review of validation measures that have been proposed to assess the predictive performance of models for binary or survival outcomes.

The next chapter reviews and evaluates some of the validation measures that have been proposed for models with independent survival outcomes and makes practical recommendations, starting with a motivation for this investigation.

Chapter 3

Measures for independent survival outcomes

3.1 Introduction

It is essential that prognostic models have good ability to make accurate predictions. Therefore, there needs to be validation measures available to evaluate the predictive ability of these models. Validation measures for models for binary outcomes are reasonably well developed; see, for example, Omar et al. [10], Steyerberg et al. [24], and Royston and Altman [25]. However, despite the proposal of several validation measures for survival outcomes, it is still unclear which measures should be adopted for general use. One common feature of survival data is that these are subject to censoring, and therefore it is essential for a validation measure to be robust to the degree of censoring [52, 53, 68]. The aim of this chapter is to review some of the validation measures proposed for survival models, to evaluate their performance through simulation studies, and to make recommendations regarding their use in practice.

Some authors have already investigated the performance of validation measures for survival risk models [52, 53, 68]. However, these papers focussed only on measures of explained variation. In addition, some of these papers did not consider the use of validation data; they validated the model using the same data that were used to

develop the model [52, 53]. However, assessing the performance of a prognostic model on the data used to develop the model can lead to overoptimistic results regarding its predictive performance [26]. This chapter will evaluate validation measures using data that have not been used for model development. The performance of validation measures for survival outcomes from all categories: discrimination, calibration and predictive accuracy, and explained variation will be investigated.

The chapter is organised as follows. Section 2 describes two real clinical datasets that are used to simulate the new data for this investigation. Section 3 describes the validation measures that were assessed, focusing on the motivation for choosing these measures, their estimation, and their properties. In Section 4, the criteria against which the measures are assessed and the simulation design are discussed. Section 5 presents and discusses the simulation results. Some recommendations are discussed in Section 6, and Section 7 ends the chapter with a general discussion.

3.2 Example data sets

3.2.1 Breast cancer data

This dataset contains information on patients with primary node positive breast cancer from the German Breast Cancer Study [71]. The outcome of interest is recurrence-free survival time and there are 686 patients, with 299 events; that is, the rest of the patients (56%) were censored. The median follow-up time was 4.5 years. All these patients had complete data for all predictors that include age, tumour size (tsize), number of positive lymph nodes (lnod), progesterone status (progest), menopausal status (menpst: pre/post), tumour grade (tgrad: 1-3), and hormone therapy (hormon: yes/no). For simulation purposes, all the continuous predictors, except age, were log-transformed. If the predictor contained zero values then a small scalar was added prior to the transformation. Age was converted into three categories: below 45 years, 45-60 years, above 60 years. The risk model based on this dataset has already been published [72]. The only focus here is to use this dataset for simulation purposes and to assess the performance of the validation measures, rather than on the clinical motivation for developing a risk model.

3.2.2 Sudden cardiac death data

This dataset contains information on a retrospective cohort of patients with hypertrophic cardiomyopathy from a single cardiac hospital in the UK. The outcome of interest is sudden cardiac death (SCD). There are 1831 patients of which 79 had recorded sudden cardiac death; the rest of the patients were censored. The median follow-up time was approximately 5 years. The predictors of interest are age, number of runs of ventricular tachycardia (runvent: 0-2 or 3+), obstruction to blood flow (BF), abnormal blood pressure response to exercise (BP: normal or abnormal), and maximum thickness of heart muscle (HM). The dataset is used in this thesis only for simulation purposes and to evaluate the performance of the validation measures. The risk model based on this dataset is under development; the aim is to guide clinical management of patients who have been suffering from hypertrophic cardiomyopathy. I would like to thank Drs Constantinos O'Mahony and Perry Elliott for allowing me to use their data for simulation purposes.

3.3 Validation measures for the Cox Proportional Hazards model

The Cox Proportional Hazards (PH) model [73] is the most commonly used regression model for the analysis of right censored survival outcomes. Note that a subject is right censored if it is known that the event of interest occurs some time after the observed follow up period. Consequently, in health care research, prognostic models for survival data are typically developed using the Cox PH model and hence validation measures will be evaluated based on this model. For this investigation, we have selected measures that can be interpreted and communicated easily for clinical purposes, have been implemented or are easy to implement in commonly used statistical softwares, and can be used routinely in practice.

Validation measures selected include the calibration slope [44] from the category of calibration measures; Harrell's C-index [8], Gönen and Heller's $K(\beta)$ [48], and Royston and Sauerbrei's D [49] from the discrimination measures; Graf et al's integrated

Brier score (IBS) from the category of predictive accuracy measures; and R^2_{IBS} [55], and Schemper and Henderson's V [23] from the explained variation category. Further motivation for choosing these measures and their estimation for the Cox PH model are given in the following sections, following a description of the Cox model and some basic notation.

3.3.1 The Cox Proportional Hazards model

Suppose we have data on N subjects, where for the ith subject, t_i is the observed time, δ_i is 1 if the event of interest is experienced at t_i or 0 otherwise (right censoring), and $\boldsymbol{x}_i$ is a vector of p predictor values. The Cox model specifies the hazard, which corresponds to the risk that the event will occur in an interval after time t given that the subject had survived to time t, as

$$h(t|\boldsymbol{x}_i, \boldsymbol{\beta}) = h_0(t) \exp(\eta_i),$$

where $h_0(t)$ is the baseline hazard that describes how the hazard changes over time at baseline levels of predictors, and $\eta_i = \beta_1 x_{i1} + \ldots + \beta_p x_{ip} = \boldsymbol{\beta}^T \boldsymbol{x}_i$ is the prognostic index, a linear combination of p predictor values weighted by regression coefficients $\beta_1, \ldots, \beta_p$. The predictive form of this model can be written in terms of the survival function as

$$S(t|\boldsymbol{x}_i, \boldsymbol{\beta}) = S_0(t)^{\exp(\eta_i)},$$

where $S(t|\boldsymbol{x}_i)$ is the probability of surviving beyond time t given predictors $\boldsymbol{x}_i$, and $S_0(t)$ is the baseline survivor function at time t, which corresponds to the baseline hazard $h_0(t)$ as $S_0(t) = \exp[-\int_0^t h_0(u)du]$. To make predictions at time t, one uses estimates $\hat{\boldsymbol{\beta}}^T$ and $\hat{S}_0(t)$ [37].

3.3.2 Measures of calibration

A calibration measure assesses whether the model makes reliable predictions by assessing how closely the predicted probability of survival for a group of subjects at a particular time point agrees with the actual outcome. When such an agreement is quantified for an individual subject, this aspect leads to a measure of predictive accuracy

(Section 3.3.4). The most commonly used calibration measure for survival models is the calibration slope proposed by van Houwelingen [44], which was originally introduced for binary outcomes by Cox [42] and was further considered by Miller et al. [74].

3.3.2.1 Calibration slope (CS)

The calibration slope (CS) assesses the degree of agreement between the observed and predicted values using a regression model. The calibration slope for survival data is obtained by fitting a Cox Model to the validation data where the predicted prognostic index $\hat{\eta}_i = \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ is included as the only predictor:

$$h(t|\hat{\eta}, \beta_\eta) = h_0(t) \exp(\beta_\eta \hat{\eta}).$$

If $\hat{\beta}_\eta$ is close to 1, it suggests that the predicted log hazard ratio is accurate. A value far away from 1 indicates that some form of re-calibration of the risk model may be necessary [44]. In particular, $\hat{\beta}_\eta \ll 1$ suggests over-fitting in the original data with the spread of predictions being too large: the predictions are too low for low risk subjects and too high for high risk subjects.

3.3.3 Measures of discrimination

Measures of discrimination assess how well a model can distinguish patients with high-risk from those with low-risk. The discriminatory ability of a survival model is commonly quantified by a measure of concordance probability that quantifies the correlation between the predicted and observed survival times. The most frequently used concordance measure is the C-index, which has been proposed by Harrell et al. [8]. However, Gonen and Heller [48] reported possible censoring bias in the C-index and proposed a new measure of concordance probability $K(\beta)$ to overcome this problem. Another measure of discrimination is the D statistic proposed by Royston and Sauerbrei [49]. This is based on the idea of prognostic separation which quantifies the spread in the observed risks between those patients predicted to be at low risk and those at high risk. All these measures are discussed in detail in the next section.

3.3.3.1 Harrell's C-index

The C-index [8] is a rank-correlation measure motivated by Kendall's τ statistic [47] which quantifies the correlation between the ranked predicted and observed survival times. For the Cox PH model, this is defined as the probability that of a randomly selected pair of subjects, the subject who fails first has the worse predicted prognosis. The overall concordance probability, or the C-index, is calculated as the proportion of all usable pairs in which the predictions and outcomes are concordant. For a randomly selected pair of subjects (i, j) with observed survival times t_i and t_j respectively, the pair is said to be usable or comparable if $t_i \neq t_j$. For censored data, a pair is usable if the shorter time corresponds to an event. With the corresponding predicted survival times $\hat{t}_i$ and $\hat{t}_j$, a usable pair is said to be concordant if $t_i > t_j$ and $\hat{t}_i > \hat{t}_j$ or $t_i < t_j$ and $\hat{t}_i < \hat{t}_j$. For a proportional hazards model, a 'one-to-one' transformation holds between the predicted survival time $\hat{t}_i$ and the predicted probability of survival $S(t|\boldsymbol{x}_i)$ for every $t > 0$ [75]. Therefore, $\hat{t}_i$ and $S(t|\boldsymbol{x}_i)$ are interchangeable. A pair is then concordant if $t_i > t_j$ and $S(t|\boldsymbol{x}_i) > S(t|\boldsymbol{x}_j)$ or $t_i < t_j$ and $S(t|\boldsymbol{x}_i) < S(t|\boldsymbol{x}_j)$. If the inequalities go in the opposite direction, that is, $t_i > t_j$ and $S(t|\boldsymbol{x}_i) < S(t|\boldsymbol{x}_j)$ or $t_i < t_j$ and $S(t|\boldsymbol{x}_i) > S(t|\boldsymbol{x}_j)$, then the pair is said to be discordant. In the presence of censoring, not all pairs of subjects are observed to be usable. If there is high degree of censoring then many subject pairs will be omitted from the calculation of the C-index.

Mathematically, the concordance probability under the Cox PH model can be defined as

$$C = \Pr\Big[S(t_i|\boldsymbol{x}_i) < S(t_j|\boldsymbol{x}_j)|t_i < t_j\Big],$$

or equivalently

$$C = \Pr\Big[\boldsymbol{\beta}^T\boldsymbol{x}_i > \boldsymbol{\beta}^T\boldsymbol{x}_j|t_i < t_j\Big].$$

Ties in the observed survival times and/or in the predicted survival probability are ignored in above definition. Indeed, the distributions of survival times and the predicted probability of survival are assumed to be continuous.

Considering all possible pair of subjects (i, j), given that at least one of them had an event, with their observed data $\{(t_i, \delta_i, \boldsymbol{x}_i), (t_j, \delta_j, \boldsymbol{x}_j)\}$ the C-index for the Cox PH model can be estimated using

$$\hat{C} = \frac{\sum_{i<j}^{N}\sum^{N}\Big[I(\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_i > \hat{\boldsymbol{\beta}}^T\boldsymbol{x}_j \;\&\; t_i < t_j \;\&\; \delta_i = 1) + I(\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_j > \hat{\boldsymbol{\beta}}^T\boldsymbol{x}_i \;\&\; t_j < t_i \;\&\; \delta_j = 1)\Big]}{\sum_{i<j}^{N}\sum^{N}\Big[I(t_i < t_j \;\&\; \delta_i = 1) + I(t_j < t_i \;\&\; \delta_j = 1)\Big]},$$

where $I(.)$ is the indicator function and $\hat{\boldsymbol{\beta}}^T$ is the partial likelihood estimator of $\boldsymbol{\beta}^T$. The C-index typically ranges between 0.5 and 1, where a value of 0.5 indicates no discriminatory ability of the model and 1 indicates perfect discrimination. Values below 0.5 are possible, but rarely occur in practice. This scenario implies that the model predicts better prognosis for the subject who fails first.

3.3.3.2 Gönen and Heller's $K(\beta)$

Gönen and Heller's $K(\boldsymbol{\beta})$ [48] is an alternative estimator of the concordance probability under the Cox PH model. It is a function of the model parameters and the predictor distribution only. Unlike Harrell's C-index, $K(\boldsymbol{\beta})$ does not use the observed event and censoring times directly. Since the effect of censoring on the partial likelihood estimator of $\boldsymbol{\beta}^T$ is negligible, $K(\boldsymbol{\beta})$ is reported to be asymptotically unbiased [48]. The $K(\boldsymbol{\beta})$ statistic has the same interpretation to the C-index.

Under the proportional hazards model, the ranking between the survival times, denoted $T(\boldsymbol{\beta}^T\boldsymbol{x}_i)$ and $T(\boldsymbol{\beta}^T\boldsymbol{x}_j)$, of a randomly selected pair of subjects (i, j) can be calculated by

$$\begin{aligned}\Pr[T(\boldsymbol{\beta}^T\boldsymbol{x}_j) > T(\boldsymbol{\beta}^T\boldsymbol{x}_i)] &= \int_0^\infty S(t|\boldsymbol{x}_j, \boldsymbol{\beta}^T)dS(t|\boldsymbol{x}_i, \boldsymbol{\beta}^T) \\ &= \frac{1}{1 + \exp\{\boldsymbol{\beta}^T(\boldsymbol{x}_j - \boldsymbol{x}_i)\}},\end{aligned}$$

where $T(\boldsymbol{\beta}^T\boldsymbol{x}_i)$ and $T(\boldsymbol{\beta}^T\boldsymbol{x}_j)$ corresponds to the prognostic index $\boldsymbol{\beta}^T\boldsymbol{x}_i$ and $\boldsymbol{\beta}^T\boldsymbol{x}_j$, respectively. Considering all pairs (i, j), the concordance probability

$$K(\boldsymbol{\beta}) = \Pr(t_j > t_i | \boldsymbol{\beta}^T \boldsymbol{x}_i \geq \boldsymbol{\beta}^T \boldsymbol{x}_j)$$

can be estimated using

$$K(\hat{\boldsymbol{\beta}}) = \frac{2}{N(N-1)} \sum_{i<j}^{N} \sum^{N} \Big[\frac{I(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i > \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_j)}{1 + \exp\{\hat{\boldsymbol{\beta}}^T (\boldsymbol{x}_j - \boldsymbol{x}_i)\}} + \frac{I(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_j > \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i)}{1 + \exp\{\hat{\boldsymbol{\beta}}^T (\boldsymbol{x}_i - \boldsymbol{x}_j)\}} \Big].$$

It seems that $K(\boldsymbol{\beta})$ can be calculated knowing the regression coefficients of the Cox model and predictor values. Unlike C-index, all pairs of subjects are used in the calculation of $K(\boldsymbol{\beta})$. The C-index uses all pairs only when there is no censoring. Therefore, for the uncensored data, $K(\boldsymbol{\beta})$ claims to be very close to the C-index [48].

3.3.3.3 Royston and Sauerbrei's D

The D-statistic [49] quantifies the observed separation between subjects with low and high predicted risk, as predicted by the model. It is calculated by first transforming each patient's predicted prognostic index $\hat{\eta}_i = \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ to give standard normal order rank statistics (rankits-formed) using Blom's approximation [50]. These rank statistics are then divided by a factor of $\sqrt{(\frac{8}{\pi})}$ to give z_i as

$$z_i = k^{-1} \Phi^{-1} \Big(\frac{i - 3/8}{N + 1/4} \Big),$$

where i is the rank order based on the predicted prognostic index $\hat{\eta}_{(i)}$, N is the number of observations, $\Phi^{-1}(.)$ is the inverse standard Normal distribution function, and $k = \sqrt{8/\pi} \approx 1.60$. The scaled normalised predicted prognostic index z_i is distributed as $\mathrm{N}(0, \pi/8)$. A Cox PH model is then fitted using z_i as the sole predictor, which takes the following form:

$$h(t|z, \beta_z) = h_0(t) \exp(\beta_z z).$$

The estimated regression coefficient for this predictor is the estimate of D statistic, $\hat{D}$. Alternatively, suppose two equal-sized prognostic groups are determined by dichotomising z_i at their median, then Cox regression on the group averaged z_i provides the same regression coefficient as Cox regression on a dummy variable distinguishing the groups.

Therefore, $\hat{D}$ is interpreted as the log hazard ratio between the two patient groups; these groups may be described as low and high risk, respectively. The null value for $\hat{D}$ is 0, with increasing values indicating greater separation.

3.3.4 Measures of predictive accuracy and explained variation

Measures of predictive accuracy quantify the squared (or absolute) distance between the predicted survival probability and the actual outcome for an individual subject at a particular time point. For example, the Brier score [55, 76] may be used to assess predictive accuracy at a particular time point t. The integrated version of the Brier score (integrated Brier score) assesses the overall predictive accuracy over the entire study period. Schemper and Henderson [23] proposed a similar measure of predictive accuracy, denoted by D_x, to the integrated Brier score (IBS).

A measure of predictive accuracy leads to a 'relative measure of predictive accuracy' or 'measure of explained variation' that has the same interpretation to R^2 measures commonly used in normal linear regression [23, 55]. Several measures of explained variation have been proposed in the literature [52, 53]. Among them, the measures which are based on the predictive accuracy approach, for example, V proposed by Schemper and Henderson [23] and R^2_{IBS} proposed by Graf et al. [55], have an interpretation closely related to the measures of discrimination and calibration. In addition, these measures have been used in practice [37, 68]. Therefore, these two measures have been chosen from the category of explained variation. All these measures are discussed in detail in the following sections.

3.3.4.1 Graf et al's Brier score

The Brier score can be calculated by comparing the predicted survival probability and the observed survival status using a quadratic loss function, then taking a weighted average over all subjects. The weights are used to compensate for the loss of information due to censoring.

For data $(t_i, \delta_i, \boldsymbol{x}_i)$, the individual contribution to the Brier score (BS) at time t can be split up into three categories: if

(i) $t_i \leq t$ and $\delta_i = 1$, $\widehat{BS}(t|\boldsymbol{x}_i) = (0 - \hat{S}(t|\boldsymbol{x}_i))^2$

(ii) $t_i > t$ and $\delta_i = 1$ or $\delta_i = 0$, $\widehat{BS}(t|\boldsymbol{x}_i) = (1 - \hat{S}(t|\boldsymbol{x}_i))^2$

(iii) $t_i \leq t$ and $\delta_i = 0$, the survival status is unknown and thus the contribution to BS cannot be calculated.

The loss of information indicated in category (iii) is compensated by adding a weight to $\widehat{BS}(t|\boldsymbol{x}_i)$ for subjects in categories (i)-(ii). These weights account for the inverse probability of censoring [77]. The resulting weighted Brier score can be calculated as

$$\widehat{BS}_x(t) = \frac{1}{N}\sum_{i=1}^{N}\Big[\frac{(0 - \hat{S}(t|\boldsymbol{x}_i))^2 I(t_i \leq t, \delta_i = 1)}{\hat{G}(t_i)} + \frac{(1 - \hat{S}(t|\boldsymbol{x}_i))^2 I(t_i > t)}{\hat{G}(t)}\Big],$$

where $\hat{G}(t)$ is the Kaplan-Meier estimate of the probability of being uncensored at time t. The Brier score at time t can be interpreted as the mean squared error of prediction for survival. Lower values of the Brier score indicate better predictive performance of the model; 0 indicates perfect predictions, which is however very unlikely to occur in practice.

The Brier score defined above is a function of time t. Therefore, to obtain a summary measure of predictive accuracy over a range of time points, say $0 < t \leq \tau$, an integrated version of the Brier score (IBS) can be estimated by integrating $\widehat{BS}_x(t)$ for all t $(0 < t \leq \tau)$ with respect to some weight functions $W(t)$. The IBS is given by

$$\widehat{IBS}_x(\tau) = \int_0^{\tau} \widehat{BS}_x(t) d\hat{W}(t).$$

where $\hat{W}(t)$ is a function to weight the contribution of the Brier score at individual time points, and τ should be chosen as any time less than or equal to the last observed failure time. The weight $\hat{W}(t)$ is implemented for $\widehat{IBS}_x(\tau)$ as a straightforward trapezoidal rule for integrating the area under the prediction curve. Following [55], we choose $\hat{W}(t) = (1 - \hat{S}(t))/(1 - \hat{S}(\tau))$, where $\hat{S}(t)$ denotes the estimated marginal survival function. The integrated Brier score was investigated in this study.

3.3.4.2 Graf et al's R^2_{IBS}

To quantify the relative gain in predictive accuracy resulting from the inclusion of predictors in the model, Graf et al. [55] also proposed R^2_{IBS} which can be estimated as

$$\hat{R}^2_{IBS} = 1 - \widehat{IBS}_x(\tau)/\widehat{IBS}_0(\tau),$$

where $\widehat{IBS}_0(\tau)$ and $\widehat{IBS}_x(\tau)$ are the estimated integrated Brier scores obtained from the null model and the model with predictors, respectively. R^2_{IBS} ranges between 0 and 1. A maximum value of 1 indicates that the predictors fully explain the variation in the outcome, whereas the minimum value 0 indicates that the predictors have failed to explain any of the outcome.

3.3.4.3 Schemper and Henderson's V

Similar to R^2_{IBS}, V [23] is a relative measure of prognostic accuracy and can be calculated as

$$\hat{V} = 1 - \hat{D}_x(\tau)/\hat{D}_0(\tau),$$

where $\hat{D}_0(\tau)$ and $\hat{D}_x(\tau)$ are the measures of predictive accuracy obtained for the null model and the model including the predictors, respectively. In principle, $D_x(\tau)$ is analogous to $IBS_x(\tau)$. However, unlike $IBS_x(\tau)$ which is based on quadratic differences, $D_x(\tau)$ quantifies the absolute difference between the predicted and observed survival status (alive/died) at each event time and averages over all subjects and event times up to the last event time τ. Additionally, subjects who are censored before the event time are allocated to alive or dead categories according to their corresponding conditional survival probability estimates at their censoring times.

Assuming that there are m distinct event times $t_{(j)}$ $(t_{(1)} < t_{(2)} \ldots < t_{(m)})$ in the observed data $(t_i, \delta_i, \boldsymbol{x}_i)$ with d_j events at $t_{(j)}$, the individual contribution to the absolute distance, $M(t_{(j)}|\boldsymbol{x}_i)$, at each event time $t_{(j)}$ falls into one of three categories:

(i) $t_i \leq t_{(j)}$ and $\delta_i = 1$, $\hat{M}(t_{(j)}|\boldsymbol{x}_i) = \hat{S}(t_{(j)}|\boldsymbol{x}_i)$

(ii) $t_i > t_{(j)}$ and $\delta_i = 1$ or $\delta_i = 0$, $\hat{M}(t_{(j)}|\boldsymbol{x}_i) = 1 - \hat{S}(t_{(j)}|\boldsymbol{x}_i)$

(iii) $t_i \leq t_{(j)}$ and $\delta_i = 0$, $\hat{M}(t_{(j)}|\boldsymbol{x}_i) = \Big(1 - \hat{S}(t_{(j)}|\boldsymbol{x}_i)\Big)\frac{\hat{S}(t_{(j)}|\boldsymbol{x}_i)}{\hat{S}(t_i|\boldsymbol{x}_i)} + \hat{S}(t_{(j)}|\boldsymbol{x}_i)\Big(1 - \frac{\hat{S}(t_{(j)}|\boldsymbol{x}_i)}{\hat{S}(t_i|\boldsymbol{x}_i)}\Big)$.

The first category corresponds to the subjects who have died before or at $t_{(j)}$ and the second to those who are alive at $t_{(j)}$. The last category relates to the subjects censored before or at $t_{(j)}$ and amounts to an extrapolation, assuming that these subjects have identical risk of death to those with known survival status at $t_{(j)}$. This assumption is quite similar to that of random censoring and is required by the standard survival methods [23]. Therefore, the estimate in the third category gives an average over alive or dead categories weighted by the corresponding conditional probability estimates at their censoring times: namely, $\hat{S}(t_{(j)}|\boldsymbol{x}_i)/\hat{S}(t_i|\boldsymbol{x}_i)$ represents the probability of survival beyond time $t_{(j)}$ given that the subject survived to at least time t_i.

The overall estimator $\hat{D}_x(\tau)$ of predictive accuracy can be obtained by taking a weighted average of $M(t_{(j)}|\boldsymbol{x}_i)$ over failure times, with weights designed to compensate for the reduction in observed deaths due to earlier censoring:

$$\hat{D}_x(\tau) \;=\; w^{-1}\sum_{j=1}^{m}\hat{G}(t_{(j)})^{-1}d_j\Big[\frac{1}{N}\sum_{i=1}^{N}\hat{M}(t_{(j)}|\boldsymbol{x}_i)\Big],$$

where $w = \sum_{j=1}^{m}\hat{G}(t_{(j)})^{-1}d_j$ is the weighting factor and $\hat{G}(t_{(j)})$ is the Kaplan-Meier estimate of the censoring times. Similarly, $\hat{D}_0(\tau)$ can be computed for the null model by replacing $\hat{S}(t_{(j)}|\boldsymbol{x}_i)$ with $\hat{S}(t_{(j)})$.

Hielscher et al. [68] showed that $IBS_x(\tau) = D_x(\tau)/2$, given that the model is correctly specified and the same method of integrating over time is used. Similarly $IBS_0(\tau) = D_0(\tau)/2$. Furthermore, since $D_x(\tau)$ uses absolute distance between the observed and predicted survival status as opposed to squared distance used by $IBS_x(\tau)$, $D_x(\tau)$ is less affected than $IBS_x(\tau)$ by unstable survival probability estimate for the largest survival time. Hence, $D_x(\tau)$ has a smaller variance than $IBS_x(\tau)$. A similar

argument can be applied to $IBS_0(\tau)$ and $D_0(\tau)$ and hence to V and R^2_{IBS}. As discussed by Hielscher et al. [68], the difference between squared distance and absolute distance is particularly large in the context of survival data, because due to censoring the uncertainty in the right tail of the survival distribution is large. This uncertainty may have influence on the quantity we use to evaluate the prediction accuracy. Therefore, a measure of predictive accuracy that is based on absolute distance might be preferred in this context.

3.4 Evaluation of the measures

Using a simulation study the validation measures are evaluated against a set of criteria that a suitable validation measure should have in the context of survival analysis. This section discusses the criteria, the simulation design, and strategies for assessing the measures against the proposed criteria.

3.4.1 Criteria for evaluation

To evaluate the suitability of the validation measures for use in practice with survival data, three aspects were considered:

(i) *Robustness to censoring*: Censoring is common for survival data. For example, in the example datasets in Section 3.2, there are 56% censoring in the breast cancer data while it is 95% in the sudden death data. An essential property for a validation measure is that it should be robust to censoring or at least not affected much by the presence of censoring.

(ii) *Sensitivity to the exclusion of important predictors*: If an important predictor is excluded from the model then the validation measures, except perhaps for the calibration slope, should demonstrate sensitivity to the exclusion. The validation measures are generally expected to move closer to their null value as important predictors are omitted from the model. However, the calibration slope may not react to this exclusion if the distribution of the predictors in the development and validation data are similar [40, 44]. For more details see Section 3.5.1.2.

(iii) *Interpretability*: The measure should be intuitive and clearly interpretable.

Each validation measure under study has already been discussed with respect to criteria (iii) in Section 3.3. In the simulation study, the measures are investigated with respect to criteria (i) and (ii).

3.4.2 Simulation design

3.4.2.1 Simulation scenarios

The simulation study was based on the two clinical datasets described in Section 2. Validation datasets were generated by simulating new outcomes for each of these datasets based on a true model and combining these with the original predictors. The validation measures were investigated over a range of scenarios to mimic real situations. For all simulations, three different risk profiles (low, medium, and high) were constructed for the patients in the validation data to reflect the fact that, in practice, the characteristics of the patients in the development and validation data may differ.

For the investigation into the effect of censoring, two types of censoring mechanism were considered, random and administrative. Random censoring is more common in clinical studies where patients are lost to follow-up throughout the course of the study, and administrative censoring is more common in population-based studies where birth cohorts are followed up until a fixed time point. The levels of censoring considered were 0%, 20%, 50%, and 80%, which combined with the risk profiles, results in a total of 24 validation scenarios for each clinical dataset. No development data were simulated for this investigation. It was assumed that the risk model had been correctly specified and perfectly estimated in order to assess the effect of censoring, rather than model development.

Censoring was not introduced into the simulations that investigated the effect of the omission of predictors, so as not to confound the results. Again, no development data were simulated for this investigation although incorrectly specified risk models were considered.

3.4.2.2 Generating new survival and censoring times

To simulate validation data, new survival outcomes were generated from a true model based on each of the real data sets. The true model was derived by fitting a Weibull proportional hazards model to each dataset including all available predictors. The estimates of the model parameters from these fitted models were then set as the "true" values to simulate new outcomes using the Weibull distribution. The Weibull survival times were simulated from the true model as

$$t_i = \left(\frac{-\log(u_i)}{\exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)} \right)^{1/\gamma} \quad (i = 1, \ldots, N),$$

where $\boldsymbol{\beta}^T \boldsymbol{x}_i$ is the true prognostic index (PI) with observed predictor vector $\boldsymbol{x}$, γ is the true value of the shape parameter, and u_i has a pseudo-random uniform distribution on (0, 1).

To introduce random censoring, an additional Weibull distributed censoring time was simulated with the same shape parameter as before but with the log hazard ratio $\boldsymbol{\beta}^T \boldsymbol{x}_i$ replaced by a scalar λ. Different choices of λ were used to give different proportions of censoring. To generate administratively censored data, it was assumed that individuals were recruited uniformly over the period from 0 to T^* and were censored at the study end date T^*, which was fixed in advance. The censoring times were simulated from a uniform distribution on (0, T^*) with different choices of T^* giving different proportions of censoring. The observed times under both types of censoring mechanism were obtained by taking the minimum of the survival and censoring times.

3.4.2.3 Generating validation data with different risk profiles

For each of the example datasets, validation data with three different risk profiles were created. To create these validation datasets, patients were split into three tertile groups based on their prognostic index PI=$\boldsymbol{\beta}^T \boldsymbol{x}_i$ derived from the true model. These groups may be viewed as low, medium, and high risk patients. Based on these risk groups, three different validation datasets were created by sampling patients (without replacement) in the following way:

(a) low risk profile: 80% of the patients were sampled from the lowest tertile, 50% from the middle tertile, and 20% from the highest tertile;

(b) medium risk profile: all patients from the 3 risk groups were used, which formed a validation sample with a mix of high and low risk patients. By definition, this dataset has the same risk profile as the observed (development) data;

(c) high risk profile: 20% of the patients were sampled from the lowest tertile, 50% from the middle tertile, and 80% from the highest tertile.

The whole procedure was done once before simulating the outcome data. Due to the sampling scheme considered, the sample sizes for the low and high risk profiles were half that of the medium risk profile. To achieve equal sample sizes we doubled the size of the low and high risk profile datasets by creating two "patients" based on each set of the observed predictor values. However, the survival and censoring times were not duplicated and were generated separately for each "patient".

Table 3.1 summarises the risk profile of the patients in the above validation scenarios in terms of the failure probability $(1 - S(t^*|x))$ estimated at a single time-point t^*. In the breast cancer simulations, the overall risk of failure was 27%, 34%, and 42% at 3 years for the patients from the low, medium, and high risk profile, respectively. The difference between the 1st and 3rd tertiles of the failure probability $(1-S(t^*|x))$ for the patients from the low risk profile was 32%, which was smaller than 42% and 43% for the patients from the medium and high risk profiles, respectively. A similar pattern of results was observed for the sudden cardiac death simulation, with an estimate of the overall risk of death of 7%, 10%, and 13% by 15 years for the patients from the low, medium, and high risk profiles, respectively.

For both datasets, the standard deviation of the true prognostic index (PI) for the patients from the medium risk profile was the largest, compared to those for the patients from the low and high risk profiles, suggesting greater separation (discrimination) between low-and high-risk patients. The distribution of the PI for the breast cancer patients from the medium risk profile was approximately symmetric (normal) while it was asymmetric for all other validation scenarios.

Table 3.1: Risk profile of patients in validation scenarios, described by the failure probability $(1 - \hat{S}(t^*|x))$ estimated at a single time point t^*: the overall probability and tertiles (mean over 500 simulations of uncensored data, maximum Monte Carlo standard error=0.0007). The distribution of the true prognostic index is also discussed.

		Failure Probability				Prognostic index		
Data	Risk profiles	Overall	tertile 1	tertile 2	tertile 3	Std.	Skew.	Kurt.
Breast cancer	low risk	0.27	0.13	0.22	0.45	0.67	0.45	3.13
	medium risk	0.34	0.15	0.30	0.57	0.75	0.06	2.55
	high risk	0.42	0.22	0.40	0.65	0.68	-0.20	3.17
Sudden death	low risk	0.07	0.04	0.06	0.12	0.51	1.22	4.63
	medium risk	0.10	0.04	0.08	0.17	0.61	0.70	3.19
	high risk	0.13	0.06	0.10	0.21	0.58	0.45	3.28

For breast cancer $t^* = 3$ years and sudden death $t^* = 15$ years. Std=Standard deviation, Skew=Skewness, and Kurt=Kurtosis

3.4.3 Assessing the effect of censoring

The aim was to investigate the effect of censoring on the performance of the validation measures, and not the effect of model development. Therefore, all validation measures were calculated for the true model, rather than for a model developed using development data. Calculation of the calibration slope and Harrell's C-index was performed using Stata packages `stcox` and `estat concordance` respectively while user written Stata codes were used for the other measures (Appendix B: Figure B.1). The results based on these codes were consistent with those with the corresponding R-packages such as `CPE` for $K(\beta)$, `pec` for IBS, and `f.surev` for V and Stata package `str2d` for D statistic. A reference value (or true value) was calculated for each validation measure by calculating its average over a large number of uncensored survival simulations (10,000), for each of the low, medium, and high risk populations. The effect of censoring was investigated by calculating bias (referred to as 'censoring bias') as the mean of the difference between the estimate of the measure and the reference value, over 500 simulations. The number of simulations required (500) was determined using the formula provided by Burton et al. [78], which is based on the true value of the measure of interest, the variability of the measure, the level of accuracy of the measure we are willing to accept (within 2% of the true value), and the normality of the estimated measure. This specification (500 simulations) provided reasonably low Monte Carlo standard error for the estimates of the measures, which was the case for each of the scenarios.

However, it is difficult to compare the different validation measures with each other due to their differing scales. Therefore, a standardised bias was calculated as follows. Let $\hat{m}_g$ $(g = 1, \ldots, G)$ be the estimate of a measure for the gth simulation, m be the corresponding reference value, and m_0 be the null value that is obtained for the null model, then the standardised bias contribution is

$$B_g = \frac{\hat{m}_g - m}{|m - m_0|} \times 100.$$

The standardised bias estimate B_g can be regarded as a random variable that follows a Normal distribution according to the central limit theorem. A confidence interval for standardised bias can be calculated assuming normality and using the empirical standard deviation of the estimated standardised bias. The empirical confidence intervals are used to make conclusions on whether the bias for a validation measure is significantly different from zero or whether the bias between two measures are significantly different.

3.4.4 Assessing sensitivity to the exclusion of important predictors

The sensitivity of the validation measures to the exclusion of important predictors from the model were also assessed. In this part, models with strong and weak predictors were specified to examine whether the validation measures are able to distinguish between the predictive ability of these models. To assess the sensitivity of the measures, first, the most important predictors were identified by fitting multivariable Cox PH models in the observed data and using the P-values calculated from likelihood ratio tests. The most important predictor identified was excluded from the full model (say Model 1 that contains all available predictors in the data), resulting in a reduced model (Model 2). The validation measures were then used to validate Model 2. Further reduced models were specified by omitting the next most important predictor, along with any others already omitted. The validation measures were calculated for each of these reduced models. All models were developed using the observed (original) data and validated by calculating validation measures using the simulated validation data. In addition, the model χ^2 values for each of the fitted models was calculated in the validation data to

assess how changes in the value of the validation measures were related to changes in the model χ^2 values.

To ensure that the results across the validation measures were comparable, the estimates (average values over 500 simulations) were re-scaled with the full model set at 100% and the null model at 0%. Furthermore, to examine how weak the reduced models were (in terms of predictive ability) relative to the full model, the R^2 values were calculated by regressing the PI derived from the full model on the PIs derived from the reduced models. This procedure is similar in idea to the 'step-down' approach proposed by Harrell [40], where a full model is approximated to a reduced model.

3.5 Results and discussion

3.5.1 Results

3.5.1.1 Effect of censoring

Figure 3.1 shows the distribution of the validation measures, over 500 simulations, for various degrees of censoring. Only the results for the medium risk profile breast cancer patients with randomly censored survival times are presented. The horizontal dashed line shows the reference value for each validation measure. The median values of C-index and IBS increased with the degree of censoring whereas the median values of R^2_{IBS} and V decreased with increased censoring. This suggests that misleading conclusions may be drawn regarding a model's predictive performance when using these measures in the presence of censoring. In particular, C-index may give an over-optimistic estimate of model discrimination in the presence of censoring, whereas IBS, R^2_{IBS} and V are likely to be conservative. The median values for the other measures were little affected by censoring. The inter-quartile range for the validation measures generally increased with increased level of censoring. This was perhaps most noticeable for IBS and R^2_{IBS}. Similar results were obtained in the other simulation scenarios (not shown).

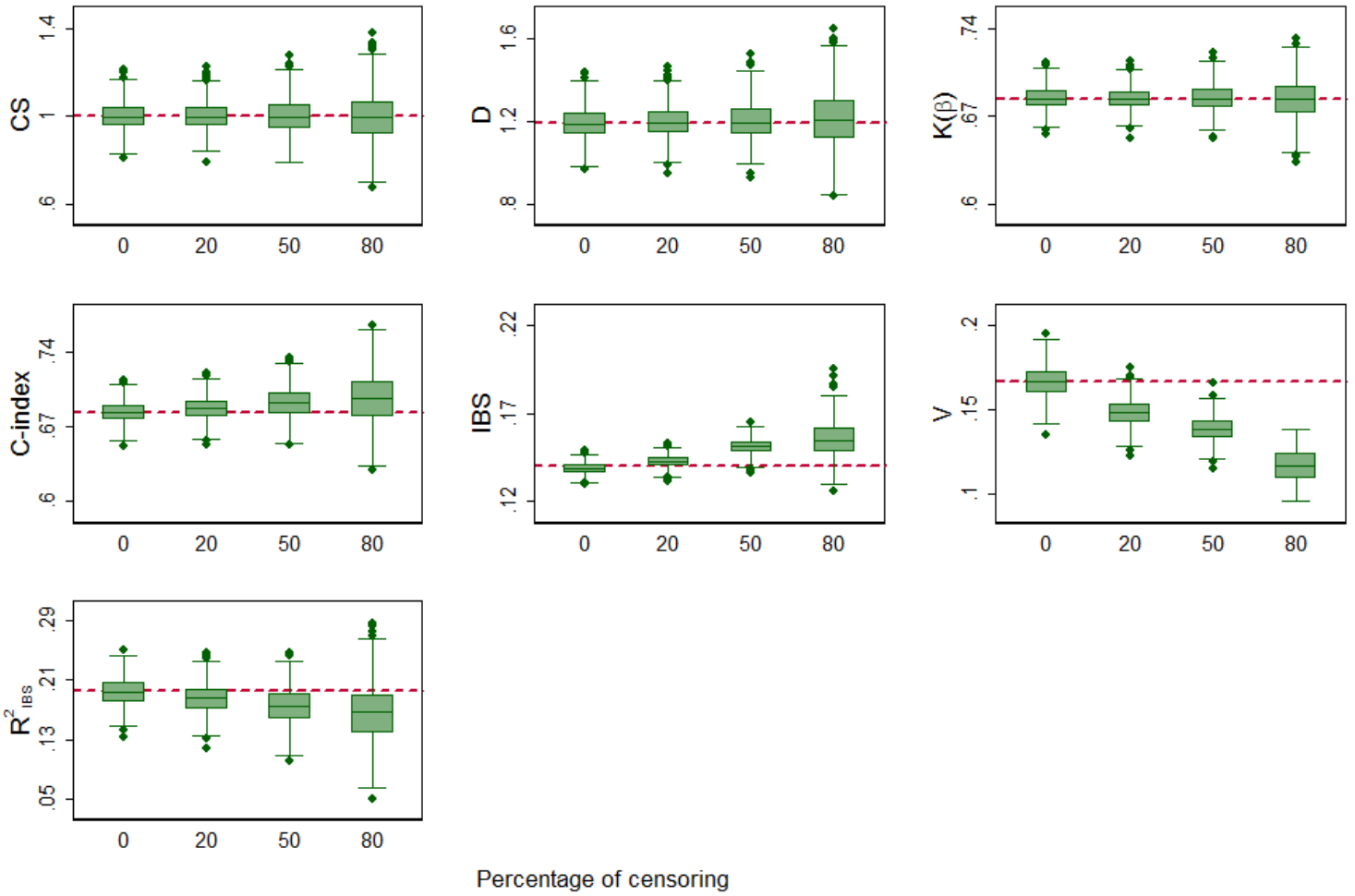

Figure 3.1: Empirical distribution of the validation measures by degree of censoring was summarised using box plots. The results are from the medium risk breast cancer simulations under the random censoring mechanism. The horizontal dashed line indicates the true/reference value of the respective measure.

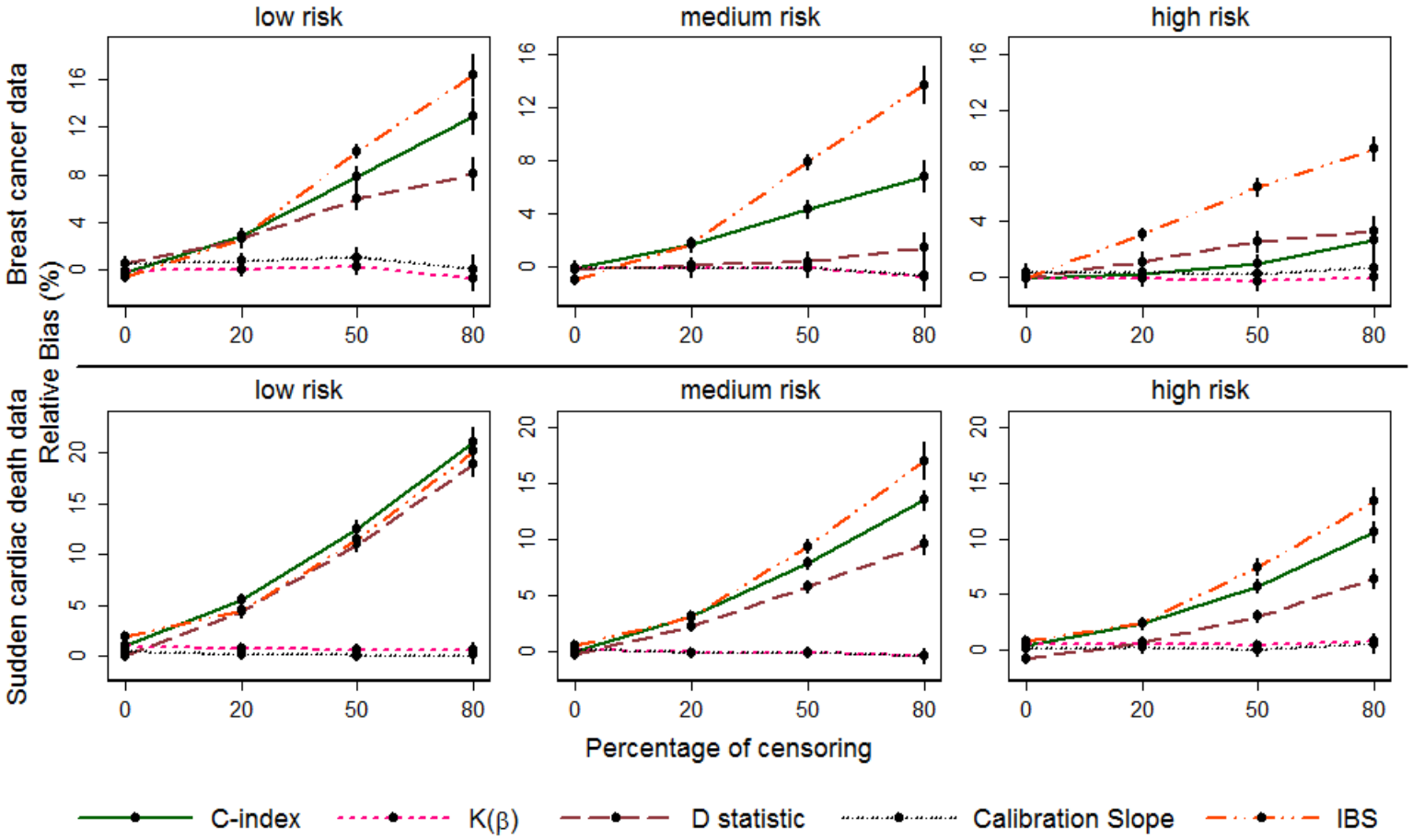

Figure 3.2: Relative bias (%) with 95% confidence intervals for the C-index, $K(\beta)$, D statistic, Calibration slope, and Integrated Brier score (IBS). The first and second rows show the results for the breast cancer and sudden cardiac death simulations with different risks profile (low, medium, and high), respectively. All simulations were under the random censoring mechanism.

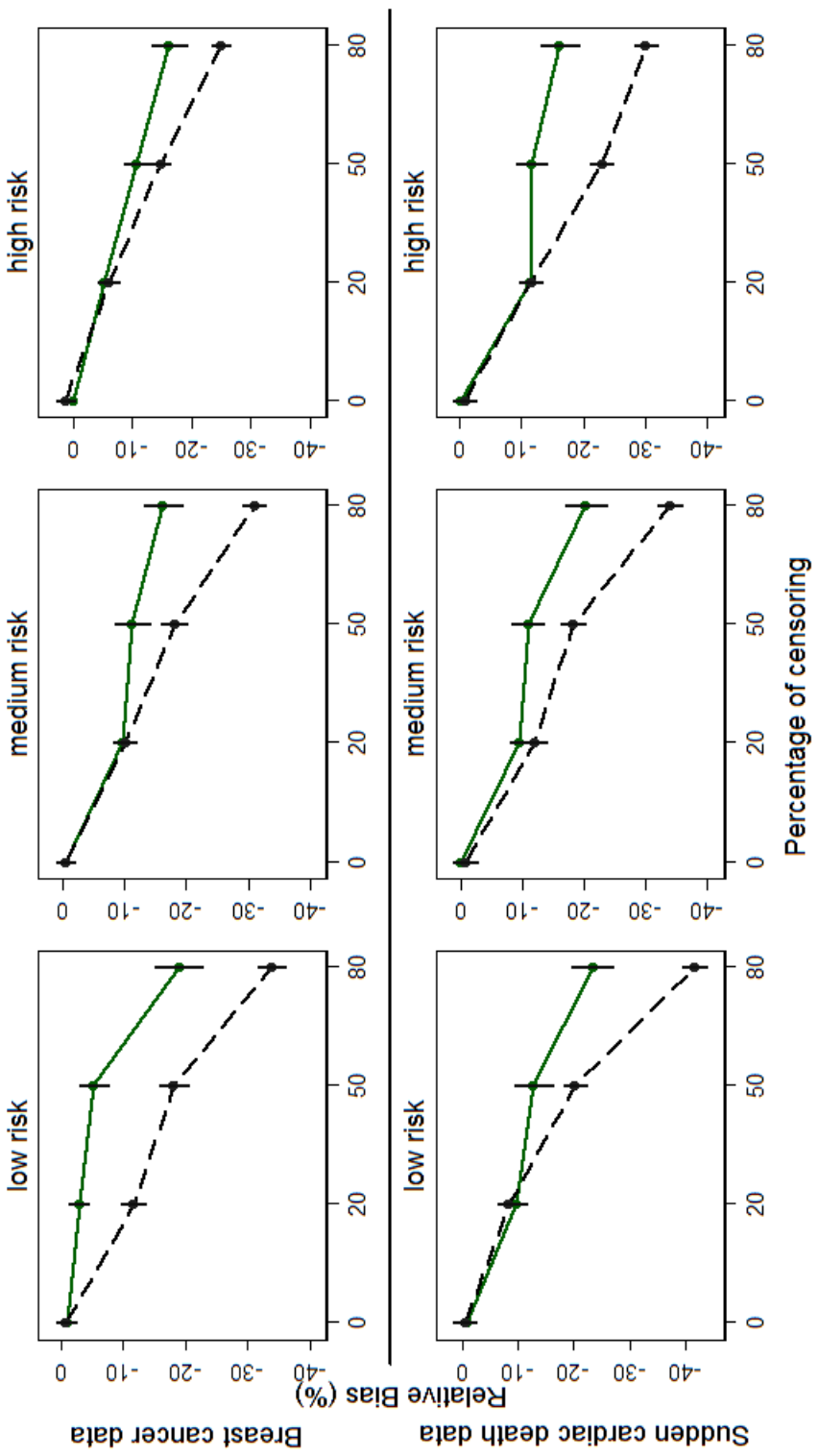

Figure 3.3: Relative bias (%) with 95% confidence intervals for R^2_{IBS} and V. The first and second rows show the results for the breast cancer and sudden cardiac death simulations with different risks profile (low, medium, and high), respectively. All simulations were under the random censoring mechanism.

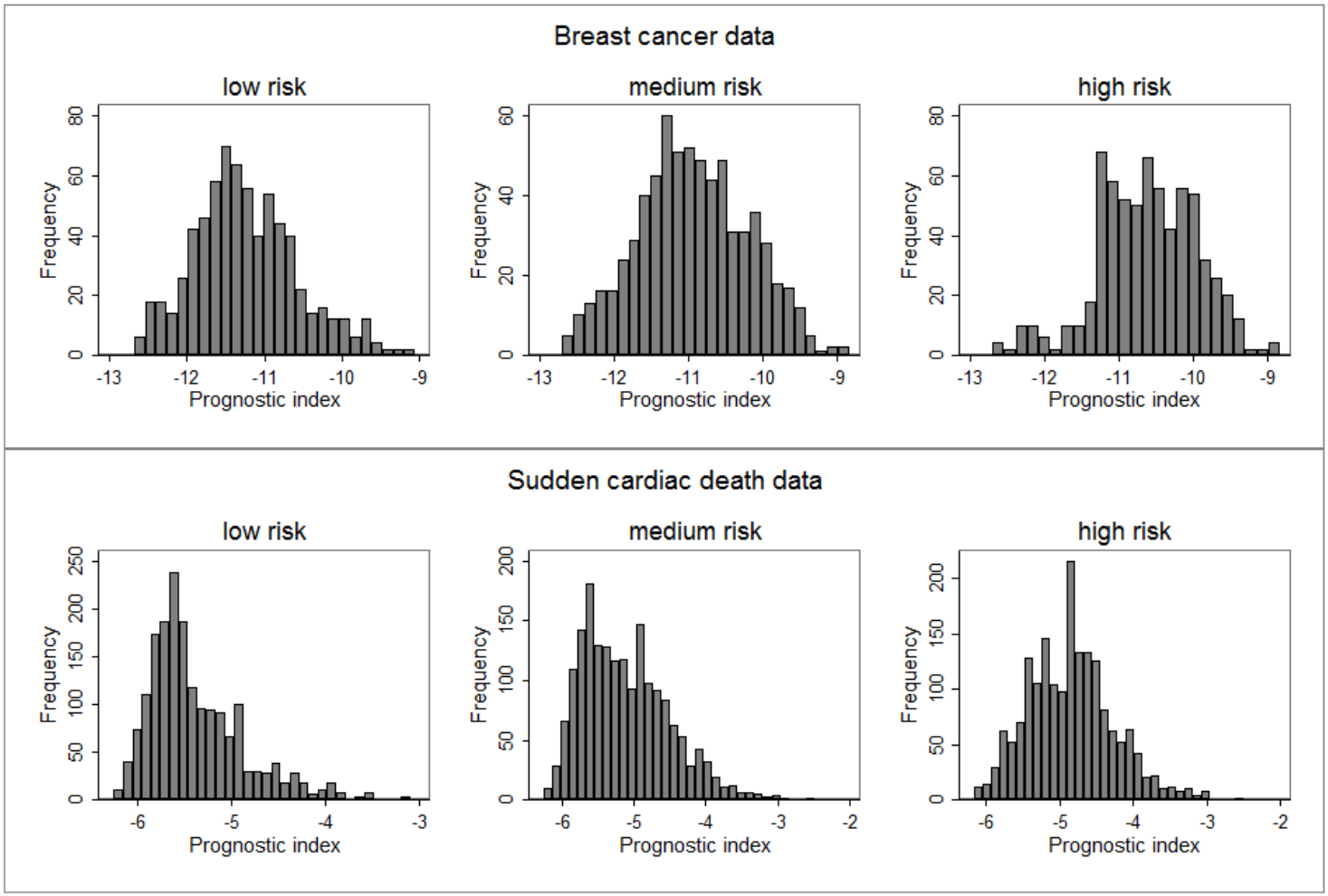

Figure 3.4: Distribution of the prognostic index derived from the true model for the breast cancer and sudden cardiac death patients with different risk profiles.

Figure 3.2 shows the standardised bias for the C-index, $K(\beta)$, D statistic, calibration slope (CS) and IBS, and Figure 3.3 shows bias for R^2_{IBS} and V measures. The results are from both the breast cancer and sudden cardiac death (SCD) simulations when the censoring is random. The bias in CS and $K(\beta)$ was negligible which is to be expected since both are derived from the Cox model. The other measure, derived from this model, the D statistic was biased in some scenarios. For example, the bias was negligible in the medium risk breast cancer scenario, whereas the bias was often high in the SCD scenarios. Further investigation suggests that the level of bias in D corresponds to the level of skewness in the distribution of the prognostic indices (Figure 3.4). Royston and Sauerbrei [49] note that the D statistic is most accurate when the prognostic index is normally distributed. The C-index, one of most widely used measures in practice, showed increasing bias as the level of censoring increased, which may be expected since it depends on the censoring mechanism. In addition, when there are high levels of censoring, the proportion of patient pairs used in the calculation of C-index is relatively small and may not be representative of the patient pairs in the population [37, 48]. Further investigation suggests that the bias in C-index may be acceptable for censoring up to 30% (additional results not shown). The measures of predictive accuracy and explained variation were most affected by censoring, even at low levels, despite their use of weighting to alleviate the effect of censoring. Similar results were observed for the administrative censoring scenarios (Appendix A: Table A.1).

3.5.1.2 Sensitivity to the exclusion of important predictors

In the breast cancer data, the Cox PH analysis identified number of lymph nodes (lnod), progesterone status (progest), hormone therapy (hormon), and menopausal status (menpst) as strong predictors and tumour grade (tgrad), age as moderate predictor, and tumour size (tsize) as weak predictor (Appendix A: Table A.2). In the sudden cardiac death data, the analysis identified number of runs of ventricular tachycardia (runvent), obstruction to blood flow (BF), and abnormal blood pressure response to exercise (BP) as strong predictors while age and maximum thickness of heart muscle (HM) as weak predictors (Appendix A: Table A.3).

Following the procedures described in Section 3.4.4, first, a full model that included all available predictors (both strong and weak predictors) was developed using the observed data, then reduced models were fitted to the same data by excluding important (strong) predictors. For each of these models, the value of the validation measures and the model χ^2 value were calculated in the simulated validation data. The predictive ability of each of these models developed using the breast cancer data (5 in total including the full model) are summarised in Table 3.2 in terms of R^2 as discussed in Section 3.4.4. It appears that the Model 5 was the weakest model, relative to the full model. The reduction in the R^2 values from the value for the full model to that obtained for the Model 5 was relatively sharp for the high risk validation data than those for the low and medium risk validation data.

Table 3.2: Models with different predictive abilities, relative to the full model, are summarised in terms of R^2 values. The results are from the breast cancer simulations with different risk profiles. No censoring was considered.

Models	Predictors in the model	Dropped predictor	R^2		
			Low	Med	High
Full Model	lnod+progest+hormon+menpst+age+tgrad+tsize	-	1.00	1.00	1.00
Model 2	progest+hormon+menpst+age+tgrad+tsize	lnod	0.61	0.64	0.57
Model 3	hormon+menpst+age+tgrad+tsize	progest	0.47	0.44	0.30
Model 4	menpst+age+tgrad+tsize	hormon	0.39	0.38	0.25
Model 5	age+tgrad+tsize	menpst	0.37	0.35	0.23

Low=Low risk, Med=Medium risk, and High=High risk

Figure 3.5 shows the results of sensitivity of the measures for the breast cancer simulations. All the validation measures, except the calibration slope, showed monotonic sensitivity to the omission of important predictors, although none were as sensitive as model χ^2 in the low and medium risk scenarios. The measures belonging to the category of predictive accuracy and explained variation (V, IBS and R^2_{IBS}) were the most sensitive, with V closely following the model χ^2 value in the high risk scenarios. This may be because these measures are calculated using the individual predictions directly and thus are more sensitive to changes in the prognostic strength of the model. The least sensitive measures were C-index which may be expected since they are both pure rank based measures and do not incorporate the actual difference between predictions. It is worth noting that there was less variation across the measures in the high risk scenarios.

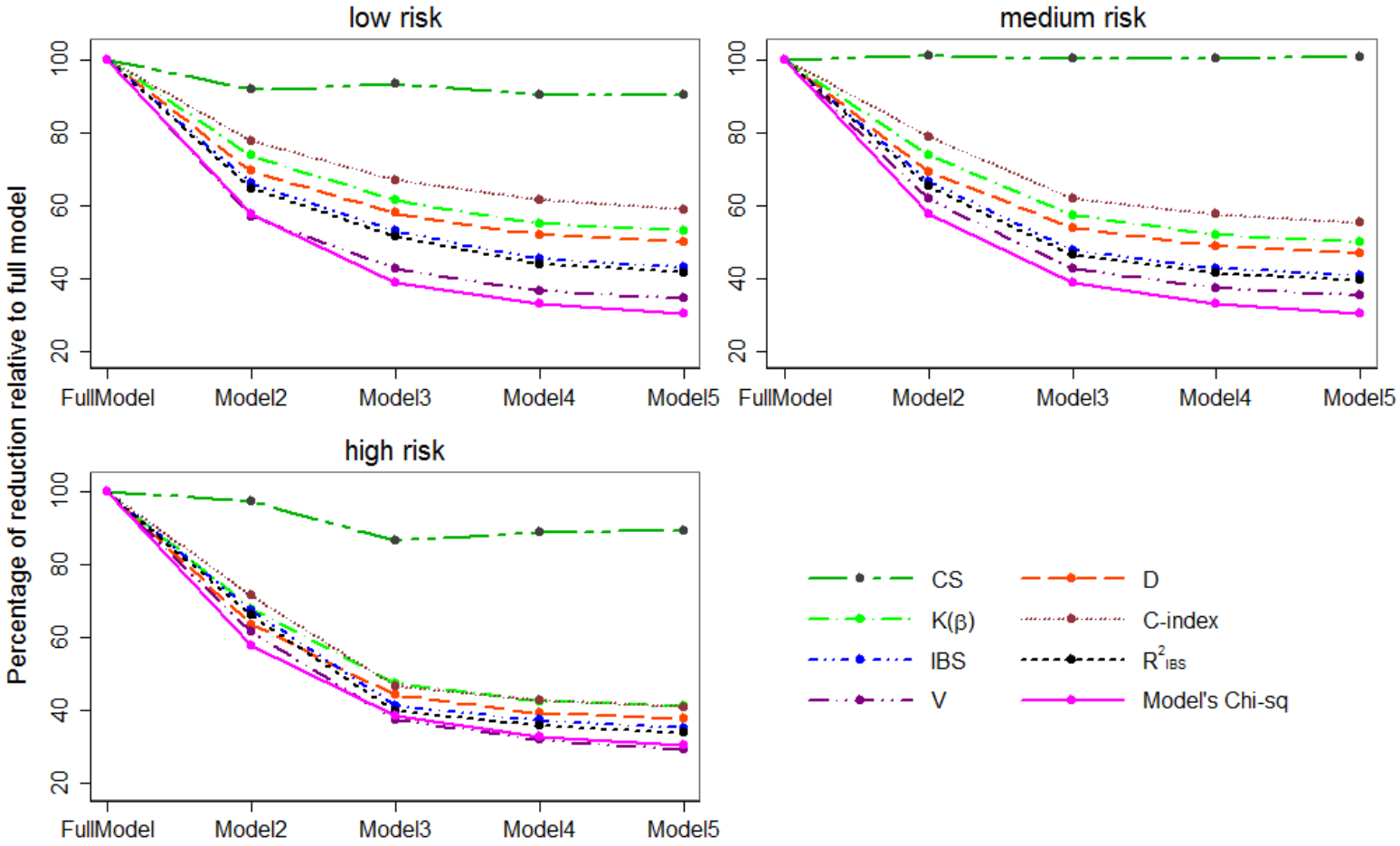

Figure 3.5: Sensitivity of the measures to the exclusion of important predictors, described as the percentage of reduction in a measure's value relative to that for the full model. The results are from the breast cancer simulations with different risk profiles. No censoring was considered.

The value of the calibration slope was little affected by the omission of important predictors when the risk profile of the validation data matched that of the development data (Figures 3.5: Medium risk). In this situation, the relationship between the outcome and the remaining predictors should be similar in both the development and validation data, and hence the calibration slope should indicate good calibration (values close to 1). If the risk profiles are different, then the relationship between the outcome and the included predictors may be different in the development and validation data due to the correlation between these predictors and the omitted predictors, and hence some sensitivity may be observed. The level of sensitivity may be difficult to predict since it depends on the strength of the predictors and the correlation between them. Similar results were seen for the sudden cardiac death simulations (not shown).

3.5.1.3 Relationship between the validation measures

The relationships between the various measures, excluding CS, are shown in Figure 3.6 for the medium risk breast cancer scenario. There was very good agreement between C-index, IBS, R^2_{IBS} and V when there was no censoring, although these relationships weakened considerably as censoring increased. Similar relationships were seen for the low and high risk breast cancer scenarios (results not shown). Generally, these relationships were weaker in the sudden cardiac death (SCD) scenarios (results not shown), perhaps reflecting the lower amount of prognostic information available. However, there was excellent agreement between $K(\beta)$ and D in the breast cancer scenarios and this relationship was robust to censoring. This relationship was weaker in the SCD scenarios which may be due to non-normality of the prognostic indices.

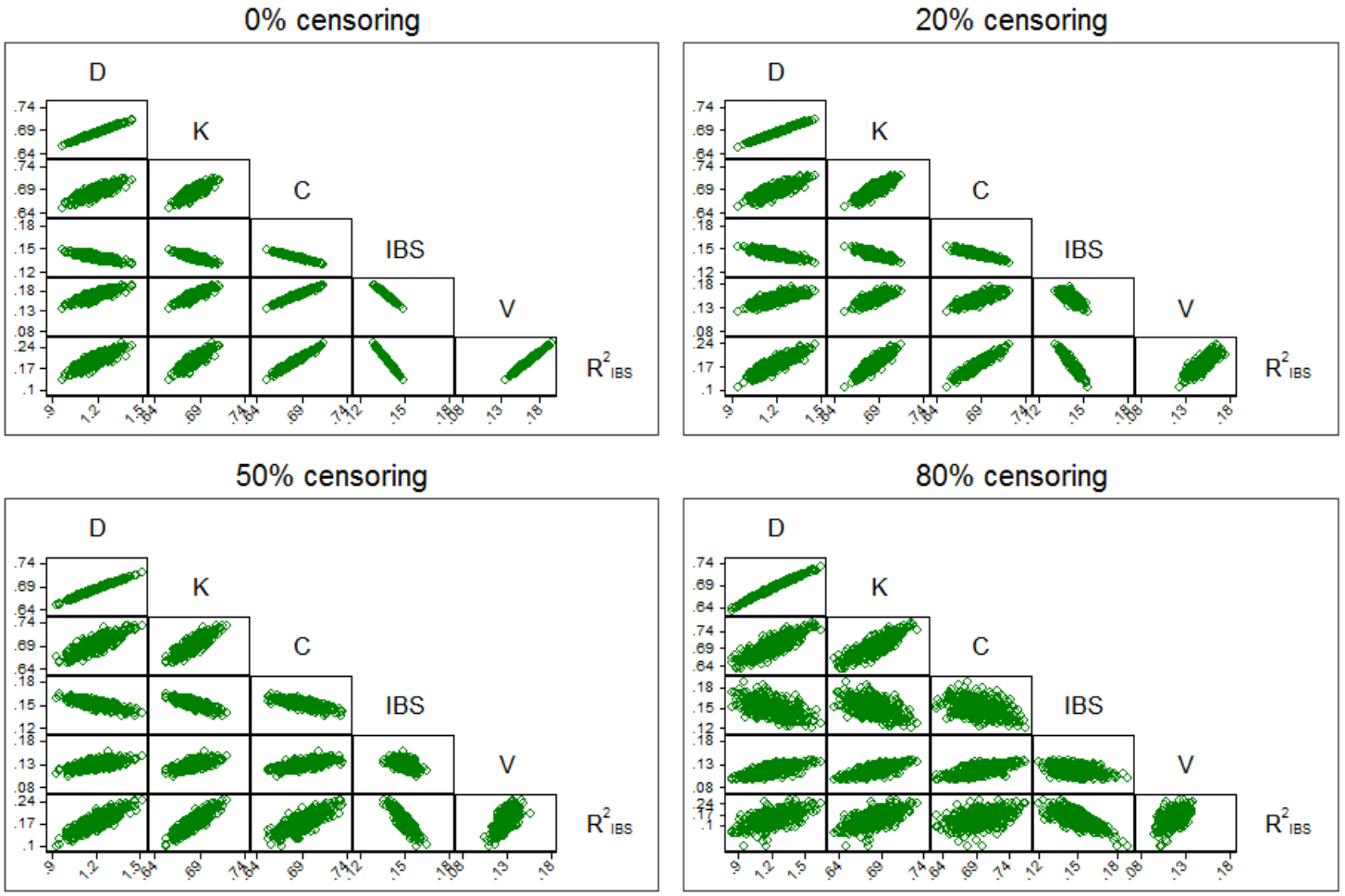

Figure 3.6: Empirical agreement between the measures by degrees of censoring. The results are from the medium risk breast cancer simulation under the random censoring mechanism.

3.5.2 Discussion and recommendations

When developing a risk prediction model for survival data it is essential that the performance of the model is evaluated using appropriate validation measures. Although a number of measures have been proposed, there is only limited guidance regarding their use in practice. The aim of this research was to perform a simulation study based on two clinical datasets with contrasting characteristics to investigate a wide range of validation measures in order to make practical recommendations regarding their use.

Based on the simulation study, the measures of predictive accuracy (IBS) and explained variation (V and R^2_{IBS}) cannot be recommended for use with survival risk models due to their poor performance in the presence of censored data. However, these measures were all conservative with censored data so that high (or low for IBS) values would still be indicative of a good risk model. Of the discrimination measures, $K(\beta)$ was not biased in the presence of censoring. The performance of D in the presence of censoring depended on the distribution of the prognostic index. Provided that the prognostic index was approximately normally distributed, the effect of censoring on the bias in D was negligible. The C-index was affected by censoring and cannot be recommended for use with data with more than 30% censoring. The sole calibration measure under investigation, CS, was unbiased in the presence of censoring.

All the measures of discrimination, predicted accuracy and explained variation showed sensitivity to the omission of important predictors from a model. However, the ranked-based measure C-index was less sensitive than the other measures. The calibration slope showed only limited sensitivity to predictor omission since the developed risk model effectively re-calibrates itself to compensate for the omitted predictors.

The validation measures differ in their flexibility regarding their assumptions and the form of the risk model. The concordance measure C-index only require that the risk model is able to rank the patients. In contrast, $K(\beta)$ requires that the risk model was fitted using the Cox proportional hazards model. The D statistic assumes that proportional hazards holds and that the prognostic index is normally distributed. The calibration slope measure, as described, also assumes proportional hazards although

more general approaches are described by van Houwelingen [44]. The measures based on predictive accuracy, IBS, R^2_{IBS}, and V, only require that a survival function can be calculated for all patients.

With respect to clinical interpretation, all of the measures considered in this paper can be easily communicated to a non-statistical health researcher, except perhaps for the calibration slope and IBS. The concordance measures can be readily communicated in terms of correctly ranking patient pairs, and explained variation measures are intuitive with their percentage scale. The D statistic also has a nice interpretation as it can be communicated as a (log) relative risk between low and high risk groups of patients.

In summary, based on the findings of this simulation study, $K(\beta)$ can be recommended for validating a risk model developed using the Cox proportional hazards model, since it is both robust to censoring and reasonably sensitive to the omission of important predictors. D can also be recommended provided that the distribution of the prognostic index is approximately normal. It is more sensitive to predictor omission than $K(\beta)$ and can be calculated for models other than those fitted using the Cox model. The calibration slope can be recommended as a measure of calibration since it is not affected by censoring although it is less sensitive than the other measures to the omission of important predictors. In practice, one might additionally investigate calibration graphically by comparing observed and predicted survival curves for groups of patients. This approach also has the benefit of being easy to communicate.

An important point to note is that the characteristics of the validation data should be investigated before choosing the validation measures. In particular, the level of censoring and the distribution of the prognostic index need to be checked, assuming that the standard model assumptions such as proportional hazards hold. It is not clear that this is routinely done in practice.

3.6 Conclusion

This chapter investigated some of the validation measures that have been used for independent survival outcomes. By means of a simulation study based on two real datasets, this investigation compared their performance against criteria for a suitable validation measure for a survival model. The results in the simulation study provided guidelines for using these measures in practice, particularly when data have censoring.

The next chapter discusses the possible extensions of validation measures that have been used for independent binary outcomes for use with correlated/clustered binary outcomes.

Chapter 4

Measures for clustered binary outcomes

4.1 Introduction

Clustered binary outcomes occur frequently in health care research. For example, subjects could be nested in larger units such as hospitals, doctors, family, or geographic regions. Due to clustering within larger units, outcomes in the same cluster often share some common cluster level characteristics and thus tend to be correlated. Various statistical models have been proposed in the last two decades to model the relationship between predictors and outcomes in the presence of clustering, particularly focusing on how to account for the effect of clustering. These models are typically grouped into two broad classes: cluster-specific and population-averaged approaches [79, 80].

In the cluster-specific approach, the probability distribution of the outcomes is modelled as a function of fixed predictors and one or more random terms. The random term represents the effect of unobserved cluster-specific characteristics, which varies across clusters following a specific distribution. This modelling approach is known as the random effects model, for example, random effects logistic model for clustered binary outcomes [81, 82]. In the population-averaged approach, the marginal or population averaged expectation is modelled as a function of predictors, treating the correlation be-

tween the outcomes within the same cluster as a nuisance parameter. Marginal logistic models, with generalized estimating equations [83] for the estimation of the model parameters, are often used for modelling clustered binary outcomes. The estimates from the random effects models have a conditional interpretation, given the cluster-specific random effect, while the estimates from the marginal models have population-averaged interpretation. The conditional estimates from a logistic model can be interpreted as the effect of a unit change in the predictors for subjects belonging to the same cluster, whereas the marginal estimates can be interpreted as the averaged effect of a unit change in the predictors for all subjects in the population. Generally, the preference for using one of these two classes of models depends on what type of inference a researcher would like to draw in practice: conditional or marginal [84]. Lee and Nelder [85] and Skrondal and Rabe-Hesketh [86] considered the random effects models as more general form of models for analysing clustered binary data, from which the marginal models can be derived by integrating out the random effects. It is thus possible to obtain both conditional and marginal predictions from the random effects models.

Although the clustering of data within larger units is usually taken into account in explanatory models in aetiological research, it is often ignored in risk prediction research, both in the process of model development and the validation of the model's performance [87]. This work focuses on the use of random effects logistic models in risk prediction for clustered binary outcomes. To understand the predictive ability of such a model, it is essential to validate its predictive performance. Validation measures for assessing the predictive ability of models for independent binary outcomes are reasonably well developed; see, for example, Omar et al. [10], Steyerberg et al. [24], Royston and Altman [25], and Harrell et al. [40]. However, very limited research has been conducted to develop validation measures for models with clustered binary outcomes. This chapter discusses possible extensions of some of the existing validation measures that could be used to assess the predictive ability of prognostic models based on the random effects logistic models.

The C-index [45], and the D-statistic [49] are commonly used validation measures to assess the discriminatory ability of prognostic models for independent binary out-

comes. The calibration slope [39, 42] is commonly used to assess whether the model predicts accurately for a group of subjects (calibration), and the Brier score [55] is often used to assess accuracy for individual predictions (predictive accuracy). In this chapter, these validation measures are extended for use with models for clustered binary outcomes. The Hosmer-Lemeshow Chi-squared test statistic [41] is also used frequently to assess a model's calibration. This test assesses whether or not the observed event rates match the expected event rates in subgroups of model population, where the groups are identified from the deciles of the predicted risk of having the event. However, it is not straightforward to evaluate this measure using a simulation study. Therefore, this measure is not investigated for the models with clustered binary outcomes.

The chapter begins with a brief description of the proposed validation measures for independent binary outcomes, then discusses the estimation of these measures for clustered data. The methods are illustrated using data on patients who had undergone heart valve surgery. A simulation study is conducted to evaluate the performance of the validation measures under various clustered data scenarios.

4.2 Validation measures for independent binary outcomes

This section briefly describes some of the commonly used validation measures for independent binary outcomes, starting with a description of notation based on the logistic regression model.

4.2.1 Logistic regression model

Let Y_i $(i = 1, \ldots, N)$ be a binary outcome (0/1) for the ith subject which follows Bernoulli distribution with the probability $\pi_i = \Pr(Y_i = 1)$. The logistic regression model can be used to model the relationship between the outcome and predictors and is defined as

$$\text{logit}[\Pr(Y_i = 1|\boldsymbol{x}_i)] = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{\beta}^T\boldsymbol{x}_i,$$

where $\boldsymbol{\beta}^T$ is a vector of regression coefficients of length $(p+1)$, and $\boldsymbol{x}_i$ is the ith row vector of the predictor matrix $\boldsymbol{X}$ which has order $N \times (p+1)$. The term $\eta_i = \boldsymbol{\beta}^T \boldsymbol{x}_i$ is known as the 'prognostic index'. The predictive form of this model, used to predict the probability of the event of interest, can be written as

$$\pi(\boldsymbol{\beta}|\boldsymbol{x}_i) = \frac{1}{1 + \exp[-\boldsymbol{\beta}^T \boldsymbol{x}_i]}.$$

Predictions from the model depend on the estimate of $\boldsymbol{\beta}^T$, which is typically obtained by the method of maximum likelihood [88].

4.2.2 The C-index: definition

The C-index is a measure of concordance probability and is numerically identical to the area under the receiver operating characteristic curve (AUC) [45], a graph of sensitivity (true positive rate) against 1-specificity (false positive rate). The C-index is widely used as a tool for assessing the discriminatory ability of standard logistic models because of its straightforward clinical interpretation. The C-index equals to the proportion of pairs in which the predicted event probability is higher for the subject who experienced the event of interest than that of the subject who did not experience the event. For a pair of subjects (i, j), where i and j correspond to those who experienced the event and those who did not respectively, with event probabilities $\{\pi(\boldsymbol{\beta}|\boldsymbol{x}_i), \pi(\boldsymbol{\beta}|\boldsymbol{x}_j)\}$, the C-index can be defined as

$$C = \Pr[\pi(\boldsymbol{\beta}|\boldsymbol{x}_i) > \pi(\boldsymbol{\beta}|\boldsymbol{x}_j)|Y_i = 1 \;\&\; Y_j = 0].$$

Since there exists a one-to-one transformation between π and $\boldsymbol{\beta}^T \boldsymbol{x}$, the above probability expression can be written as

$$C = \Pr[\boldsymbol{\beta}^T \boldsymbol{x}_i > \boldsymbol{\beta}^T \boldsymbol{x}_j|Y_i = 1 \;\&\; Y_j = 0].$$

The C-index from standard logistic regression models can be estimated using both parametric and nonparametric approaches. Generally, under the parametric approach, a distributional assumption is required for the prognostic index for the population who

had experienced the event and for those who did not. Under the assumption of normal distribution, the method of maximum likelihood may be used to estimate the C-index [89, 90].

The widely used non-parametric approach to estimate the C-index is based on the Mann-Whitney U statistic [91] and does not require any distributional assumptions regarding the prognostic index. The C-index or AUC has been shown to be equal to the U statistic when it (the area) is calculated using the trapezoidal rule [45, 92]. The U statistic is usually computed to test whether the levels of a quantitative variable in one population tend to be greater than those in a second population, without making any distributional assumptions for the variable. In this chapter, both the parametric and nonparametric approaches for estimating the C-index are discussed.

4.2.3 Non-parametric estimation of the C-index

Let $\eta_i^{(1)} = \boldsymbol{\beta}^T \boldsymbol{x}_i | Y_i = 1$ and $\eta_j^{(0)} = \boldsymbol{\beta}^T \boldsymbol{x}_j | Y_j = 0$ be the prognostic index derived by the model for subject i who had experienced the event and for subject j who did not, respectively. Further, let N_1 and N_0 be the number of events and non-events, respectively. Considering all pairs (i, j), the C-index can be estimated by analogy to the U statistic formulation [45, 91, 92] as

$$C^{np} = \frac{1}{N_1 N_0} \sum_{i=1}^{N_1} \sum_{j=1}^{N_0} I\big(\eta_i^{(1)}, \eta_j^{(0)}\big), \tag{4.1}$$

where

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

The value of C^{np} ranges between 0.5 and 1: a value of 0.5 indicates that the model has no ability to discriminate between low and high risk subjects, whereas a value of 1 indicates that the model can perfectly discriminate between these two groups.

4.2.4 Parametric estimation of the C-index

Based on the central limit theorem, the prognostic index is likely to follow normal distribution as the dimension of the parameter vector $\boldsymbol{\beta}$ increases [52]. The estimation of the parametric C-index is as follows.

Let us assume that $\eta_i^{(1)} = \boldsymbol{\beta}^T \boldsymbol{x}_i | Y_i = 1 \sim \mathrm{N}(\mu_1, \sigma^2)$ and $\eta_j^{(0)} = \boldsymbol{\beta}^T \boldsymbol{x}_j | Y_j = 0 \sim \mathrm{N}(\mu_0, \sigma^2)$. Therefore, $\eta_i^{(1)} - \eta_j^{(0)} \sim \mathrm{N}(\mu_1 - \mu_0, 2\sigma^2)$. By definition, the parametric C-index is

$$\begin{aligned} C^p &= \Pr[\eta_i^{(1)} > \eta_j^{(0)}] \\ &= \Pr[(\eta_i^{(1)} - \eta_j^{(0)}) > 0]. \end{aligned}$$

After standardising the term $\eta_i^{(1)} - \eta_j^{(0)}$, C^p can be obtained as

$$\begin{aligned} C^p &= \Pr\Big[Z < \frac{\mu_1 - \mu_0}{\sqrt{2\sigma^2}}\Big], \quad Z \sim \mathrm{N}(0,1) \\ &= \Phi\left(\frac{\mu_1 - \mu_0}{\sqrt{2\sigma^2}}\right), \end{aligned} \tag{4.2}$$

where Φ denotes the standard normal cumulative distribution function. The estimate of C^p can be obtained by replacing μ_1, μ_0, and σ^2 by their sample estimates $\bar{x}_1$, $\bar{x}_0$, and S^2, respectively.

4.2.5 D statistic

The D statistic [49] is a measure of prognostic separation and quantifies the separation between two equal-sized prognostic groups obtained by dichotomising the predicted prognostic indices at their median value. The D statistic for the logistic regression model can be calculated by transforming the predicted prognostic index $\hat{\eta}_i = \boldsymbol{\beta}^T \boldsymbol{x}_i$ to a standard normal order statistic z_i, in a manner similar to that for the Cox PH model. A logistic model is then fitted to the validation data with z as the sole predictor:

$$\mathrm{logit}(Y_i = 1 | z_i) = \beta_z z_i.$$

The estimated coefficient of z is an estimate of the D statistic, $\hat{D}$, and the corresponding estimated standard error of z is the standard error of $\hat{D}$. $\hat{D}$ is interpreted as the log odds ratio of having the event of interest between low-and high-risk groups, where the groups represent the lower and upper half of the predicted prognostic index, respectively. The null value for D is 0, with increasing values indicating greater separation (discrimination) between these two groups.

4.2.6 Relationship between the C-index and D statistic

The C-index and D statistic are closely related under the assumption of normality of the prognostic index $\eta_i = \boldsymbol{\beta}^T \boldsymbol{x}_i$. Based on this assumption, an analytical relationship between the parametric C-index and D-statistic is derived as follows.

Let us assume that

$$\eta_i^{(1)} = \boldsymbol{\beta}^T \boldsymbol{x}_i | Y_i = 1 \sim \mathrm{N}(\mu_1, \sigma^2) \text{ with } \Pr(Y_i = 1) = \pi_1$$

and

$$\eta_j^{(0)} = \boldsymbol{\beta}^T \boldsymbol{x}_j | Y_j = 0) \sim \mathrm{N}(\mu_0, \sigma^2) \text{ with } \Pr(Y_j = 0) = \pi_0.$$

Further, suppose that the conditional distribution of η given $Y = y$ is

$$\eta | Y = y \sim \mathrm{N}\Big(\mu_y, 2\sigma^2\Big).$$

The above formulation corresponds to linear discriminant analysis (LDA) [93], which is equivalent to logistic regression model [94]. In LDA, we assign subject i with prognostic score $\eta_i = \boldsymbol{\beta}^T \boldsymbol{x}_i$ to the population who had experienced the event with probability $\Pr(Y_i = 1 | \eta_i)$. This probability can be expressed in terms of a logistic model as

$$\Pr(Y_i = 1 | \eta_i) = \frac{1}{1 + \exp[-(\beta_0 + \beta_\eta \eta_i)]}, \tag{4.3}$$

where $\beta_0 = -\log \dfrac{\pi_1}{\pi_0} + \dfrac{1}{2} \dfrac{(\mu_1^2 - \mu_0^2)}{2\sigma^2}$ and $\beta_\eta = \dfrac{(\mu_1 - \mu_0)}{2\sigma^2}$.

Standardising the prognostic index η and then multiplying it by $\sqrt{\pi/8}$ gives the term Z' (say), which is distributed as $N(0, \pi/8)$. This standardised statistic is approximately equivalent to the standard normal order statistic Z which we obtained

from η through a transformation when calculating the D statistic (see Section 4.2.5). Therefore, the standardised versions of $\eta^{(1)}$ and $\eta^{(0)}$ can be written as

$$Z'^{(1)} \sim \mathrm{N}\left(\frac{\mu_1 - E(\eta)}{\sqrt{\mathrm{var}(\eta)}}, \frac{2\sigma^2}{\mathrm{var}(\eta)}\pi/8\right) = \mathrm{N}\left(\frac{\mu_1 - E(\eta)}{\sqrt{2\sigma^2}}, \pi/8\right)$$

and

$$Z'^{(0)} \sim \mathrm{N}\left(\frac{\mu_0 - E(\eta)}{\sqrt{\mathrm{var}(\eta)}}, \frac{2\sigma^2}{\mathrm{var}(\eta)}\pi/8\right) = \mathrm{N}\left(\frac{\mu_0 - E(\eta)}{\sqrt{2\sigma^2}}, \pi/8\right),$$

respectively. This formulation also corresponds to the LDA with the transformed variable Z' and can be expressed in terms of a logistic regression model for the binary outcome Y with Z' as a predictor:

$$\Pr(Y_i = 1 | Z'_i = z'_i) = \frac{1}{1 + \exp[-(\beta_0 + \beta_{z'} z'_i)]}.$$

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 in equation (4.3) as

$$\begin{aligned} D &\approx \frac{E[Z'^{(1)}] - E[Z'^{(0)}]}{\mathrm{var}(Z')} \\ &= \frac{\frac{\mu_1 - E(\eta)}{\sqrt{2\sigma^2}} - \frac{\mu_0 - E(\eta)}{\sqrt{2\sigma^2}}}{\pi/8} \\ &= (8/\pi)\left(\frac{\mu_1 - \mu_0}{\sqrt{2\sigma^2}}\right). \end{aligned} \tag{4.4}$$

By using equation (4.2), equation (4.4) can be written as

$$D \approx (8/\pi)\Phi^{-1}(C^p), \tag{4.5}$$

where $\Phi^{-1}(.)$ is the inverse standard normal distribution function. An illustration of the above relationship is as follows. Under the null situation, if $C^p = 0.5$ indicating no ability of the model (possibly the null model) to discriminate between the low and high risk subjects, then D from equation (4.5) is equal to 0, indicating no separation (discrimination) between those two groups. Similarly, if $C^p = 0.75$, the approximate

equivalent value of D is 1.72, indicating reasonably good separation. Both the C-index and D statistic have their own clinical interpretations: the former can be readily communicated in terms of correctly ranking patient pairs and the latter can be communicated as a (log) relative risk between low and high risk groups of patients. Therefore, to have different clinical interpretation in practice, one can quickly obtain the value of the D statistic knowing the value the C-index and vice-versa, rather than calculating from the model.

4.2.7 Calibration slope

The calibration slope (CS) assesses the calibration of the model by quantifying the agreement between the observed outcome and prediction for a group of subjects. The calibration slope can be obtained by fitting a logistic model with the prognostic index $\hat{\eta}_i = \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$, calculated from the validation sample, as the only predictor in the model [39, 42]:

$$\text{logit}[\Pr(Y_i = 1|\hat{\eta}_i)] = \beta_0 + \beta_\eta \hat{\eta}_i, \tag{4.6}$$

where $\widehat{CS}$ is equal to $\hat{\beta}_\eta$. If $\hat{\beta}_\eta$ is close to 1 then it suggests that the prognostic indices (log odds) derived from the model are accurate. If $\hat{\beta}_\eta$ is somewhat different from 1, it suggests that some form of re-calibration is necessary [24, 38, 40, 59, 95]. In particular, a value much smaller than 1 indicates over-fitting, where risk estimates are too low for low risk subjects and too high for high risk subjects (for more details, see Chapter 2).

4.2.8 Brier score

The Brier score (BS) assesses whether the predictions of the model for each subject are accurate, by quantifying the averaged squared difference between the predicted event probability and the actual outcome [55, 69]. For the logistic model with predicted probability $\hat{\pi}(\hat{\boldsymbol{\beta}}|\boldsymbol{x}_i)$ for subject i, the Brier score is defined as

$$BS = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{\pi}(\hat{\boldsymbol{\beta}}|\boldsymbol{x}_i))^2. \tag{4.7}$$

If the model is predicting perfectly then $BS = 0$, which is however unlikely to occur in practice. Inaccuracy in predictions is indicated by positive value of the BS, and higher values indicate greater inaccuracy. A Brier score value of about 0.33 indicates that predictive ability of the model is not better than random guessing [55, 69].

4.3 Extension of the validation measures for clustered data

This section discusses possible approaches to extend the validation measures discussed above for use with models for clustered binary data. Here a random effects logistic model is considered, where the intercept is the only random parameter. This type of model is usually referred to as a 'random-intercept logistic model', which assumes equal correlation between pairs of subjects in the same cluster. The section begins with describing a random-intercept logistic model and approaches to make predictions using this model, and then discusses how to obtain the validation measures for this model.

4.3.1 Random-intercept logistic model

Let Y_{ij} be a binary outcome variable (1/0) for the ith subject in the jth cluster of size n_j ($i = 1, \ldots, n_j; j = 1, \ldots, J$) and $\sum_{j=1}^{J} n_j = N$. It is assumed that $Y_{ij} \sim \text{Bernoulli}(\pi_{ij})$, where $\pi_{ij} = \Pr(Y_{ij} = 1)$ is the probability of having the event of interest. The random-intercept logistic model is an extension of the standard logistic model with an additional cluster-specific random effect u_j, where u_j acts as an additive component with the intercept of the model and varies randomly between clusters. The random effects u_js represent the effects of cluster-specific unobserved predictor information and are independent and identically distributed random variables. Typically u_js are normal with mean 0 and variance σ_u^2. The variance parameter σ_u^2 is interpreted as the variation in the log-odds of having the event of interest between clusters. The random-intercept logistic regression model is given by:

$$\text{logit}[\Pr(Y_{ij} = 1 | u_j, \boldsymbol{x}_{ij})] = \log\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right) = \boldsymbol{\beta}^T \boldsymbol{x}_{ij} + u_j,$$

where $\boldsymbol{\beta}^T$ is the vector of regression coefficients of length $(p+1)$, and $\boldsymbol{x}_{ij}$ is the ith row vector of the p-predictors.

4.3.2 Predictions from the model

The predictive form of the random effect logistic model, to predict the probability of having the event, for subject i in cluster j is given by

$$\pi(\boldsymbol{\beta}|u_j, \boldsymbol{x}_{ij}) = \frac{\exp[\eta(\boldsymbol{\beta}, \boldsymbol{x}_{ij}, u_j)]}{1 + \exp[\eta(\boldsymbol{\beta}, \boldsymbol{x}_{ij}, u_j)]}, \tag{4.8}$$

where $\eta(\boldsymbol{\beta}, \boldsymbol{x}_{ij}, u_j)$=$\boldsymbol{\beta}^T\boldsymbol{x}_{ij} + u_j$ is referred to as the prognostic index. Predictions from the model depend on the estimates of the model parameters $(\boldsymbol{\beta}^T, \sigma_u^2)$ and the random effect u_j.

The model parameters can be estimated using adaptive Gaussian quadrature (AGQ) [96–99] or penalized quasi-likelihood (PQL) [100–102]. Using the estimates of the model parameters, the random effect u_j for the jth cluster can be obtained by empirical Bayes approach [86, 103–105], which is the most commonly used method for estimating random effects. The empirical Bayes estimates are the means of the empirical posterior distribution of u_j, $p(u_j|y_{ij}, \boldsymbol{x}_{ij}; \hat{\boldsymbol{\beta}}^T, \hat{\sigma}_u^2)$ with the parameters estimates $(\hat{\boldsymbol{\beta}}^T, \hat{\sigma}_u^2)$ plugged in, and are given by:

$$\hat{u}_j = E(u_j|y_{ij}, \boldsymbol{x}_{ij}; \hat{\boldsymbol{\beta}}^T, \hat{\sigma}_u^2) = \int u_j p(u_j|y_{ij}, \boldsymbol{x}_{ij}; \hat{\boldsymbol{\beta}}^T, \hat{\sigma}_u^2) du_j, \tag{4.9}$$

where $p(u_j|y_{ij}, \boldsymbol{x}_{ij}; \hat{\boldsymbol{\beta}}^T, \hat{\sigma}_u^2)$ can be derived using Bayes theorem. The Bayes theorem combines the prior distribution of u_j, which is essentially $\mathrm{N}(0, \sigma_u^2)$, and the data $(y_{ij}, \boldsymbol{x}_{ij})$. The above integrals do not have analytical solution and need to be solved numerically. The estimated random effects may be useful to make inferences about particular clusters and to identify outlying clusters [106, 107].

The random effects logistic model formulated in the above way can be used to make both conditional (cluster-specific) and marginal (population-averaged) predictions. The conditional predictions can be made either by using $\hat{\boldsymbol{\beta}}^T$ and plugging in the estimated

random effects $\hat{u}$ or by using $\hat{\boldsymbol{\beta}}^T$ and setting the random effects at their mean value zero ($u = 0$). Marginal predictions can be made by integrating the conditional prediction $\pi(\boldsymbol{\beta}|u, \boldsymbol{x})$ given in equation (4.8) over the (prior) random effects distribution. For convenience, these three forms of model prediction are denoted as $\pi_{ij}(u)$, $\pi_{ij}(0)$, and $\pi_{ij}(\text{pa})$, respectively. Similarly, the prognostic indices derived from these predictive functions are denoted by $\eta_{ij}(u)$, $\eta_{ij}(0)$, and $\eta_{ij}(\text{pa})$, respectively. Note that $\pi_{ij}(0) \neq \pi_{ij}(\text{pa})$, which holds for most models with non-linear link function.

As an alternative to $\pi_{ij}(u)$, a clustered-averaged or posterior mean probability $\bar{\pi}_{ij}(u)$ can be obtained by integrating $\pi(\boldsymbol{\beta}|u, \boldsymbol{x})$ given in equation (4.8) over the posterior distribution of the random effects for cluster j [86]. However, Skrondal and Rabe-Hesketh [86] showed via simulation studies that both $\pi_{ij}(u)$ and $\bar{\pi}_{ij}(u)$ perform equally and have equal mean squared error of predictions for a range of conditions in a clustered data setting. This research considers $\pi_{ij}(u)$ instead of $\bar{\pi}_{ij}(u)$ as it can be obtained from most standard softwares.

The decision to make either cluster-specific or population-averaged predictions should depend on the research question. Some examples of cluster-specific predictions can be found in [108–110] and population-averaged predictions in [111]. Generally, the use of the above three approaches to prediction may also depend on whether the subjects for whom predictions will be made belong to an existing cluster or to a new cluster. Skrondal and Rabe-Hesketh [86] and Oirbeek and Lesaffre [112] suggested that if subjects are from an existing cluster, $\pi_{ij}(u)$ is preferred as the effect of clustering (random effect) for that cluster is known. If subjects are from a new cluster on which information is usually unknown, either $\pi_{ij}(0)$ or $\pi_{ij}(\text{pa})$ should be used, assuming that the new cluster is sampled randomly.

This research discusses possible extensions of the standard validation measures described in Section 4.2 for use with each of the above three different approaches to prediction. The following sections discuss the calculation of these validation measures, starting with an overview of the approaches proposed to calculate the measures for clustered data.

4.3.3 Approaches for the calculation of the validation measures for clustered data

For clustered data, the naïve use of the existing validation measures for independent outcomes may lead to misleading conclusions regarding the model's predictive performance. The naïve approach assesses the effects of the fixed predictors only, and the predictive performance may change if clustering effects are considered in addition to the effects of the fixed predictors. Furthermore, assessing the model's performance within each cluster may be of interest, particularly to identify outlying clusters, where, for example, a cluster might represent a hospital.

Only limited research has been carried out to date to address these issues. One approach suggested by Oirbeek and Lesaffre [112] is an adaptation of the concordance measure for clustered survival outcomes (Harrell's C-index [8] Chapter 3). Their approach results in three concordance measures each with its own interpretation. In turn, these measures are based on a comparison of subjects: between clusters ('between cluster concordance' or Q_B); within clusters ('within cluster concordance' or Q_W); and both between and within clusters ('overall concordance' or Q_O). Q_O is calculated as a weighted sum of Q_B and Q_W, with weightings given by the proportion of between-and within-cluster usable pairs, denoted by π_B and π_W respectively. Between-cluster pairs consist of pair of subjects from different clusters only, whereas within-cluster pairs consist of pair of subjects from the same cluster only. The 'overall weighted concordance measure', Q_O, is given by:

$$Q_O = \pi_B Q_B + \pi_W Q_W, \tag{4.10}$$

where

$$\pi_B = \frac{N_{B,\text{usbl}}}{N_{T,\text{usbl}}}, \; \pi_W = \frac{N_{W,\text{usbl}}}{N_{T,\text{usbl}}},$$

$$Q_B = \frac{N_{B,\text{conc}}}{N_{B,\text{usbl}}}, \quad Q_W = \frac{1}{J}\sum_{j=1}^{J} Q_{W,j} = \frac{1}{J}\sum_{j=1}^{J} \frac{n_{W,\text{conc},j}}{n_{W,\text{usbl},j}}, \tag{4.11}$$

$N_{T,\text{usbl}}$ is the total number of usable pairs in the data, $N_{B,\text{usbl}}$ and $N_{W,\text{usbl}}$ are the number of between-and within-cluster usable pairs respectively, $N_{B,\text{conc}}$ and $N_{W,\text{conc}}$ are the number of between-and within-cluster concordance pairs respectively, and $Q_{W,j}$ is the 'within-cluster concordance measure' for cluster j. As discussed by Oirbeek and Lesaffre [112] and Chebon [87], the 'overall weighted measure' Q_O depends on the number and size of the clusters. Therefore, its value is difficult to compare across studies with different clustering designs. The 'within-cluster concordance measure' Q_W is the simple arithmetic mean of the cluster-specific concordance measure $Q_{W,j}$ and hence may be affected by the precision of the cluster-specific estimates of the measure.

This research proposes two approaches to calculate validation measures in the clustered data setting, which results in an 'overall' and a 'pooled cluster-specific' measure. In the 'overall' approach, one calculates the validation measure from a comparison of subjects within and between clusters, and the resulting measure assesses the overall predictive ability of the model. For example, the 'overall C-index' for clustered data can be calculated by comparing all possible pairs of subjects in the data, where subjects in a pair may come from the same cluster or from different clusters. Using the above notation, the 'overall C-index', C_O, can be written as:

$$C_O = \frac{N_{B,\text{conc}} + N_{W,\text{conc}}}{N_{T,\text{usbl}}}. \tag{4.12}$$

C_O has the same interpretation as the 'overall weighted measure' of Oirbeek and Lesaffre Q_O in that it assesses the overall discriminatory ability of the model. In the rest of the chapter, the notation C_O will be replaced by $C_{re(u)}$, $C_{re(0)}$, and C_{pa} based on the model predictions $\pi_{ij}(u)$, $\pi_{ij}(0)$, and $\pi(\text{pa})$, respectively.

In the 'pooled cluster-specific' approach, one calculates the validation measure for each cluster based on its original definition for standard logistic model along with a measure of precision. These measures are then pooled across clusters using the random-effects summary statistic method often used in meta analysis [113] (for more details, see Section 4.3.5). This approach yields a weighted average of the cluster-specific values, referred to as a 'pooled estimate'. The 'pooled cluster-specific' measure assesses the

predictive ability of the predictors whose values vary within clusters. For example, a 'pooled estimate' of the cluster-specific C-indices of 0.75 can be interpreted as that the ability of the model to discriminate between low-and high risk subjects is reasonable, given that the subject-pairs are drawn from the same cluster. This 'pooled' measure is similar to the 'within cluster measure' of Oirbeek and Lesaffre. However, unlike Oirbeek and Lesaffre's approach, this approach provides a weighted estimate, weighted by the precision of the cluster-specific estimates of the validation measure. Therefore, the 'pooled estimate' of the cluster-specific measures is less affected by clusters which produce extreme estimates.

The calculations of the validation measures for each of these approaches are discussed in the following sections.

4.3.4 Estimation: Overall measure

4.3.4.1 The C-index for clustered data: definition

Based on the model's three different approaches to prediction, three different definitions of the C-index can be obtained as follows. For a pair of subjects (i, k) from clusters (j, l) respectively, where i and k correspond to subject who had an event and those who did not respectively, with event probability $\{\pi_{ij}(u), \pi_{kl}(u)\}$, the concordance probability or C-index for the random-intercept logistic model can be defined as

$$C_{re(u)} = \Pr[\pi_{ij}(u) > \pi_{kl}(u)] \Leftrightarrow \Pr[\eta_{ij}(u) > \eta_{kl}(u)].$$

This applies to all possible pairs (i, k) in the data, where a pair may consist of subjects from the same cluster or from different clusters. If subjects are from different clusters, the cluster-specific random effect u values contribute in determining whether a pair is concordant and in $C_{re(u)}$, even if both subjects have the same predictor values. The random effects u however do not contribute in determining a concordant pair if both subjects are from the same cluster, as they share the same value of the random effect.

Based on the conditional event probabilities $\{\pi_{ij}(0), \pi_{kl}(0)\}$ (where the random

effects u are set to zero), the above probability become

$$C_{re(0)} = \Pr[\pi_{ij}(0) > \pi_{kl}(0)] \Leftrightarrow \Pr[\eta_{ij}(0) > \eta_{kl}(0)].$$

Similarly, based on population average probabilities $\{\pi_{ij}(\text{pa}), \pi_{kl}(\text{pa})\}$, the C-index can be defined as

$$C_{pa} = \Pr[\pi_{ij}(\text{pa}) > \pi_{kl}(\text{pa})] \Leftrightarrow \Pr[\eta_{ij}(\text{pa}) > \eta_{kl}(\text{pa})].$$

Note that $\pi_{ij}(\text{pa})$ is simply a transformed or re-scaled value of $\pi_{ij}(0)$, re-scaled by integrating out the random effect u in $\pi_{ij}(u)$ to obtain population-averaged probability. This has a one-to-one relationship with $\pi_{ij}(0)$, and hence the rank orders based on both $\pi_{ij}(\text{pa})$ and $\pi_{ij}(0)$ will be identical. Therefore, C_{pa} is equal to $C_{re(0)}$.

4.3.4.2 Nonparametric estimation of the C-index

Let $\eta_{ij}^{(1)}(u) = \eta_{ij}(u)|Y_{ij} = 1$ be the prognostic index for the ith subject with an event in the jth cluster, derived from $\pi_{ij}(u)$. Similarly, let $\eta_{kj}^{(0)}(u) = \eta_{kj}(u)|Y_{kj} = 0$ be the prognostic index for the kth subject without an event in the jth cluster. Let n_{1j} and n_{0j} be the number of subjects with an event and without an event respectively in the jth cluster. The total number of subjects with an event is $N_1 = \sum_j n_{1j}$, and the total number of subjects without an event is $N_0 = \sum_j n_{0j}$. Further, let J_1 and J_0 be the total number of clusters with at least one subject with an event and one without an event, respectively. Note that $J \leq (J_1 + J_0) \leq 2J$.

Extending equation (4.1), the non-parametric C-index for clustered binary outcomes can be defined as

$$C_{re(u)}^{np} = \frac{1}{N_1 N_0} \sum_{j=1}^{J} \sum_{l=1}^{J} \sum_{i=1}^{n_j} \sum_{k=1}^{n_l} I\Big(\eta_{ij}^{(1)}(u), \eta_{kl}^{(0)}(u)\Big), \tag{4.13}$$

where $I(.)$ can be defined similarly as in Section 4.2.3. $C_{re(u)}^{np}$ is analogous to the U statistic derived by Obuchowski [114] for clustered data. The use of the U statistic

in the context of clustered data has been further discussed in other studies; see, for example, Rosner and Grove [115], Lee and Rosner [116], and Lee and Dehling [117].

The C-index based on $\pi_{ij}(0)$ and $\pi_{ij}(\text{pa})$ can be obtained using the same approach to that described in equation (4.13) but by replacing $\eta_{ij}^{(1)}(u)$ and $\eta_{kl}^{(0)}(u)$ by the corresponding prognostic indices derived from $\pi_{ij}(0)$ and $\pi_{ij}(\text{pa})$. The resulting C-indices are denoted by $C_{re(0)}^{np}$ and C_{pa}^{np}, respectively. Since the rank orders based on $\pi_{ij}(\text{pa})$ and $\pi_{ij}(0)$ are identical, $C_{pa}^{np} = C_{re(0)}^{np}$.

The indices $C_{re(u)}^{np}$ and $C_{re(0)}^{np}$ are referred to as 'conditional indices', conditioned on the random effect u, and assess the predictive ability of predictor effects $\boldsymbol{\beta}$ and the random effects u, although $C_{re(0)}^{np}$ is based on the mean value of the random effects at zero. The C_{pa}^{np} does not include the contribution of the random effects u, assesses the predictive ability of the predictor effects $\boldsymbol{\beta}$ only, and has a marginal interpretation.

Note that $C_{re(u)}^{np} > C_{pa}^{np}$ if clustering exists in the data. If there is no clustering, $C_{re(u)}^{np} = C_{pa}^{np}$. This relationship is analogous to those derived by Oirbeek and Lesaffre [112] for a concordance measure for clustered survival data. The relationship could be explained using the following arguments. Let us consider a model with p predictors, where its discriminatory ability is quantified by C_{pa}^{np}. Let this model be extended by adding at least one predictor (hence $p+1$ predictors altogether in the new model) and the discriminatory ability of the extended model is quantified by $C_{re(u)}^{np}$. For example, if there is p fixed predictors in the model and an additional predictor represents the the effect of clustering then $C_{re(u)}^{np}$ is based on a model of $p+1$ predictors. If the additional predictor (that is, clustering) adds discriminative ability, then $C_{re(u)}^{np}$ based on the $p+1$ predictor model is greater than C_{pa}^{np} obtained from the p predictor model. If the additional predictor has no discriminative ability, both indices will be equal. In the random-intercept logistic model, the random effects u estimate the clustering effect, that is, the effect of unmeasured cluster level predictors that have not been included in the model. $C_{re(u)}^{np}$ is the result of combining both the random effects u and the predictor effects $\boldsymbol{\beta}$, whereas C_{pa}^{np} is the result of the predictor effects $\boldsymbol{\beta}$ only. This implies that

$C^{np}_{re(u)}$ is expected to be greater than or equal to C^{np}_{pa}, depending on whether clustering exists or not. Similarly, $C^{np}_{re(u)} > C^{np}_{re(0)}$ as $C^{np}_{pa} = C^{np}_{re(0)}$.

Confidence interval for $C^{np}_{re(u)}$

Several approaches have been proposed to estimate the variance of the area under ROC curve in the absence of clustering [45, 118, 119]. However, Rockette et al. [120] showed that all these approaches are approximately equivalent when sample size is large. Obuchowski [114] extended the method of DeLong et al. [118] for use with clustered data, following the concept of the design effect and effective sample size for the clustered design proposed by Rao and Scott [121]. In this research, the method of Obuchowski [114] is adapted to derive the variance expression for $C^{np}_{re(u)}$.

Let us define following two components as

$$V_1[\eta^{(1)}_{ij}(u)] = \frac{1}{N_0}\sum_{l=1}^{J_0}\sum_{k=1}^{n_l} I\Big(\eta^{(1)}_{ij}(u), \eta^{(0)}_{kl}(u)\Big) \tag{4.14}$$

for all $\eta^{(1)}_{ij}(u)$, and

$$V_0[\eta^{(0)}_{kl}(u)] = \frac{1}{N_1}\sum_{j=1}^{J_1}\sum_{i=1}^{n_j} I\Big(\eta^{(1)}_{ij}(u), \eta^{(0)}_{kl}(u)\Big) \tag{4.15}$$

for all $\eta^{(0)}_{kl}(u)$, where $V_1[\eta^{(1)}_{ij}(u)]$ is the proportion of subjects without an event who had prognostic index smaller than that of each subject with an event, and $V_0[\eta^{(0)}_{kl}(u)]$ is the proportion of subjects with an event who had prognostic indices larger than that of each subject without an event. It is obvious that $\sum_{j=1}^{J_1}\sum_{i=1}^{n_j} V_1[\eta^{(1)}_{ij}(u)]/N_1 = C^{np}_{re(u)}$ and similarly, $\sum_{l=1}^{J_0}\sum_{k=1}^{n_l} V_0[\eta^{(0)}_{kl}(u)]/N_0 = C^{np}_{re(u)}$.

Following Obuchowski [114] and Rao and Scott [121], the sum of squares of the proportions defined in equations (4.14)-(4.15) are computed as follows. Let $V_1[\eta^{(1)}_{.j}(u)]$ and $V_0[\eta^{(0)}_{.j}(u)]$ be the sums of the components defined in (4.14) and (4.15), respectively. Note that $V_1[\eta^{(1)}_{.j}(u)]$ is equal to zero if $n_{1j} = 0$, and similarly, $V_0[\eta^{(0)}_{.j}(u)]$ is equal to

zero if $n_{0j} = 0$. Using the notations of Obuchowski [114] and DeLong et al. [118], the sum of squares of the components in (4.14) and (4.15) can be defined as

$$S_1 = \frac{J_1}{(J_1 - 1)N_1} \sum_{j=1}^{J_1} \left[V_1[\eta_{.j}^{(1)}(u)] - n_{1j} \hat{C}_{re(u)}^{np} \right]^2 \tag{4.16}$$

and

$$S_0 = \frac{J_0}{(J_0 - 1)N_0} \sum_{j=1}^{J_0} \left[V_0[\eta_{.j}^{(0)}(u)] - n_{0j} \hat{C}_{re(u)}^{np} \right]^2, \tag{4.17}$$

respectively, where $n_{1j}\hat{C}_{re(u)}^{np}$ and $n_{0j}\hat{C}_{re(u)}^{np}$ are the mean sum of the components defined in (4.14) and (4.15), respectively. Further, let us define the following cross-product of these two components as

$$S_{10} = \frac{J}{(J-1)} = \sum_{j=1}^{J} \left[\{ V_1[\eta_{.j}^{(1)}(u)] - n_{1j} \hat{C}_{re(u)}^{np} \} \{ V_0[\eta_{.j}^{(0)}(u)] - n_{0j} \hat{C}_{re(u)}^{np} \} \right],$$

which takes into account the correlation between subjects with an event and those without an event within the same cluster [114]. Finally, the variance of $\hat{C}_{re(u)}^{np}$ can be estimated as

$$\widehat{\text{var}}[\hat{C}_{re(u)}^{np}] = \frac{1}{N_1} S_1 + \frac{1}{N_0} S_0 + \frac{2}{N_1 N_0} S_{10}. \tag{4.18}$$

As discussed by DeLong et al. [118], it can be shown by the central limit theorem that $\left(\hat{C}_{re(u)}^{np} - C_{re(u)}^{np}\right) / \sqrt{\widehat{\text{var}}[\hat{C}_{re(u)}^{np}]}$ is asymptotically $N(0,1)$ if $\lim_{J \to \infty} J_1/J_0$ is bounded and nonzero. The $(1-\alpha)\%$ confidence interval for $C_{re(u)}^{np}$ can be obtained as $\hat{C}_{re(u)}^{np} \pm Z_{\alpha/2} \sqrt{\widehat{\text{var}}[\hat{C}_{re(u)}^{np}]}$, where $Z_{\alpha/2}$ is the $\alpha/2$ percentile of standard normal distribution. The confidence interval for C_{pa}^{np} and $C_{re(0)}^{np}$ can be obtained using the same approach as described above.

4.3.4.3 Parametric estimation of the C-index

Similar to those for the standard logistic model, the parametric C-index for the random effects logistic model can be estimated under the assumption of normality of the prognostic index, $\eta_{ij}(u) = \boldsymbol{\beta}^T \boldsymbol{x}_{ij} + u_j$, for the population who had experienced the event and for those who did not.

For the fixed effects component of $\eta_{ij}(u)$, let us assume that $\boldsymbol{\beta}\boldsymbol{x}_{ij}|Y_{ij} = 1 \sim \mathrm{N}(\mu_1, \sigma^2)$ and $\boldsymbol{\beta}\boldsymbol{x}_{kl}|Y_{kl} = 0 \sim \mathrm{N}(\mu_0, \sigma^2)$. Since the random effects u_js are assumed to vary across clusters following a normal distribution with mean zero and variance σ_u^2, the outcome prevalence (number of events) is expected to vary across clusters. The greater the level of clustering greater the variation in the prevalence is expected across clusters. This could lead to a scenario where some of the clusters may appear with a prevalence close to 100% while others with a prevalence close to 0%. For simplicity, let us assume that there is one subject in a cluster. Further consider the notation u_{ij} instead of u_j and define $u_{ij}^{(1)}$ and $u_{kl}^{(0)}$ as the random effects for the cluster with a subject who had experienced the event and one who did not, respectively. If one sketches the distribution of $u_{kl}^{(0)}$ and $u_{ij}^{(1)}$, their location parameters are expected to shift to some extent from zero towards $-\infty$ and $+\infty$ respectively, depending on the level of clustering. Based on this premise, assume that $u_{ij}^{(1)} \sim \mathrm{N}(\gamma_1, \sigma_u^2)$ and $u_{kl}^{(0)} \sim \mathrm{N}(\gamma_0, \sigma_u^2)$.

Therefore,

$$\eta_{ij}^{(1)}(u)|Y_{ij} = 1 \sim \mathrm{N}(\mu_1 + \gamma_1, \sigma^2 + \sigma_u^2)$$

and

$$\eta_{kl}^{(0)}(u)|Y_{kl} = 0 \sim \mathrm{N}(\mu_0 + \gamma_0, \sigma^2 + \sigma_u^2).$$

The C-index based on $\pi_{ij}(u)$ can be defined as

$$\begin{aligned} C_{re(u)}^{p} &= \Pr[\eta_{ij}^{(1)}(u) > \eta_{kl}^{(0)}(u)] \\ &= \Pr[(\eta_{ij}^{(1)}(u) - \eta_{kl}^{(0)}(u)) > 0]. \end{aligned} \tag{4.19}$$

After standardising the term $\eta_{ij}^{(1)}(u) - \eta_{kl}^{(0)}(u)$, $C_{re(u)}^{p}$ can be obtained as

$$\begin{aligned} C_{re(u)}^{p} &= \Pr\Big[Z < \frac{(\mu_1 + \gamma_1) - (\mu_0 + \gamma_0)}{\sqrt{2\sigma^2 + 2\sigma_u^2}}\Big], \quad Z \sim \mathrm{N}(0,1) \\ &= \Phi\left(\frac{(\mu_1 - \mu_0) + (\gamma_1 - \gamma_0)}{\sqrt{2\sigma^2 + 2\sigma_u^2}}\right), \end{aligned}$$

where Φ denotes the standard normal cumulative distribution function. Replacing the parameters $(\mu_1, \mu_0, \gamma_1, \gamma_0, \sigma^2, \sigma_u^2)$ by the corresponding sample estimates $(\bar{x}_1, \bar{x}_0, \bar{u}_1, \bar{u}_0, S^2, \hat{\sigma}_u^2)$, $C_{re(u)}^{p}$ can be estimated as

$$\hat{C}_{re(u)}^{p} = \Phi\left(\frac{\bar{x}_1 - \bar{x}_0 + \bar{u}_1 - \bar{u}_0}{\sqrt{2S^2 + 2\hat{\sigma}_u^2}}\right), \tag{4.20}$$

The indices $C_{re(0)}^{p}$ and C_{pa}^{p} for $\pi_{ij}(0)$ and $\pi_{ij}(\mathrm{pa})$ respectively can be derived using a similar approach to that discussed above, but replacing $\eta_{ij}(u)$ by the corresponding prognostic indices $\eta_{ij}(0)$ and $\eta_{ij}(\mathrm{pa})$. All these versions of parametric C-indices have the same interpretation to those with the non-parametric indices.

Confidence interval for $C_{re(u)}^{p}$

Let us define $\hat{\delta} = \dfrac{\bar{x}_1 - \bar{x}_0 + \bar{u}_1 - \bar{u}_0}{\sqrt{2S^2 + 2\hat{\sigma}_u^2}}$ so that $\hat{C}_{re(u)}^{p} = \Phi(\hat{\delta})$. Since Φ is a monotonically increasing function of $\hat{\delta}$, finding the variance for $\hat{C}_{re(u)}^{p}$ using the Delta method [122, 123] is equivalent to finding one for $\hat{\delta}$ [124].

According to the properties of normal distribution, $\bar{x}_1$, $\bar{x}_0$, $\bar{u}_1$, and $\bar{u}_0$ are independent normal random variables with means and variances μ_1 and σ^2/N, μ_0 and σ^2/N, γ_1 and σ_u^2/J, and γ_0 and σ_u^2/J, respectively. Therefore,

$$\hat{\mu} = (\bar{x}_1 - \bar{x}_0) + (\bar{u}_1 - \bar{u}_0) \sim \mathrm{N}\Big(\mu_1 - \mu_0 + \gamma_1 - \gamma_0, \frac{2\sigma^2}{N} + \frac{2\sigma_u^2}{J}\Big), \tag{4.21}$$

and

$$\frac{(N-1)S^2}{\sigma^2} \sim \chi^2_{N-1} \qquad \text{and} \qquad \frac{(J-1)\hat{\sigma}_u^2}{\sigma_u^2} \sim \chi^2_{J-1} \tag{4.22}$$

are mutually independent. Let $\hat{\sigma}_p^2 = 2S^2 + 2\hat{\sigma}_u^2$ so that $\hat{\delta} = \dfrac{\hat{\mu}}{\hat{\sigma}_p}$. Assuming $\hat{\sigma}_p^2$ and $\hat{\mu}$ to be independent, the Delta method yields the following approximate variance expression for $\hat{\delta}$:

$$\text{var}(\hat{\delta}) \approx \left(\frac{\partial\hat{\delta}}{\partial\hat{\mu}}\right)^2 \text{var}(\hat{\mu}) + \left(\frac{\partial\hat{\delta}}{\partial\hat{\sigma}_p}\right)^2 \text{var}(\hat{\sigma}_p) = \frac{1}{\hat{\sigma}_p^2}\text{var}(\hat{\mu}) + \frac{\hat{\mu}^2}{\hat{\sigma}_p^4}\text{var}(\hat{\sigma}_p). \tag{4.23}$$

$\text{Var}(\hat{\mu})$ is given in equation (4.21), whereas the Delta method is applied again to obtain $\text{var}(\hat{\sigma}_p)$ as:

$$\begin{aligned}
\text{var}(\hat{\sigma}_p) &= \text{var}(\hat{\sigma}_p^2)^{\frac{1}{2}} \approx \left(\frac{\partial(\hat{\sigma}_p^2)^{\frac{1}{2}}}{\partial\hat{\sigma}_p^2}\right)^2 \text{var}(\hat{\sigma}_p^2) = \frac{1}{4\hat{\sigma}_p^2}\text{var}(\hat{\sigma}_p^2) \\
&= \frac{1}{4\hat{\sigma}_p^2}\Big[4\text{var}(S^2) + 4\text{var}(\hat{\sigma}_u^2)\Big] \\
&= \frac{1}{4\hat{\sigma}_p^2}\Big[\frac{8(\sigma^2)^2}{N-1} + \frac{8(\sigma_u^2)^2}{J-1}\Big]; \ \text{[using equation (4.22)]}.
\end{aligned} \tag{4.24}$$

Using equation (4.21) and (4.24) in equation (4.23) yields,

$$\text{var}(\hat{\delta}) \approx \frac{1}{\hat{\sigma}_p^2}\Big[\frac{2\sigma^2}{N} + \frac{2\sigma_u^2}{J}\Big] + \frac{\hat{\mu}^2}{4(\hat{\sigma}_p^2)^3}\Big[\frac{8(\sigma^2)^2}{N-1} + \frac{8(\sigma_u^2)^2}{J-1}\Big]. \tag{4.25}$$

Substituting the estimates for the unknown parameters in equation (4.25) results in

$$\begin{aligned}
\widehat{\text{var}}(\hat{\delta}) &\approx \Big[\frac{2S^2}{N} + \frac{2\hat{\sigma}_u^2}{J}\Big](2S^2 + 2\hat{\sigma}_u^2)^{-1} \\
&+ \frac{(\bar{x}_1 - \bar{x}_0 + \bar{u}_1 - \bar{u}_0)^2}{4(2S^2 + 2\hat{\sigma}_u^2)^3}\Big[\frac{8S^4}{N-1} + \frac{8\hat{\sigma}_u^4}{J-1}\Big].
\end{aligned} \tag{4.26}$$

The $(1-\alpha)\%$ CI for $\hat{C}^{p}_{re(u)}$ is then given by

$$\Phi\Big(\hat{\delta} \pm Z_{\alpha/2}\sqrt{\widehat{\text{var}}(\hat{\delta})}\Big), \tag{4.27}$$

where $Z_{\alpha/2}$ is the $\alpha/2$ percentile of the standard normal distribution. The confidence interval for $C^{p}_{re(0)}$ and C^{p}_{pa} can be obtained using a similar approach to that discussed above.

4.3.4.4 D statistic

The D statistic for the random effects logistic model can be obtained by transforming the prognostic index $\hat{\eta}_{ij}(u)$ to z_{ij} using the same approach as described for the standard Cox model and then fitting a random-intercept logistic model to the validation sample with z_{ij} as the only predictor. The model takes the following form:

$$\text{logit}(Y_{ij}=1|u_j, z_{ij}) = \beta_z z_{ij} + u_j, \tag{4.28}$$

where $\hat{D}_{re(u)}$ is equal to the coefficient of z and the standard error of $\hat{D}_{re(u)}$ is equal to standard error of $\hat{\beta}_z$. It is also equivalent to obtain $\hat{D}_{re(u)}$ by fitting a standard logistic model with z_{ij} as the only predictor, because the random effects are already included in z_{ij}.

For $\pi_{ij}(0)$ and $\pi_{ij}(\text{pa})$, $\hat{D}_{re(0)}$ and $\hat{D}_{pa}$ respectively can be obtained in a similar manner to that described above by transforming the corresponding prognostic index to z_{ij}. All these versions of D statistic have the same interpretation to those for C-index.

4.3.4.5 Calibration slope

The calibration slope (CS) for clustered binary outcomes can be obtained using the same way to the standard logistic model but by fitting a random-intercept logistic model with the prognostic index $\hat{\eta}_{ij}(u)$, derived from $\pi_{ij}(u)$, as the only predictor. The

resulting model takes the following form:

$$\text{logit}(Y_{ij} = 1|u_j, \hat{\eta}_{ij}(u)) = \beta_u \hat{\eta}_{ij}(u) + u_j. \tag{4.29}$$

The estimated calibration slope, $\widehat{CS}_{re(u)}$, is equal to the estimate of β_u. Similar to $\hat{D}_{re(u)}$, one can also obtain $\widehat{CS}_{re(u)}$ by fitting a standard logistic model.

The calibration slope $\widehat{CS}_{re(0)}$ and $\widehat{CS}_{pa}$ based on $\pi_{ij}(0)$ and $\pi_{ij}(\text{pa})$ respectively can be obtained using the same approach to that discussed above but by replacing $\hat{\eta}_{ij}(u)$ by the corresponding prognostic indices. All these versions of calibration slope have the same interpretation to the standard calibration slope, based on the reference value of one (see, Section 4.2.7).

4.3.4.6 Brier score

The Brier score (BS) for the random-intercept logistic model can be obtained by averaging the squared differences between the predicted probabilities $\pi_{ij}(u)$ and the observed outcomes y. Extending equation (4.7), the Brier score for $\pi_{ij}(u)$ can be obtained as

$$BS_{re(u)} = \frac{1}{N} \sum_{j=1}^{J} \sum_{i=1}^{n_j} \Big(y_{ij} - \hat{\pi}_{ij}(u) \Big)^2. \tag{4.30}$$

Similarly, for $\pi_{ij}(0)$ and $\pi_{ij}(\text{pa})$, the Brier score can be obtained by replacing $\hat{\pi}_{ij}(u)$ by their corresponding predicted probabilities $\hat{\pi}_{ij}(0)$ and $\hat{\pi}_{ij}(\text{pa})$, respectively. The resulting Brier scores are denoted by $BS_{re(0)}$ and BS_{pa}, respectively. Unlike the same versions of the rank-based validation measures, $BS_{re(0)} \neq BS_{pa}$ as $\pi(0) \neq \pi(\text{pa})$. In addition, it can be shown that $BS_{re(u)} \leq BS_{pa}$ using the same explanation as discussed for showing that $C^{np}_{re(u)} \geq C^{np}_{pa}$, keeping in mind that the Brier score has an inverse relationship with the C-index. For example, the Brier score for a model with p predictors can decrease to some extent towards its minimum value of zero with the inclusion of a predictor that adds predictive strength in the model, whereas the C-index can increase to some extent towards it maximum value of one due to a similar inclusion.

4.3.5 Estimation: Pooled cluster-specific measure

The 'pooled cluster-specific' measure involves estimation of validation measures for each cluster and then pooling of these across clusters to obtain a weighted average. The weights can be calculated based on the inverse of both the within cluster-and between-cluster variances of the cluster-specific validation measures. The within cluster variance is simply the estimated variance of the cluster-specific estimates of the validation measures. For the between cluster variance, several estimation techniques have been proposed in the literature of meta-analysis including the method of moments [113] and maximum likelihood [125, 126]. In this thesis, the method of moment has been used to estimate the between cluster variance, because of its simplicity. The estimated between cluster variance is incorporated in the calculation of the pooled estimate of the cluster-specific validation measures to take into account for the heterogeneity between the clusters. This approach is commonly used in meta analysis to combine the results of several studies. The detailed calculation of the pooled estimate of the cluster-specific validation measures is described as follow.

Let $\hat{\theta}_j$ $(j = 1, \ldots, J)$ be the estimate of a validation measure for the jth cluster, and $\hat{\sigma}_j^2$ be the corresponding estimated variance. The weighted average (pooled estimate) of the cluster specific estimates can be calculated as

$$\hat{\theta}_w = \bar{w}^{-1} \sum_{j=1}^{J} \hat{\theta}_j \hat{w}_j, \tag{4.31}$$

where $\hat{w}_j = 1/(\hat{\sigma}_j^2 + \hat{\tau}^2)$, $\bar{w} = \sum_{j=1}^{J} \hat{w}_j$, and $\hat{\tau}^2$ is the estimate of the between cluster variance and can be obtained as

$$\hat{\tau}^2 = \max\left\{0, \frac{\left[\sum_{j=1}^{J} \hat{a}_j(\hat{\theta}_j - \bar{\theta})^2\right] - (J-1)}{\sum_{j=1}^{J} \hat{a}_j - \sum_{j=1}^{J} \hat{a}_j^2 / \sum_{j=1}^{J} \hat{a}_j}\right\},$$

where $\hat{a}_j = 1/\hat{\sigma}_j^2$ and $\bar{\theta} = \sum_{j=1}^{J} \hat{a}_j \hat{\theta}_j / \sum_{j=1}^{J} \hat{a}_j$.

Assuming that the clusters are sufficiently large and there is at least a moderate

number of clusters, confidence intervals can be obtained by using the following approximation:

$$\hat{\theta}_w \sim \mathrm{N}\Big(\theta_w, 1/\textstyle\sum_{j=1}^{J} w_j\Big).$$

The $100(1-\alpha)\%$ confidence intervals for θ_w can be obtained as

$$\hat{\theta}_w \pm Z_{\alpha/2}\Big(\sum_{j=1}^{J} \hat{w}_j\Big)^{-1/2}, \tag{4.32}$$

where $Z_{\alpha/2}$ is the $\alpha/2$ percentile of the standard normal distribution.

Using the above approach, the 'pooled estimate' of each of the validation measures can be obtained. Similar to the 'overall measure', three different definitions for each of the 'pooled cluster-specific' measures can be obtained based on the model predictions $\pi_{ij}(u)$, $\pi_{ij}(0)$, and $\pi_{ij}(\text{pa})$. The resulting nonparametric C-indices, for example, are denoted by $C^{np}_{w,re(u)}$, $C^{np}_{w,re(0)}$, and $C^{np}_{w,pa}$, respectively. However, $C^{np}_{w,re(u)}$=$C^{np}_{w,re(0)}$=$C^{np}_{w,pa}$=C^{np}_{w} (say). This is because the C-index is a rank-based statistic, and that the rank orders between the subjects within a cluster for these three types of prediction are identical as subjects from the same cluster share the same random effect u. This argument holds for the parametric C-index and also for any other rank-based statistic, for example, the D statistic. The resulting parametric C-index and D statistic are denoted by C^{p}_{w} and D_w, respectively.

Although the calibration slope is not a rank-based statistic, the 'pooled estimate' of the cluster-specific calibration slopes for all the three approaches to prediction are also equal. The reason is as follows. Among these approaches, only $\pi_{ij}(u)$ uses the random effect u values. When calculating the calibration slope for a cluster j by fitting a standard logistic model, $\hat{u}_j$ for that cluster is treated as a constant as all subjects in a cluster have the same $\hat{u}_j$. Therefore, the slope (calibration slope) of that logistic model is not affected by $\hat{u}_j$, except for the intercept which is essentially equal to $\hat{u}_j$. Therefore, the 'pooled cluster-specific' calibration slope, say CS_w, for $\pi_{ij}(u)$ is equal to those for $\pi_{ij}(0)$ and $\pi_{ij}(\text{pa})$.

Similarly, the 'pooled estimate' of the cluster-specific Brier score can be obtained for each of these predictions $\pi_{ij}(u)$, $\pi_{ij}(0)$, and $\pi_{ij}(\text{pa})$. The resulting measures are denoted by $BS_{w,re(u)}$, $BS_{w,re(0)}$, and $BS_{w,pa}$, respectively. Unlike the other measures, these are not equal. This is because the Brier score quantifies the accuracy of individual predictions, but the predictions from these three approaches are not equal. However, these three 'pooled' Brier scores have their own interpretation based on $\pi_{ij}(u)$, $\pi_{ij}(0)$, and $\pi_{ij}(\text{pa})$. Note that the analytical expression for the variance of the Brier score is not available, and therefore bootstrap-based standard errors can be used to obtain the 'pooled estimate' of the cluster-specific Brier score.

4.4 Application to clustered binary data

In this section, an application of the above methods is illustrated using a real dataset of patients undergoing heart valve surgery at different hospitals in the UK. The section starts with a description of the data, which is followed by the analysis and results.

4.4.1 Heart valve surgery data

This dataset was based on patients who underwent aortic and/or mitral heart valve surgery at 30 different hospitals in the UK. The clinical outcome of interest was in-hospital mortality (alive/dead). The dataset consists of 32,839 patients, with a total of 2,089 (6.3 percent) in-hospital deaths. The predictors of interest were age, gender, body mass index (BMI), hypertension (no/yes), diabetes (no/yes), renal failure (none or functioning transplant/ creatinine > 200 μmoL/ dialysis dependency), concomitant CABG surgery (no/yes), concomitant tricuspid surgery (no/yes), preoperative arrhythmias (no/atrial fibrillation or heart block/ventricular tachycardia or fibrillation), ejection fraction (<30%/30%-50%/>50%), operative priority (elective/urgent/emergency), operation sequence (previous sternotomy; first/second/third or more), and the year of surgery. The median cluster size was 1517 with an interquartile range (IQR): 1168 to 2098. The intra-cluster correlation (ICC) calculated using the method of analysis of variance (ANOVA) [127, 128] was 0.06. The risk model based on this dataset has already been developed by Ambler et al. [1]. The main focus here is to illustrate the validation measures for clustered binary data.

4.4.2 Analysis and results

4.4.2.1 Model development

The dataset was split into two parts: one part was used to develop the model and the other to validate the model. The development data included all patients who underwent surgery during the first five years, and a temporal validation was conducted by including patients who underwent surgery in the subsequent three years. In this validation exercise, both the development and validation datasets consisted of the same hospitals but different patients. A prognostic model was developed based on the random-intercept logistic regression model with normally distributed random effects and all available predictors. Maximum likelihood estimation of the model parameters was performed using adaptive Gaussian quadrature [97, 99] with 20 quadrature points per level. The `gllamm` package in Stata version 11 [129] was used to fit the model. Inspection of the residual plots suggested that the assumption of normality regarding the random effects was reasonably satisfied.

The estimated model parameters are not reported here, except for the variance parameter of the random effects, σ_u^2, which was estimated as 0.18. This corresponds to an ICC $= \sigma_u^2/(\sigma_u^2 + \pi^2/3) = 0.05$, indicating weak correlation between patients within a hospital, after accounting for the fixed predictors.

4.4.2.2 Model validation

The model was used to predict the probability of in-hospital mortality using three different approaches $\pi_{ij}(u)$, $\pi_{ij}(0)$, and $\pi_{ij}(\text{pa})$ in the validation data. These predicted probabilities are plotted in Figure 4.1 to observe the spread in predictions and to see whether there are any differences between them. All three approaches showed reasonable spread in predictions, with relatively high proportion of patients predicted to have a low risk of in-hospital mortality and low proportion of patients predicted as high risk. This reflects the observed risk, that is, about 6 percent of patients had experienced in-hospital mortality following a heart valve surgery. The spread in predictions for all the three approaches were similar; however, slightly a greater spread was observed for predictions based on $\pi_{ij}(u)$.

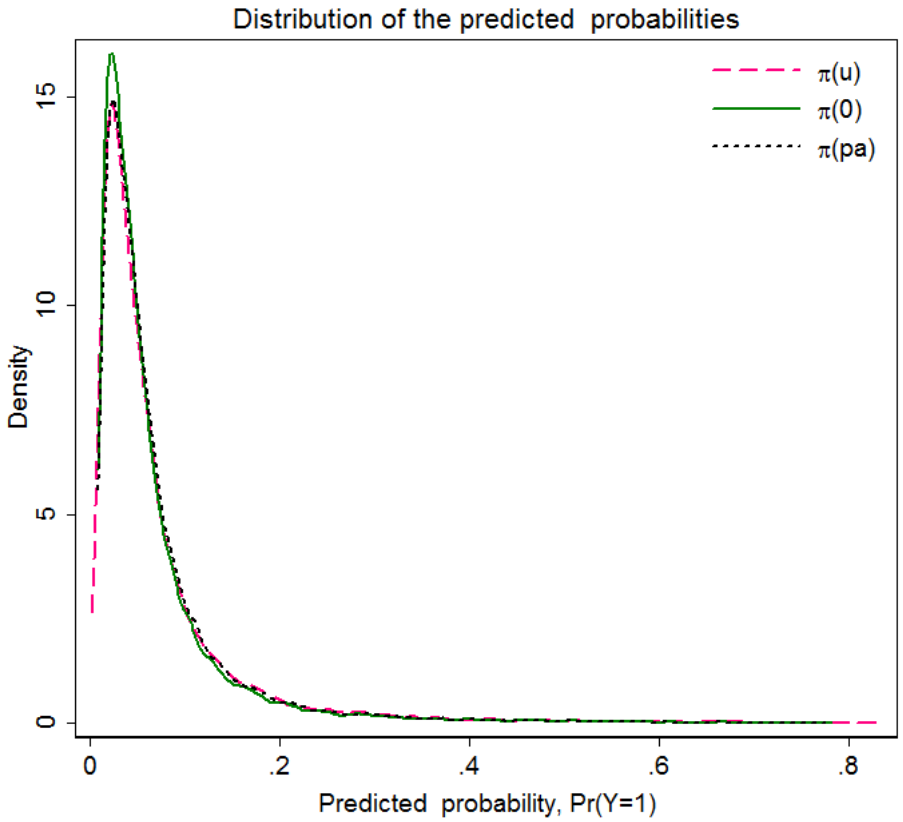

Figure 4.1: Distribution of the predicted probability, $\Pr(Y = 1)$, by types of prediction such as $\pi_{ij}(u)$, $\pi_{ij}(0)$, and $\pi_{ij}(\text{pa})$.

The predictive performance of the model in the validation data was evaluated by using the validation measures described in Section 4.3. To calculate the validation measures, the prognostic index $\eta_{ij}(u)$, $\eta_{ij}(0)$, and $\eta_{ij}(\text{pa})$ based on the model's three different approaches to prediction were derived in the validation data. Some of the validation measures, for example, the D statistic and the parametric C-index are based on the assumption of normality of the prognostic index (PI). Therefore, the distributions of the predicted PI for the patients who survived and those who died are presented graphically in Figure 4.2, by types of prediction. It appears that the distributions of the PI for the two groups of patients are approximately normal, which holds for all types of prediction. Furthermore, there is a reasonable discrimination (or separation) between these two groups of patients. The discriminatory ability for $\pi_{ij}(u)$ appeared to

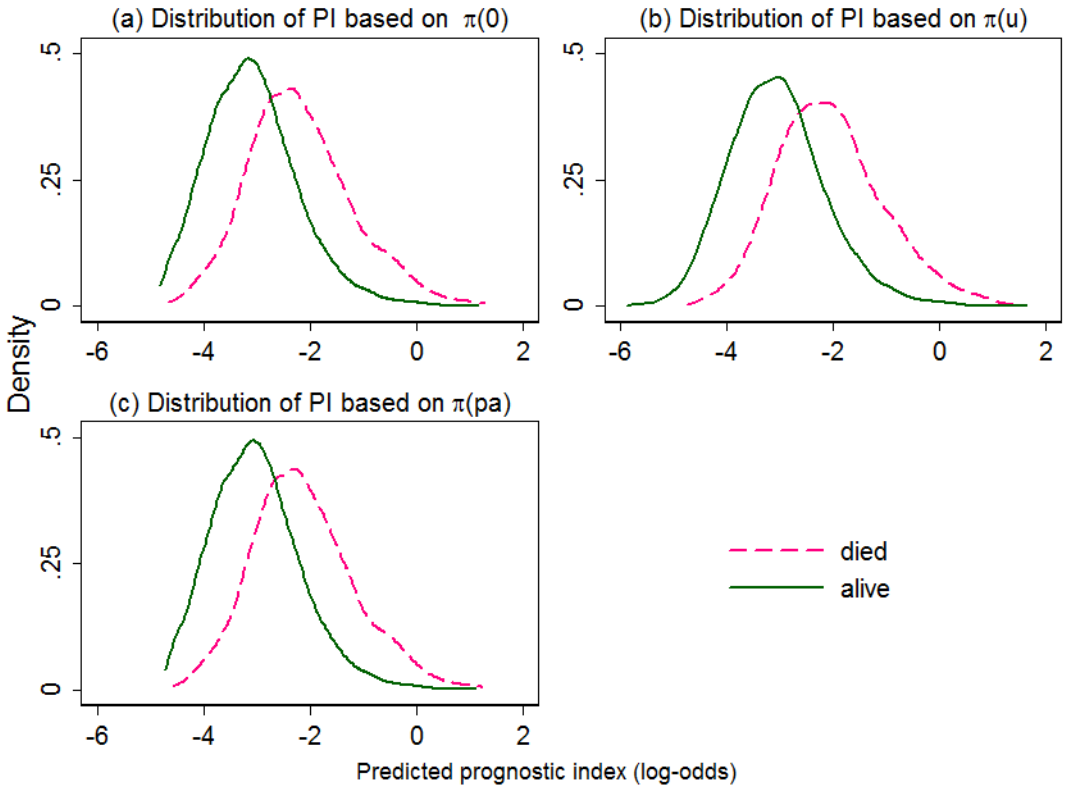

Figure 4.2: Distribution of the predicted prognostic index (PI) or log odds for the population who survived and those who died by types of prediction: (a) $\pi_{ij}(0)$, (b) $\pi_{ij}(u)$, and (c) $\pi_{ij}(\text{pa})$.

be approximately equal to those for $\pi_{ij}(0)$ and $\pi_{ij}(\text{pa})$. This is because the clustering effect in these data is not strong.

Calculation of the validation measures was performed using user written Stata code (Appendix B: Figure B.2), and the results are presented in Table 4.1. The 'overall estimates' for all types of non-parametric C-index $C^{np}_{re(u)}$, $C^{np}_{re(0)}$, and C^{np}_{pa} suggest reasonably good discrimination between the high and low risk patients. The point estimate $C^{np}_{re(u)}$ was slightly greater than that of $C^{np}_{re(0)}$ and C^{np}_{pa}, although the 95% CIs of the indices overlap each other. This is because the effect of clustering in these data was weak. In addition, the estimates of C^{np}_{pa} and $C^{np}_{re(0)}$ were equal, indicating identical

discrimination for both $\hat{\pi}_{ij}(0)$ and $\hat{\pi}_{ij}(\text{pa})$. Similar findings were observed for the D statistics and the parametric C-indices. The non-parametric C-index was also calculated based on Oirbeek and Lesaffre's Q_O approach. The estimate was 0.784, which is very close to that obtained for the analogous version $C^{np}_{re(u)}$.

Table 4.1: Estimates of the validation measures for the model predicting in-hospital mortality following heart valve surgery in the validation sample.

		Overall Measures	
Standard measures	Adapted measures	Estimates	95% CIs
Non Parametric C-index	$C^{np}_{re(u)}$	0.785	[0.776, 0.793]
	$C^{np}_{re(0)}$	0.774	[0.759, 0.789]
	C^{np}_{pa}	0.774	[0.759, 0.789]
Parametric C-index	$C^{p}_{re(u)}$	0.785	[0.775, 0.794]
	$C^{p}_{re(0)}$	0.775	[0.758, 0.790]
	C^{p}_{pa}	0.775	[0.758, 0.790]
D statistic	$D_{re(u)}$	1.85	[1.78, 1.92]
	$D_{re(0)}$	1.76	[1.63, 1.87]
	D_{pa}	1.76	[1.63, 1.87]
Calibration slope	$CS_{re(u)}$	1.01	[0.94, 1.08]
	$CS_{re(0)}$	0.98	[0.91, 1.06]
	CS_{pa}	0.99	[0.93, 1.07]
Brier score	$BS_{re(u)}$	0.049	-
	$BS_{re(0)}$	0.052	-
	BS_{pa}	0.051	-
		Pooled Measures	
Non-parametric C-index	C^{np}_{w}	0.775	[0.757, 0.791]
Parametric C-index	C^{p}_{w}	0.774	[0.756, 0.790]
D statistic	D_w	1.77	[1.63, 1.89]
Calibration slope	CS_w	0.99	[0.92, 1.07]
Brier score	$BS_{w,re(u)}$	0.051	[0.046, 0.056]
	$BS_{w,re(0)}$	0.053	[0.047, 0.059]
	$BS_{w,pa}$	0.052	[0.046, 0.058]

The 'overall' calibration slope $CS_{re(u)}$ was estimated to be 1.01 (95% CI : 0.94 to 1.08), which suggests that overall calibration for $\hat{\pi}_{ij}(u)$ was reasonably good. Similar results were observed for $\hat{\pi}_{ij}(0)$ and $\hat{\pi}_{ij}(\text{pa})$. The estimates of $BS_{re(u)}$, $BS_{re(0)}$, and BS_{pa} suggest that all the approaches showed reasonably good accuracy in predicting

in-hospital mortality. The estimate of $BS_{re(u)}$ for $\hat{\pi}_{ij}(u)$ was slightly smaller than those for $\hat{\pi}_{ij}(\text{pa})$ and $\hat{\pi}_{ij}(0)$, again suggesting weak clustering in these data.

The 'pooled estimates' of the cluster-specific measures are also presented in Table 4.2. The estimates of both the parametric and non-parametric C-indices $\hat{C}_w^{np}$ and $\hat{C}_w^{p}$ suggest that the model has reasonable ability to discriminate between patients who died in the hospital and those who survived, given that both patients in the pair considered in the calculation belong to the same hospital. A similar result was observed for D_w. The non-parametric C-index based on Oirbeek and Lesaffre's Q_W approach was 0.773, which is similar to $\hat{C}_w^{np}$. The 'pooled' calibration slope CS_w was estimated to be 0.99 (95% CI: 0.92 to 1.07), which indicates that the model has good calibration when predicting within a cluster. The 'pooled estimates' of the Brier scores suggest that the prediction error of $\hat{\pi}_{ij}(u)$, $\hat{\pi}_{ij}(0)$, and $\hat{\pi}_{ij}(\text{pa})$ were reasonably low. As with the 'overall estimates', the 'pooled estimate' of the cluster-specific Brier score based on $\hat{\pi}_{ij}(u)$ is slightly smaller than those based on $\hat{\pi}_{ij}(0)$ and $\hat{\pi}_{ij}(\text{pa})$.

The 'pooled cluster-specific' approach based on the random effects summary statistic method provided the method-of-moments estimates of the between-cluster variances of the cluster-specific measures, τ^2, as 0.003, 0.036, 0.001, 0.001 for C_w^{np}, D_w, CS_w, and BS_w, respectively. To examine whether the method-of-moments provided comparable results with other available methods, τ^2 was also estimated using the method of maximum likelihood (ML) and restricted maximum likelihood (REML). Both approaches showed results similar to that obtained from the method-of-moments. Furthermore, the random effects summary statistic method is usually preferred to a fixed effect method as it may be considered to encompass the fixed effects method when τ^2 zero.

The cluster (hospitals)-specific estimates (with their 95% CIs) of the validation measures are plotted in Figure 4.3. Each of the four plots shows the rank order of the hospitals based on the hospital-specific estimates of the validation measures. The horizontal solid line based on the 'pooled estimate' represents the average performance of the model within a hospital. The plots show the results of 25 hospitals, because the model could not be applied to 5 of the hospitals as they did not contribute to

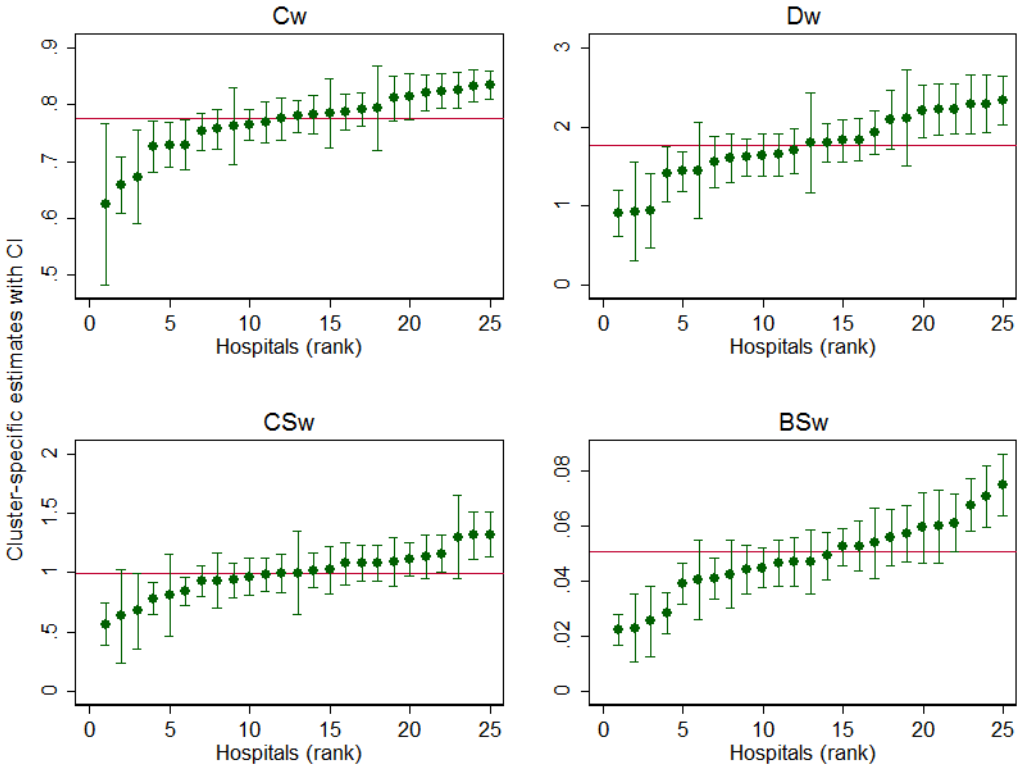

Figure 4.3: Cluster (hospital)-specific estimates against their rank order: the C-index (non-parametric), D statistic, calibration slope, and Brier score. Each horizontal solid line indicates the 'pooled estimate' of the respective measures.

the validation data due to lack of events. This type of plot may be used to make a comparison between hospitals and to identify hospitals where model performance is good or poor, relative to the averaged performance. This type of comparison may also shed some light on monitoring hospital performances. One could also compare the observed and predicted deaths to evaluate hospital performance [130].

It can be seen in Figure 4.3 that the predictive ability of the model for some of the hospitals were significantly worse (better) than the pooled averaged as the points estimates of the validation measures, except the calibration slope, for these hospitals were smaller (greater) than the 'pooled estimate' and the 95% CIs did not include the average value. The point estimates of the calibration slope for some of the hospitals somewhat different from 1 and 95% CI did not include this value, which indicates

poor calibration of the model for these hospitals. The heterogeneity in the model performance between hospitals may be caused by the unobserved patient or hospital level characteristics. This may also suggest a mis-specification of the model for these hospitals. Therefore, it would be important to investigate the factors which explain this heterogeneity.

One issue that may be raised before making a comparison between hospitals based on the hospital-specific estimates of the validation measures is to examine whether these estimates are associated with the hospital sizes or hospital-specific prevalence (mortality rate). It appears in Figure 4.3 that the estimates of the validation measures for some of the hospitals with narrow CIs, which indicate some of the larger hospitals or hospitals with higher prevalence, are still below the averaged line (horizontal solid line). Furthermore, a scatter plot between the hospital-specific estimates of the validation measures and the prevalence did not suggest an association between these two (results not shown).

In summary, this illustration has showed that the 'overall' and 'pooled' estimates of the validation measures have meaningful interpretations when assessing the predictive ability of a model for clustered binary outcomes. In the next section, performance of these validation measures for clustered data are evaluated using simulation studies.

4.5 Simulation study

In this section, the properties of both the point estimates and confidence intervals of the validation measures such as bias, root Mean Squared Error (rMSE), and coverage were investigated by simulation studies. Both development and validation data were simulated from a true model. Prognostic models were developed using the simulated development data and then evaluated using the corresponding simulated validation data. The properties of the validation measures were investigated in a range of scenarios, created by varying the number of clusters and their size and the intra-cluster correlation coefficient (ICC) between subjects within the same cluster in the validation data, to see how these measures perform across these scenarios. The aim was to identify scenarios

where the validation measures did not perform adequately, for example, whether the validation measures were affected by number of clusters, cluster size, and the level of clustering. The section begins by describing the simulation design and is followed by describing the strategies for evaluating the measures and the results.

4.5.1 Simulation design

4.5.1.1 True model

Clustered binary data were generated from a true model based on the random-intercept logistic model with normally distributed random effects and one fixed predictor that has a fixed effect. One of the aims was to generate data under different values of ICC, to mimic scenarios with no, moderate, and high levels of clustering. Accordingly the subject level variability (represented by the fixed predictor) was varied and the total predictive variability that combines the fixed and random effects to represent both the subject and cluster level characteristics has been fixed to a specific value over the different ICC scenarios. For a sample of size N with J clusters, the predictor value x_{ij} for the ith subject in the jth cluster ($i = 1, \ldots, n_j; j = 1, \ldots, J$) was generated from $\mathrm{N}(0, 1)$, and the true random effects u_j were from $\mathrm{N}(0, \sigma_u^2)$. Then the outcomes y_{ij} were generated from the Bernuolli distribution with probability calculated from the true random-intercept logistic model using

$$\pi(\beta_0, \beta_1 | x_{ij}, u_j) = \frac{\exp[\eta(\beta_0, \beta_1, x_{ij}, u_j)]}{1 + \exp[\eta(\beta_0, \beta_1, x_{ij}, u_j)]}. \tag{4.33}$$

where $\eta(\beta_0, \beta_1, x_{ij}, u_j) = \beta_0 + \beta_1 x_{ij} + u_j$ is the true prognostic index with intercept β_0 and slope β_1. As $X \sim \mathrm{N}(0, 1)$, $\beta_1 X \sim \mathrm{N}(0, \beta_1^2)$, and therefore $\eta(\beta_0, \beta_1, x_{ij}, u_j)$ follows $\mathrm{N}(\beta_0, \beta_1^2 + \sigma_u^2)$, assuming one subject per cluster. Note that $\beta_1^2 + \sigma_u^2$ represent the total predictive variability in the log-odds of having the event, which can be decomposed into subject level variability (β_1^2) and cluster level variability (σ_u^2). Then the intra-cluster correlation (ICC) between subjects within a cluster can be specified as $\sigma_u^2/(\beta_1^2 + \sigma_u^2)$, where relatively high values of σ_u^2 indicate high ICC.

To simulate data under different ICC scenarios, the values of σ_u^2 were varied keeping

the total predictive variability fixed to 1.4^2, and β_1 was determined from $\beta_1^2+\sigma_u^2 = 1.4^2$. The choice of the value of the total predictive ability is arbitrary, but the aim was to assess the performance of the validation measures for a model with reasonably strong predictive ability. In addition, β_0 was set to a fixed value of -1.8 to generate data with a prevalence of approximately 20% for each of the ICC scenarios.

4.5.1.2 Simulation scenarios

A total of four ICC values such as 0%, 5%, 10%, and 20% were considered, to mimic scenario with low, medium, and high level of clustering. Under each ICC value, development datasets each with 100 clusters of size 100 were generated. For each development set, validation datasets from several scenarios were generated, to represent scenarios with small number of large clusters and large number of small clusters. The validation scenarios considered were (i) 10 clusters of sizes 10 and 300, and (ii) 100 clusters of sizes 10, 30, and 100. For each of the four ICC values, there are one development and five validation scenarios, and in total four development and twenty validation scenarios. For each of the development and validation scenarios, 500 datasets were generated. This specification (500 replications) was determined following Burton et al. [78] and provided very low Monte Carlo standard error for the validation measures for clustered binary outcome. The level of clustering in the development and validation data were kept equal, generating both data from the same ICC value. This would represent a scenario where subjects in development and validation data are sampled from the same population of clusters, where the level of clustering in both datasets are equal.

4.5.2 Strategies for evaluating the measures

4.5.2.1 Model fitting and calculation of the measures

A random-intercept logistic model with normally distributed random effects was fitted to each of the development datasets. Maximum likelihood estimation based on adaptive Gaussian quadrature [97, 99] was employed to obtain the estimates of the model parameters, $(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}_u^2)$. The `gllamm` package in Stata version 11 [129] was used to obtain these estimates. To calculate the validation measures, the estimated event

probabilities based on $\pi_{ij}(u)$, $\pi_{ij}(0)$, and $\pi_{ij}(\text{pa})$ and the associated predicted prognostic indices $\eta_{ij}(u)$, $\eta_{ij}(0)$, and $\eta_{ij}(\text{pa})$ were obtained in the corresponding simulated validation datasets by plugging in the estimates of the model parameters from the development data. The `gllapred` package was used to obtain these predictions.

To make predictions based on $\pi_{ij}(u)$, the random effects u were estimated from the validation data. The `gllapred` package calculates the empirical Bayes estimates of the random effects in the validation data using equation (4.9), without fitting a model, but using the estimates $(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}_u^2)$ from the development data. This can be considered as a re-calibration of the model based on the random effects. Finally the point estimates and confidence intervals of the validation measures were calculated using user written Stata code (Appendix B: Figure B.2).

4.5.2.2 Assessing the properties

The effects of the ICC, the number of clusters and their size on the validation measures were investigated through simulation by estimating the empirical bias and rMSE of the point estimates and coverage of the nominal 90% confidence intervals. The true values of the 'overall' and 'pooled' validation measures were obtained empirically by averaging the estimates of the measures over 100 very large simulated datasets (N=300,000 with clusters J=1000). The 'overall' validation measures were calculated using the true values of the regression parameters and the random effects. The rank-based 'pooled' measures (C_w and D_w) and the calibration slope (CS_w) were calculated using the true values of the regression parameters only as the random effects do not contribute to the calculation of these measures (for more details see Section 4.3.5). However, the true value of the 'pooled' Brier score was calculated using the true value of the regression parameters and the random effects.

Bias in the estimate of the validation measure was calculated as the mean of the differences between the true and estimated values for each validation measure, over 500 simulations. The rMSE was calculated as the square root of the mean of the squared differences between the true and estimated values for each validation measure. Coverage was calculated as the percentage of simulations where the estimated confidence interval

contained the true value of the validation measure. Coverage was calculated for both analytical and bootstrap based confidence intervals for each validation measure. In the bootstrapping approach, 200 bootstrap samples were used, where the sample drawn during each replication was a bootstrap sample of subjects within each cluster.

The validation measures have different scales and hence their bias and rMSE are not directly comparable. Therefore, the bias was rescaled to a percentage in a similar manner to that discussed in Chapter 3. Similarly, the rMSE was rescaled to a percentage as

$$\text{rMSE} = \sqrt{\frac{1}{G}\sum_{g=1}^{G}\left(\frac{\hat{m}_g - m}{|m - m_0|} \times 100\right)^2},$$

where $\hat{m}_g$ is the estimate for the gth simulation ($g = 1, \ldots, G$), m is the true value and m_0 is the null value.

4.5.3 Results

4.5.3.1 The Overall validation measures

The relative bias in the 'overall' estimates of the validation measures were plotted against different ICC values, for all the simulation scenarios. Figure 4.4 shows the results for the validation measures based on the different approaches to prediction. When there was no clustering in the data (ICC=0%), the validation measures in general showed approximately unbiased estimates for all simulation scenarios, though the D statistic and calibration slope showed a small amount of bias when both the number of clusters and their sizes were small. In the presence of clustering (ICC $> 0\%$), the validation measures $C^{np}_{re(u)}$, $D_{re(u)}$, $CS_{re(u)}$, and $BS_{re(u)}$ showed approximately unbiased estimates when the clusters were large. However, they showed bias for the small clusters. The bias associated with $C^{np}_{re(u)}$, $D_{re(u)}$, and $CS_{re(u)}$ increased with increasing ICC while for $BS_{re(u)}$, it decreased. The results for the parametric C-index were similar to those for the non-parametric C-index (not shown). Since both these C-indices had similar results for all the stimulation scenarios, no results for the parametric C-index will be shown in the rest of the chapter.

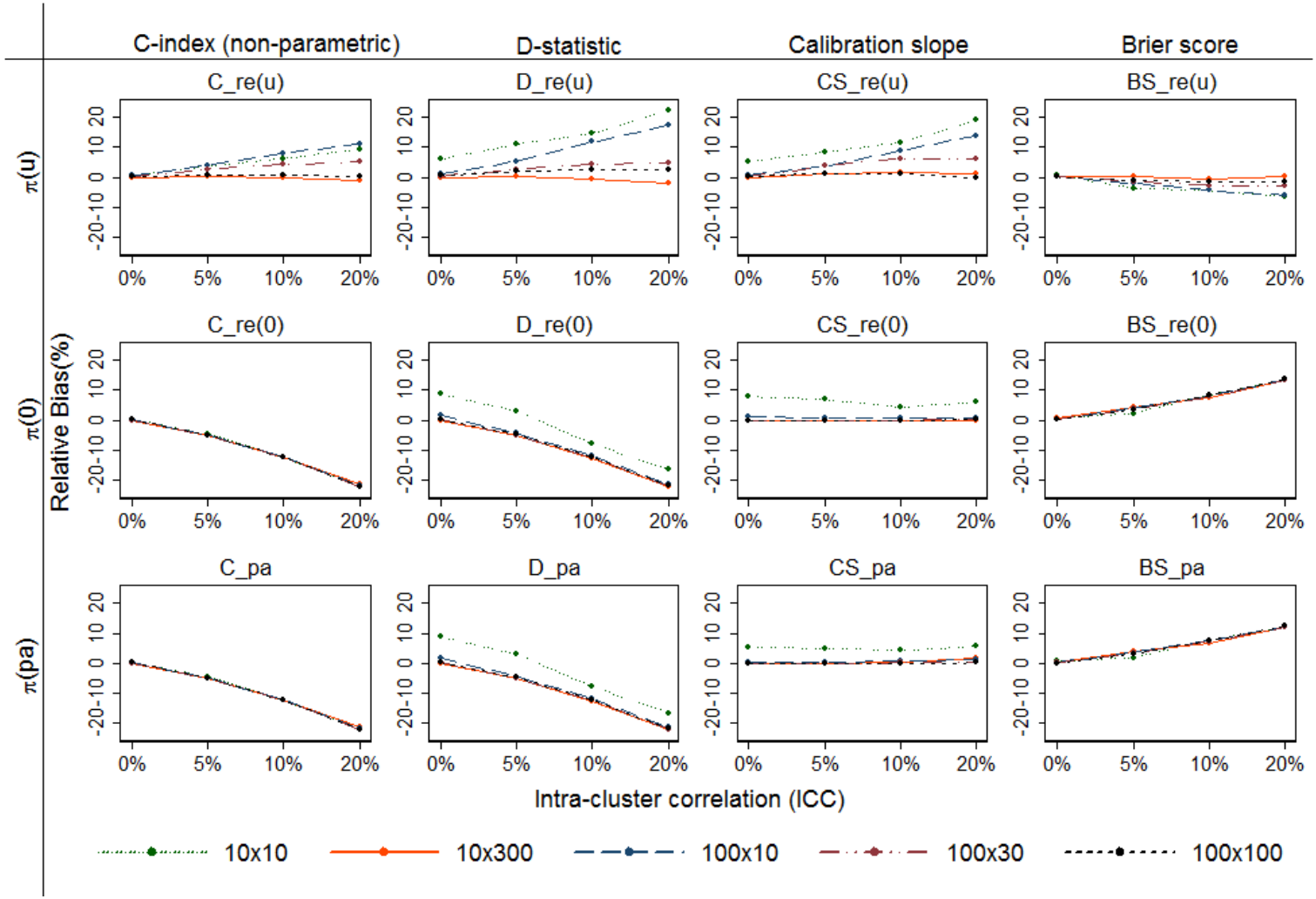

Figure 4.4: Relative bias (%) in the 'overall' estimates of the validation measures for different ICC values. The results are from the different simulation scenarios based on the number of clusters and their size (clusters×size). Each column represents plots of bias for the different estimates of a validation measure based on the model prediction $\hat{\pi}_{ij}(u)$, $\hat{\pi}_{ij}(0)$, and $\hat{\pi}_{ij}(\text{pa})$.

For all simulation scenarios, both $C^{np}_{re(0)}$ and C^{np}_{pa} showed substantial negative bias (but of equal amount) in the presence of clustering, and the bias increased with increasing ICC values. Similar results were observed for $D_{re(0)}$ and D_{pa}. Furthermore, for all simulation scenarios, the bias associated with $BS_{re(0)}$ and BS_{pa} were positively correlated with the ICC values. The calibration slopes $CS_{re(0)}$ and CS_{pa} were not affected by the level of clustering, but were affected by the number and size of the clusters, for example, 10 clusters of size 10.

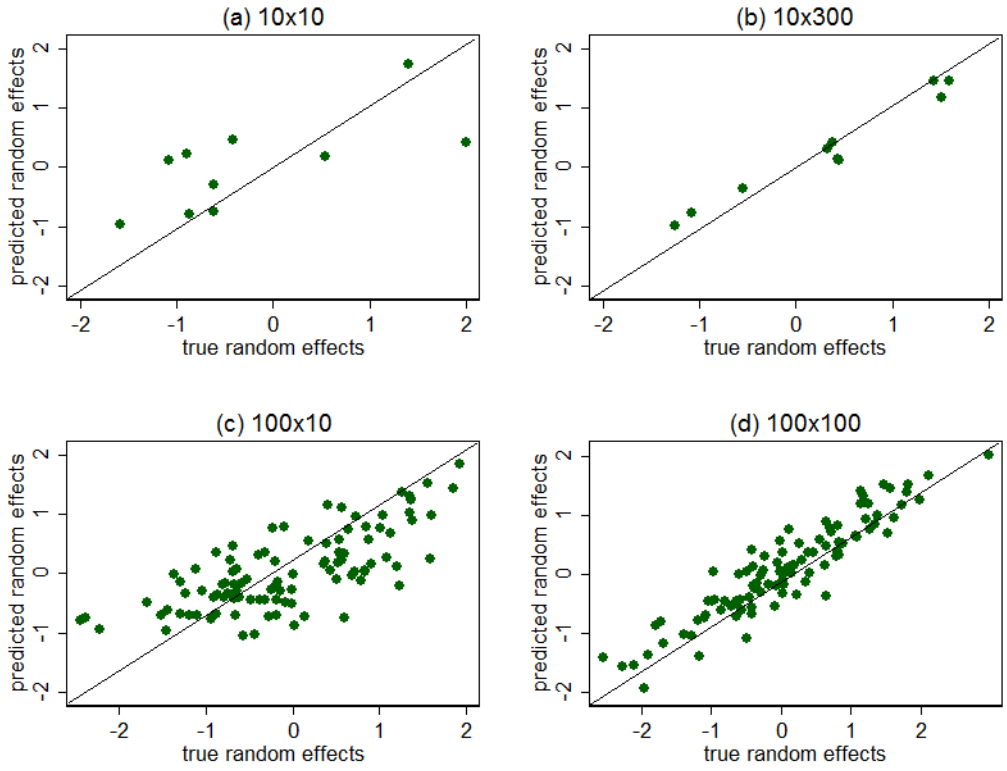

Figure 4.5: Agreement between the estimated ($\hat{u}$) and the true random effects u in the validation data. The results are from the different simulation scenarios under ICC=20%: number of clusters (a) 10 of size 10, (b) 10 of size 300, (c) 100 of size 10, and (d) 100 of size 100. Figure 2b shows nine points, because two points amongst the ten correspond to the same values and hence represents one point.

The reason for bias in the validation measures based on $\hat{\pi}_{ij}(u)$ when the clusters are small is possibly due to the poor estimation of the random effects. To investigate this, the empirical Bayes estimates of the random effects from the validation data were

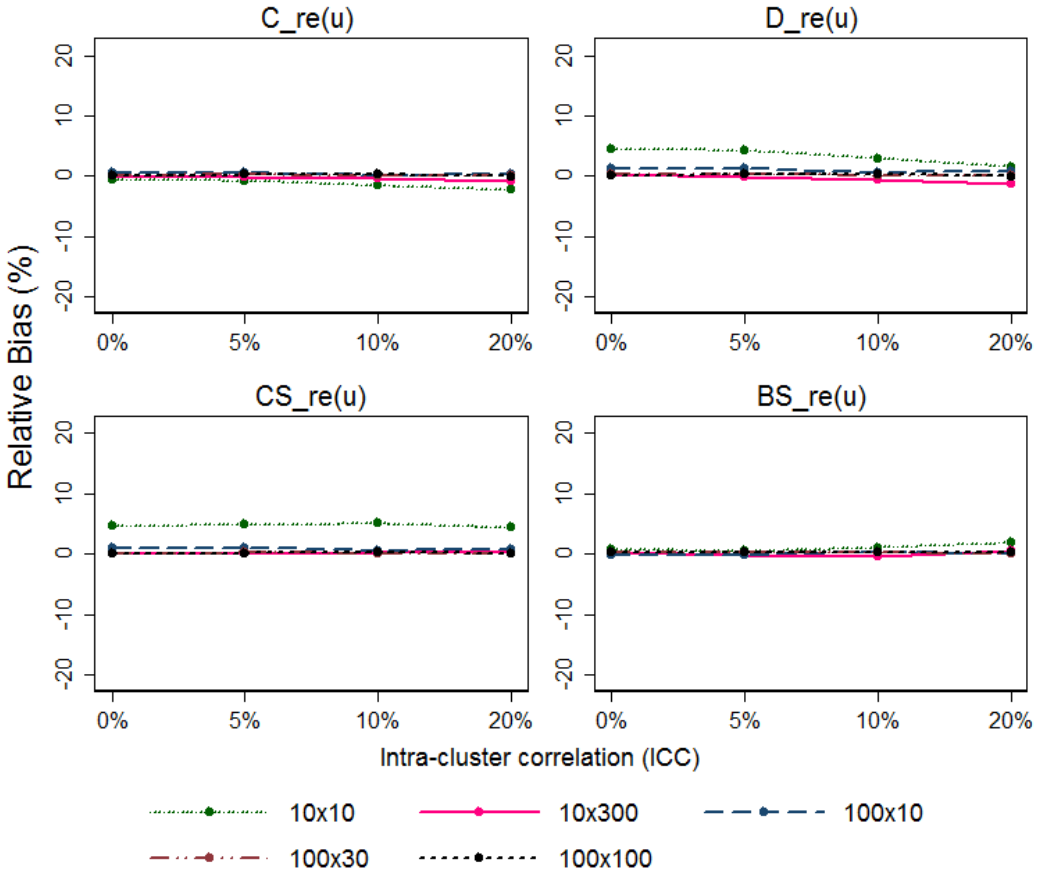

Figure 4.6: Relative bias (%) in the 'overall' estimates of the validation measures for $\pi_{ij}(u)$ when they were calculated using the true values of the random effects u, rather than the estimates. The results are from the different simulation scenarios (clusters×size).

plotted against their true values in Figure 4.5. It appears that the random effects were poorly estimated especially when the cluster sizes were small and the level of clustering was high. Figure 4.5 shows that there was poor agreement between the estimated and the true values of the random effects when the clusters were small (Figure 4.5(a) and (c)), but there was close agreement when the clusters were large (Figure 4.5(b) and (d)). The empirical Bayes estimates are conditionally biased, that is, conditional expectation of random effects given the population value of the random effects $E(\hat{u}_j|u_j, x_{ij}\hat{\sigma}_u) \neq 0$, which pull the empirical Bayes towards 0, the mean of the prior distribution [86]. This is because the prior dominates the likelihood when cluster sizes are small.

When the empirical Bayes estimates of the random effects were replaced by their

true values in the calculation of the validation measures while still using the estimates of the fixed predictors, the measures showed a reasonably good performance, even for the small clusters (Figure 4.6). These results are analogous to those derived by Oirbeek and Lesaffre [131]. However, $D_{re(u)}$ and $CS_{re(u)}$ were slightly biased when the number and size of the clusters were small, even if clustering did not exist. This implies that these two measures are affected by small sample size. The validation measures based on $\hat{\pi}_{ij}(0)$ and $\hat{\pi}_{ij}(\text{pa})$, excluding the calibration slope (CS), showed bias in the presence of clustering. This is because all these measures ignore the actual contribution of the random effects and therefore underestimate the true value.

The relative rMSE of the 'overall' estimates of the validation measures are presented for different ICC values, for the various simulation scenarios in Figure 4.7. Figure 4.7 shows the results for all validation measures based on the model's different approaches to prediction. The validation measures in general had high rMSE for small clusters. The measures based on $\hat{\pi}_{ij}(u)$ had low rMSE for all ICC values when the clusters were large. The validation measures based on $\hat{\pi}_{ij}(0)$ and $\hat{\pi}_{ij}(\text{pa})$ had low rMSE when there was no clustering and the clusters were large. However, the rMSE associated with these measures, except for the calibration slope, increased with increasing ICC values.

Coverage of nominal 90% confidence intervals (CIs) for each of the validation measures based on both analytical and bootstrap standard errors (SEs) are reported in Table 4.2. The table shows the results for the validation measures based on $\hat{\pi}_{ij}(u)$. Coverage for $BS_{re(u)}$ based on analytical CIs is not reported as it is not available. The estimated coverage for $C^{np}_{re(u)}$, $D_{re(u)}$, and $CS_{re(u)}$, based on both analytical and bootstrap CIs, were approximately close to the nominal 90% value when the clusters were large. When the clusters were small, both the analytical and bootstrap CIs had poor coverage, because the point estimates of the measures were biased. Similar results were observed for $BS_{re(u)}$ based on bootstrap based CIs. In general, coverage for all the simulation scenarios decreased slightly with increasing ICC as their SE decreased. All the validation measures based on $\hat{\pi}_{ij}(0)$ and $\hat{\pi}_{ij}(\text{pa})$ had good coverage when the clusters were large and there was no clustering, but they had poor coverage when the level of clustering was high as their point estimates were biased (results not shown).

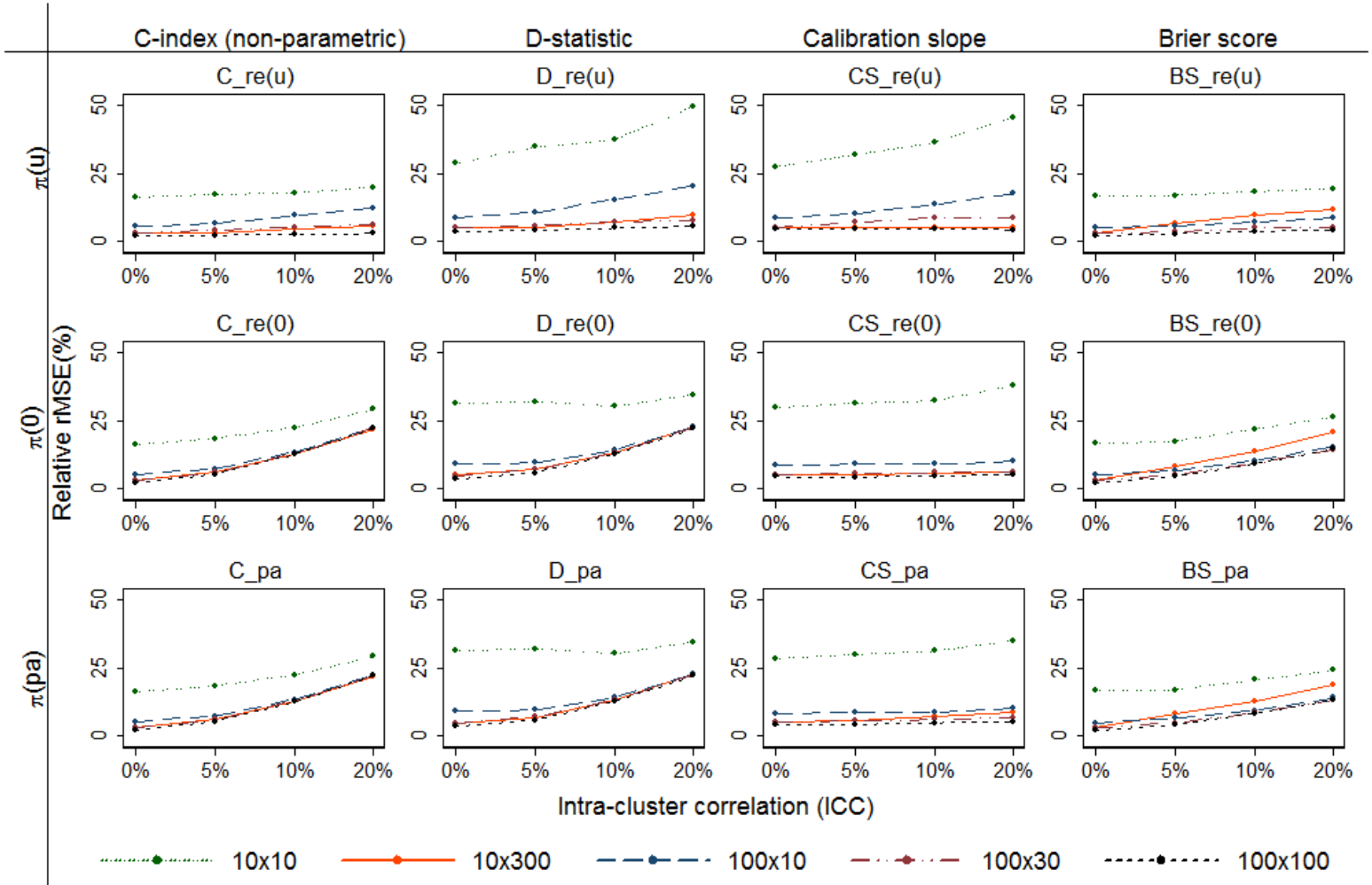

Figure 4.7: Relative rMSE (%) of the 'overall' estimates of the validation measures for different ICC values. The results are from the different simulations scenarios (clusters×size). Each column represents plots of rMSE for different estimates of a validation measure based on $\hat{\pi}_{ij}(u)$, $\hat{\pi}_{ij}(0)$, and $\hat{\pi}_{ij}(\text{pa})$.

Table 4.2: Coverage (%) of nominal 90% confidence intervals (CIs) of the 'overall' validation measures. The confidence interval are based on both analytical and bootstrap standard errors. Maximum Monte Carlo Standard Error=2.25%.

Coverage (analytical CIs)										
			$C^{np}_{re(u)}$				$D_{re(u)}$			
Cluster	Size	ICC	0%	5%	10%	20%	0%	5%	10%	20%
10	10		91	87	78	77	92	84	81	79
	300		90	88	87	85	90	86	83	81
	10		89	79	65	58	88	83	72	57
100	30		87	78	68	60	90	72	63	54
	100		90	86	85	84	89	83	80	78
			$CS_{re(u)}$				$BS_{re(u)}$			
10	10		93	75	86	81	-	-	-	-
	300		89	87	88	86	-	-	-	-
	10		88	69	49	35	-	-	-	-
100	30		88	56	45	30	-	-	-	-
	100		87	83	84	82	-	-	-	-
Coverage (normal-based bootstrap CIs)										
			$C^{np}_{re(u)}$				$D_{re(u)}$			
10	10		84	85	84	82	93	95	95	87
	300		90	89	88	87	90	88	87	86
	10		85	84	83	75	86	82	78	70
100	30		86	82	74	66	85	80	73	64
	100		89	85	86	84	91	85	84	80
			$CS_{re(u)}$				$BS_{re(u)}$			
10	10		93	95	94	97	94	91	88	87
	300		88	87	87	86	89	90	88	87
	10		88	87	84	82	88	87	84	82
100	30		86	78	72	76	87	87	85	83
	100		89	88	88	84	88	89	86	85

The above simulation study was performed using data with equal cluster sizes. However, most real datasets have clusters of unequal sizes. Therefore, further simulation studies were performed to investigate the performance of the validation measures in this scenario. Two validation scenarios were considered with 30 clusters of either median size 50 (IQR: 29 to 90) or 145 (IQR: 54 to 365). The same ICC values were considered as before. The relative biases of the 'overall' validation measures are presented in Figure 4.8. In general, these results are similar to those obtained for the simulations based on equal cluster sizes. The results for rMSE and coverage were also similar to those obtained before (not shown).

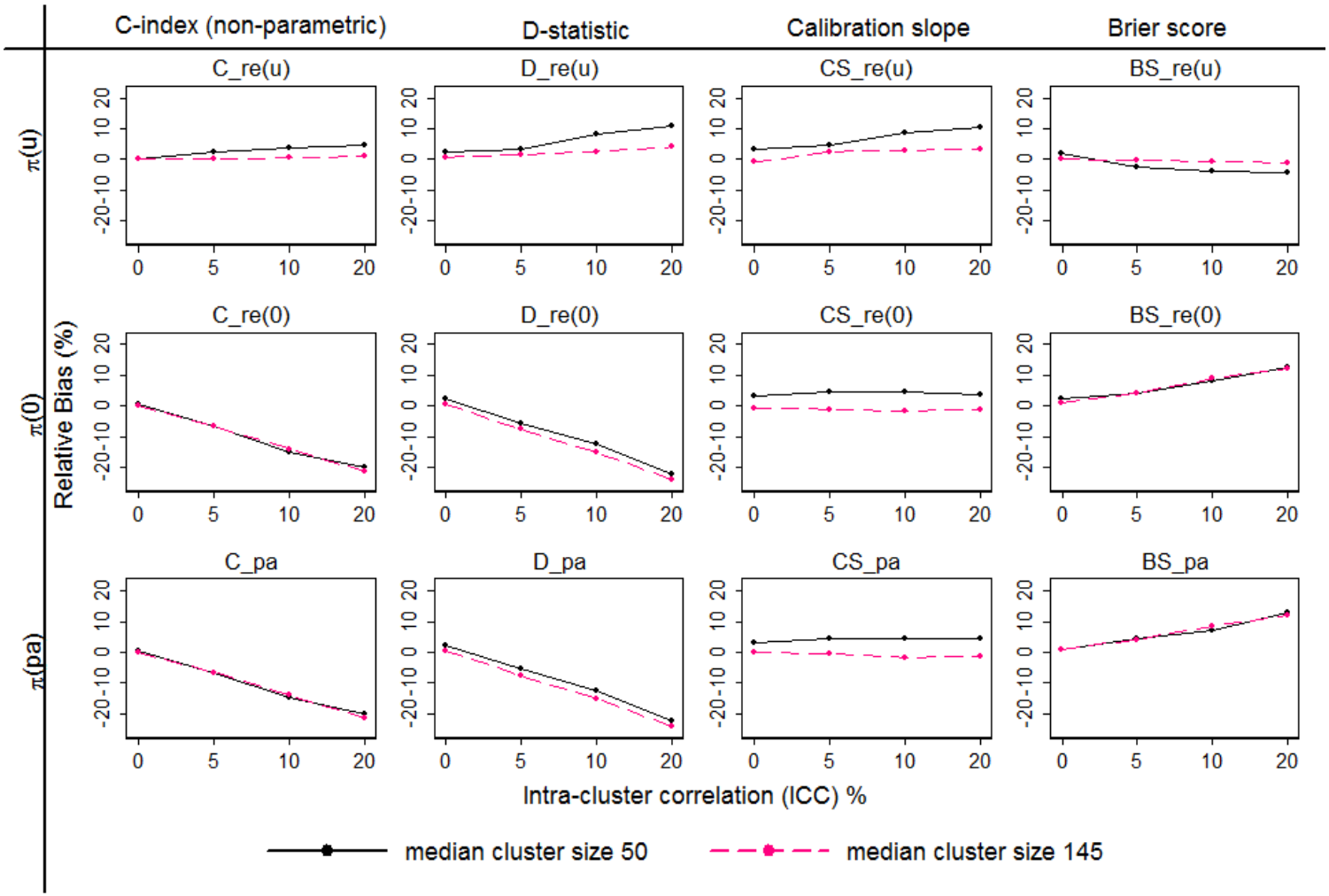

Figure 4.8: Relative bias (%) in the 'overall' estimates of the validation measures for different ICC values. The results are from the different simulation scenarios based on unequal cluster sizes: 30 clusters with median sizes 50 or 145. Each column represents plots of bias for the different estimates of a validation measure based on the model prediction $\hat{\pi}_{ij}(u)$, $\hat{\pi}_{ij}(0)$, and $\hat{\pi}_{ij}(\text{pa})$.

4.5.3.2 The Pooled cluster-specific validation measures

The bias in the 'pooled' estimates of the cluster-specific validation measures were plotted for different values of the ICC, for various simulation scenarios in Figure 4.9. The rank-based measures (C_w, D_w) and the calibration slope (CS_w) were unbiased when clusters were large, but they showed large bias for small clusters. The Brier score $BS_{w,re(u)}$ based on $\pi_{ij}(u)$ includes the random effects and was therefore affected by the ICC when the clusters were small, as the random effects were poorly estimated for these clusters. However, $BS_{w,re(u)}$ was unbiased when the clusters were large. Both $BS_{w,re(0)}$ and $BS_{w,pa}$ based on $\pi_{ij}(0)$ and $\pi_{ij}(\text{pa})$ respectively also showed bias in the presence of clustering, even when the clusters were large (results not shown).

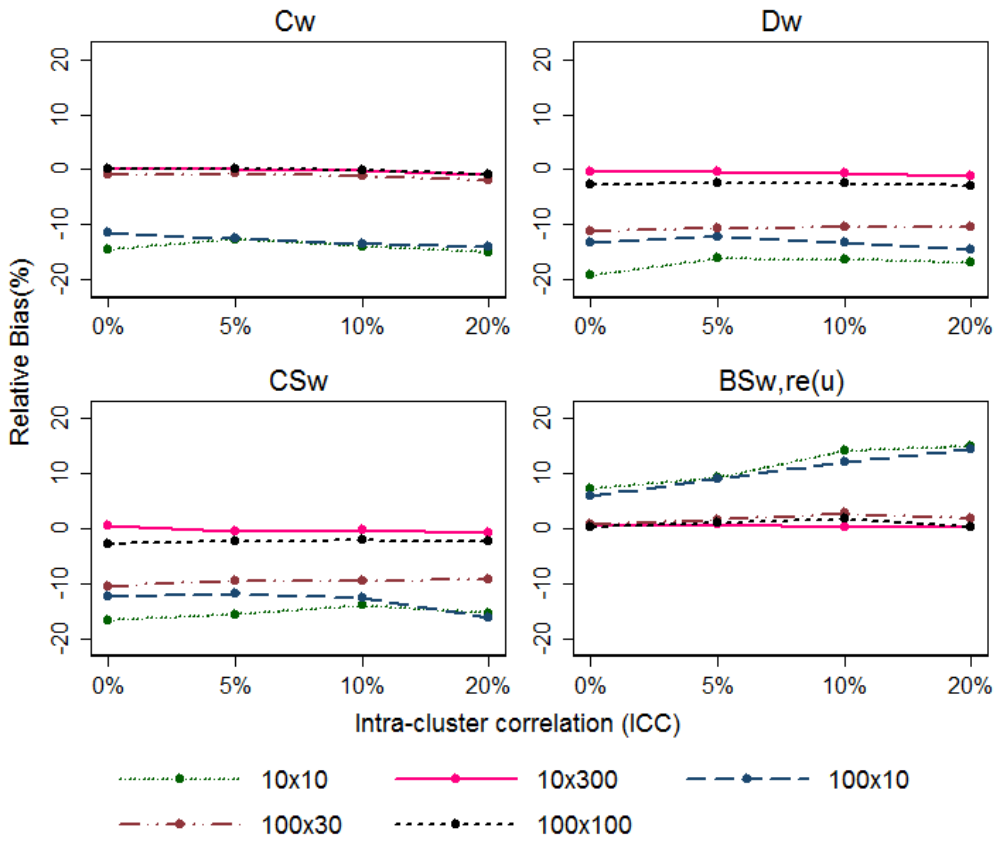

Figure 4.9: Relative bias (%) in the 'pooled' estimate of the validation measures for different ICC values. The results are from the different simulations scenarios (clusters×size).

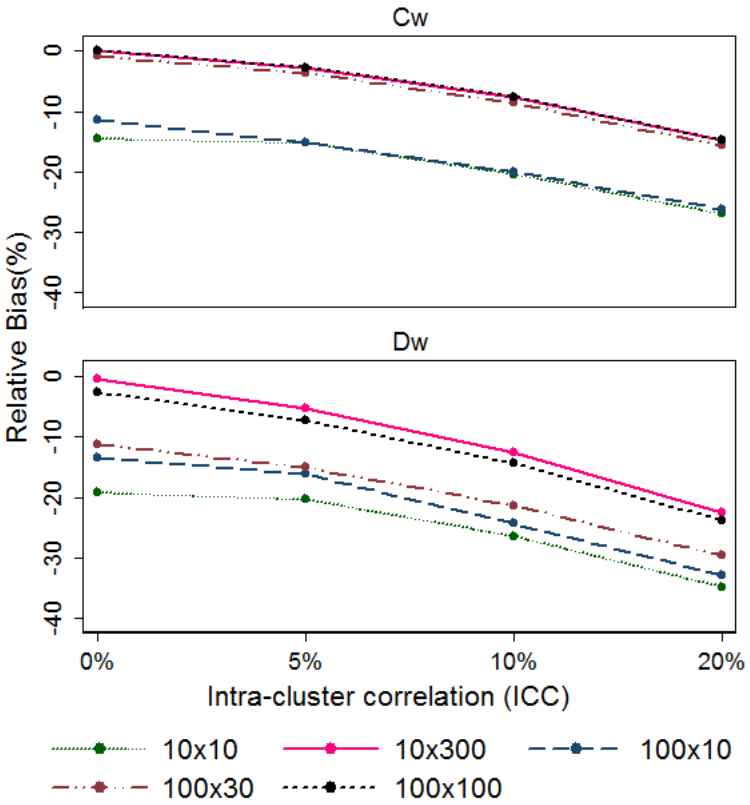

Figure 4.10: Relative bias (%) in the 'pooled' estimates of the C-index and D statistic when calculating bias against the 'overall' true values. The results are from the different simulations scenarios (clusters×size).

The extent of bias in the 'pooled' estimates of the validation measures was also compared with the 'overall' true values, since these values are able to capture the variability between the clusters (cluster characteristics) in addition to the subject-level variability (subject characteristics). Only results for the C-index (C_w) and D statistic (D_w) are presented in Figure 4.10. The measures were approximately unbiased when the clusters were large and there was no clustering. However, the bias increased with increasing ICC values, even with the large clusters.

The possible reason for bias in the 'pooled' estimates of the cluster specific measures when the clusters are small is as follows. The prevalence of the outcome was set at 20% for the simulations. However, the number of events varied between the clusters for high

values of the ICC. The minimum number of events required per cluster to calculate the non-parametric C-index and the Brier score is one and is two for the parametric C-index, D statistic, and calibration slope. When calculating a validation measure based on small clusters, if the number of events for a cluster was too low, the cluster was ignored. Thus the calculation of the 'pooled estimate' was often based on a reduced number of clusters, resulting in bias. In Table 4.3, the number of dropped clusters is reported. This shows that approximately 12-20% of small clusters were dropped as they did not have at least one event to calculate C_w^{np} and $BS_{w,re(u)}$, whereas 50-55% clusters were dropped for the calculation of D_w and CS_w. Consequently, for the simulation scenarios with small clusters, the bias in D_w and CS_w was larger than that for C_w^{np} and $BS_{w,re(u)}$. However, when the clusters were large, hardly any clusters were dropped, resulting in unbiased pooled estimates.

Table 4.3: Distribution of the number of clusters dropped when calculating validation measures within a cluster. The results are presented by the number of events required to calculate a measure. Each figure is the average over 500 simulations.

		Number of events required								
			One event				Two events			
Clusters	Size	ICC	0%	5%	10%	20%	0%	5%	10%	20%
10	10		1.2	1.5	1.6	1.8	4.9	4.9	5.1	5.2
	300		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01
100	10		12.2	14.5	17.4	20.2	49.3	50.2	51.6	53.5
	30		0.1	0.4	0.9	1.3	1.9	3.6	6.3	10.1
	100		0.0	0.0	0.01	0.11	0.0	0.01	0.08	0.34

The estimated rMSE for the 'pooled estimate' of the cluster-specific measures are presented for different ICC values in Figure 4.11. All the 'pooled' cluster-specific measures had very low rMSE when clusters were large, but had high rMSE for small clusters. The coverage of nominal 90% CIs for the 'pooled estimate' of the cluster-specific measures based on analytical SEs are reported in Table 4.4. For all simulation scenarios with different ICC values, the measures had good coverage when the clusters were large. However, the coverage was poor when the clusters were small, because the point estimates of the measures were biased.

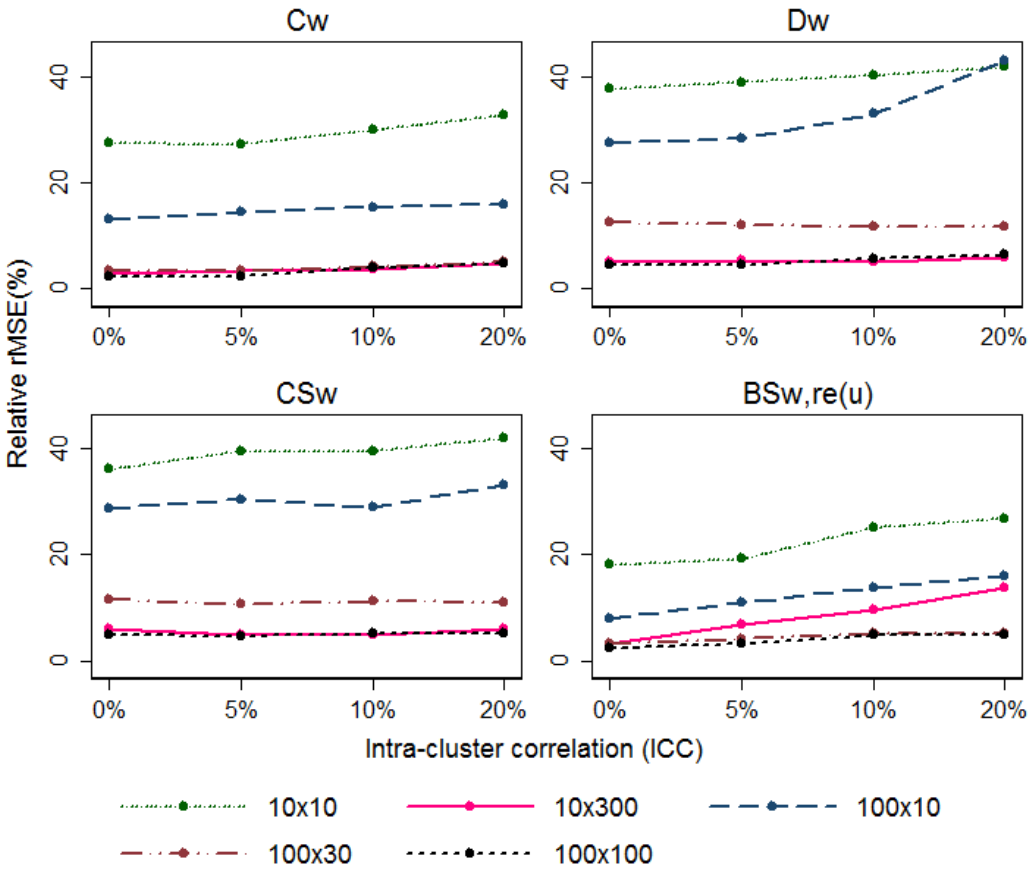

Figure 4.11: Relative rMSE (%) of the 'pooled' estimates of the validation measures for different ICC values. The results are from the different simulations scenarios (clusters×size).

The relative biases in the 'pooled' estimates of the cluster-specific measures obtained from the simulations based on unequal cluster sizes are presented in Figure 4.12. These results are similar to those observed in the simulations that used equal cluster sizes. The results for rMSE and coverage were also similar to those obtained before (not shown).

Table 4.4: Coverage (%) of nominal 90% confidence intervals (CIs) for the 'pooled' estimates of the cluster-specific measures. The CIs are based on analytical standard errors of the measure. Maximum Monte Carlo Standard Error = 2.37%.

			Coverage (analytical CIs)							
			C_w^{np}				D_w			
Clusters	Size	ICC	0%	5%	10%	20%	0%	5%	10%	20%
10	10		92	89	88	88	89	86	88	91
	300		90	91	90	89	90	91	90	91
100	10		55	56	58	57	97	91	95	96
	30		89	90	90	91	55	56	55	59
	100		89	90	89	89	84	85	87	88
			CS_w				$BS_{w,re(u)}$			
10	10		90	97	92	91	82	75	71	67
	300		87	89	89	88	90	89	88	89
100	10		97	92	97	87	80	75	70	63
	30		62	68	67	70	86	85	83	84
	100		86	84	85	84	87	88	84	86

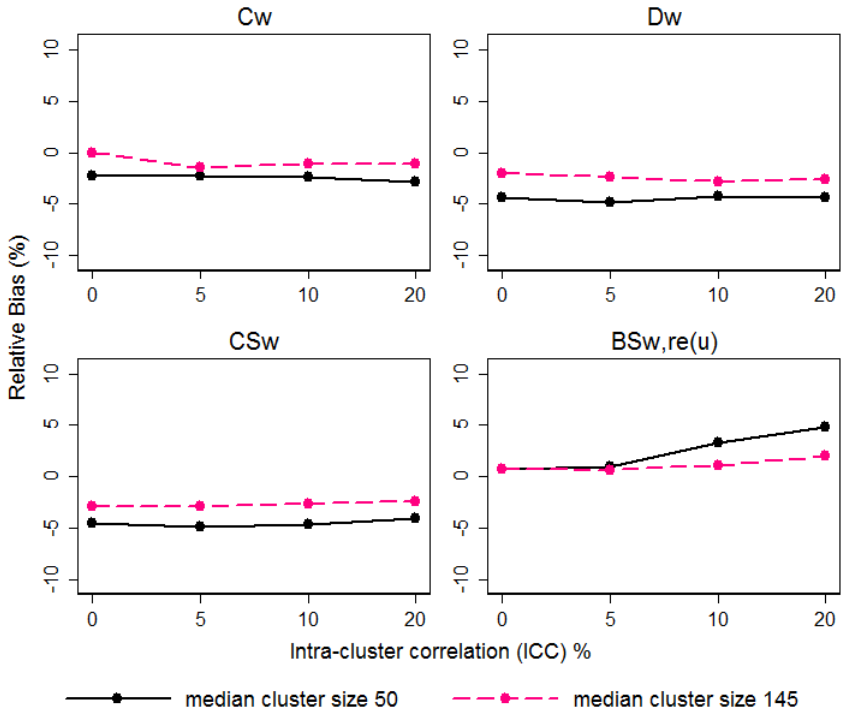

Figure 4.12: Relative bias (%) in the 'pooled' estimate of the validation measures for different ICC values. The results are from the simulations based on unequal cluster sizes.

4.6 Conclusion

This chapter has described an adaptation of the C-index, D statistic, calibration slope, and Brier score for use with models for clustered binary outcomes. Two approaches are proposed: an 'overall' and a 'pooled cluster-specific' measures. Each approach produces three different values depending on the model predictions $\hat{\pi}_{ij}(u)$, $\hat{\pi}_{ij}(0)$, and $\hat{\pi}_{ij}(\text{pa})$. The decision regarding which predictions to use should depend on the research objective.

The new validation measures were illustrated using a dataset of patients who underwent heart valve surgery. The results showed that both the 'overall' and 'pooled cluster-specific' validation measures have a meaningful interpretation in a clustered data setting. The properties of the measures were evaluated by a simulation study in a range of clustered data scenarios. The simulation results showed that the 'overall' validation measures based on $\hat{\pi}_{ij}(u)$ showed reasonable performance when there was clustering in the data and the clusters were reasonably large, possibly due to the fact that the random effects were better estimated in larger clusters. The empirical Bayes estimates of the random effects are poorly estimated when the clusters are small, in other words, do not have sufficient number of events [86]. This is because the prior dominates the likelihood, which pulls the empirical Bayes towards 0, the mean of the prior distribution. When the empirical Bayes estimates were replaced by the true values of the random effects while still using the fixed predictor effects, the measures showed good performance even for the small clusters. The 'overall' measures based on $\hat{\pi}_{ij}(0)$ and $\hat{\pi}_{ij}(\text{pa})$ performed poorly when there was a moderate level of clustering in the data, because they ignore the effect of clustering. The 'pooled cluster-specific' measures showed bias when the cluster sizes were small. This is because this approach ignores information from some of these clusters due to lack of events to calculate the measures.

In general, both the 'overall' and 'pooled cluster-specific' measures are recommended to use to assess the predictive ability of the cluster-data model. However, one needs to check whether the clusters are sufficiently large (for example, greater than 30) and each of these contains at least two events before using the 'pooled' measures.

Similar to the measures for independent survival outcomes, the validation measures for binary outcome differ in their flexibility regarding their assumptions and the form of the prognostic model. Both the parametric C-index and D statistic assume that the prognostic index derived from the model is distributed as normal. In contrast, the non-parametric C-index only requires that the prognostic model is able to rank the patients. The calibration slope assumes that the model is correctly specified. The Brier score only requires that a risk algorithm can be calculated for all patients. One needs to be aware of these before choosing the measures. In practice, the non-parametric C-index, calibration slope, and Brier score are recommended since they are free from a distributional assumption of the prognostic index. The parametric C-index and D statistic can be used only if the prognostic index is normally distributed.

In practice, when validating the model using subjects from the same cluster as that of the development data, predictions using the estimate of the random effects, $\hat{\pi}_{ij}(u)$, and the validation measures based on this approach are recommended. This is because the random effects for the clusters are known and validation measures based on this approach showed reasonable performance in the simulation study. It would not be straightforward to use this approach for validating model using subjects from new clusters, since the random effects of the new cluster are unknown. In this situation, firstly, one may inspect the characteristics of the new clusters to see whether these are similar to those of the development data. Then it may be reasonable to assume that the clusters in the development and validation data come from the same population of clusters and thus the level of clustering in both datasets are approximately equal. In this case, one could assess the equality in the level of clustering between development and validation data by using the confidence intervals for the variance parameters of the random effects estimated from both datasets or using F-test, provided that the number of cluster is reasonably large and the random effects are normally distributed. If equality holds then one could make predictions based on $\hat{\pi}_{ij}(u)$ and use the validation measures based on this approach. In this case, the random effects can be estimated from the validation data using the estimates of the variance parameter of the random effects from the development data. Then one could consider this as a form of model re-calibration. However, equality in the level of clustering between two datasets is unlikely in practice.

If the type of between cluster heterogeneity is different between the validation and development datasets, marginal predictions $\hat{\pi}_{ij}(\text{pa})$ or conditional predictions that set the random effects at their mean value of zero, $\hat{\pi}_{ij}(0)$, could be used if ICC is less than 0.05.

In summary, it is important to investigate the validation data before choosing the validation measures. In particular, one needs to check whether the validation data involve the same (or different) clusters as the development data, the level of clustering, cluster size, prevalence, and the distribution of prognostic index.

Using a similar approach to that discussed in this chapter, the next chapter discusses possible extensions of some of the validation measures for independent survival outcomes discussed in Chapter 3 for use with models for clustered survival outcomes.

Chapter 5

Measures for clustered survival outcomes

5.1 Introduction

The last chapter has investigated the use of validation measures for clustered binary outcomes. This chapter focuses on validation measures for clustered survival outcomes. Although a number validation measures for standard survival models have been developed (see, Chapter 3), very limited work has been done validation measures for models with clustered survival outcomes. This chapter discusses possible extensions of some of the standard validation measures for use with risk models that can handle clustered survival outcomes.

Frailty models are extensions of standard survival models with a frailty term or random effect included in the models [132–134]. These models are often used to analyse clustered survival data and have a cluster-specific or conditional interpretation, given the frailty. A possible alternative to the frailty models are the standard survival models with an adjustment, for the clustering of the data, for standard errors of the regression parameters [135–137]. These models have a population-averaged or marginal interpretation and are referred to as 'marginal models'. Generally, preference for using one of these two classes of models depends on the research question. However, frailty

model may be considered to be a more general type of model for analysing clustered survival data, because marginal interpretation of the predictors, can be derived from the frailty model by integrating out the frailty term [134]. This research discusses the use of frailty models in risk predictions for clustered survival data.

Some of the more commonly used validation measures for standard survival models have been considered in Chapter 3. For example, the calibration slope [44] is used to assess the calibration of a standard survival model. Similarly, Harrell's C-index [40], Gönen and Heller's $K(\beta)$ [48], and Royston and Sauerbrei's D [49] have been developed to assess discrimination, and Graf et al'e IBS and its R^2 extension assess both calibration and discrimination. In this chapter, these measures are extended for use with frailty models.

This chapter is organised as follows. Section 5.2 discusses the extensions of the validation measures mentioned above for use with clustered survival data. In Section 5.3, an application of the methods is illustrated using child mortality data from Bangladesh. Section 5.4 discusses simulation studies to evaluate the performance of the measures, and Section 5.5 ends this chapter with a discussion and conclusion.

5.2 Extension of the validation measures for use with clustered survival data

This section begins with a description of basic notation based on the Proportional Hazards (PH) frailty model and is followed by the detailed calculation of the validation measures for use with the PH frailty models.

5.2.1 The Proportional Hazards frailty model

Let us suppose that we have data $(t_{ij}, \delta_{ij}, \boldsymbol{x}_{ij})$ $(i = 1, \ldots, n_j; j = 1, \ldots, J)$ on N subjects from J different clusters of size n_j and $\sum_{j=1}^{J} n_j = N$, where for the ith subject belonging to the jth cluster, t_{ij} is the observed time, δ_{ij} is 1 if the event of interest is experienced at t_{ij} or 0 otherwise (right censoring), and $\boldsymbol{x}_{ij}$ is the ith row vector of the p-predictors.

To take account of the clustering effect, the standard proportional hazards (PH) model is extended to the PH frailty model by introducing a frailty term ω_j for the jth cluster. These frailties ω_js represent the effect of unobserved cluster-level predictors and vary across clusters. Since all subjects in the same cluster share the same frailty, this model is also called a shared frailty model. The hazard function of the PH shared frailty model takes the following form:

$$\begin{aligned} h(t|\boldsymbol{x}_{ij}, \omega_j) &= \omega_j h_0(t) \exp(\boldsymbol{\beta}^T \boldsymbol{x}_{ij}) \\ &= h_0(t) \exp(\boldsymbol{\beta}^T \boldsymbol{x}_{ij} + \ln \omega_j) \\ &= h_0(t) \exp(\boldsymbol{\beta}^T \boldsymbol{x}_{ij} + \varpi_j) \end{aligned} \tag{5.1}$$

where $\varpi_j (= \ln \omega_j)$ is the log frailty, and $\boldsymbol{\beta}^T \boldsymbol{x}_{ij} + \varpi_j$ is known as the prognostic index. The frailties are independent and identically distributed random variables that have a probability distribution $f(\omega|\theta)$, called the frailty distribution. Popular frailty distributions are the Gamma distribution and the inverse Gaussian distribution, which are all well-known members of the power variance family [138]. In this Chapter, the one parameter Gamma distribution [139] is considered as the frailty distribution because of its computational convenience. The frailties ω_js follow a Gamma distribution with mean 1 and variance θ, which is estimated from the data. The variance parameter θ is interpreted as a measure of heterogeneity in the risk of failures across clusters. If $\theta = 0$, then values of ω are all identical to 1, which implies that there is no effect of clustering and the survival times are independent within as well as between clusters. When θ is large, values of ω are more dispersed, indicating greater heterogeneity in the cluster specific baseline hazards $\omega_j h_0(t)$. The variance parameter θ can also be used to estimate the intra-cluster correlation coefficient Kendall's τ, which is equal to $\theta/(2+\theta)$.

Various estimation methods have been proposed for estimating the model parameters, the fixed predictor effects $\boldsymbol{\beta}^T$, the variance parameter of the frailty, θ, and the cumulative baseline hazard function $H_0(t) = \int_0^t h_0(u)du$. These include the expectation maximisation (EM) algorithm [134, 140, 141], and the penalised likelihood approach [133, 134]. For a semiparametric shared gamma frailty model, both approaches have been shown to provide similar results [134].

5.2.2 Predictions from the frailty model

The predictive form of the PH frailty model can be written in terms of the survival function as

$$S(t|\boldsymbol{x}_{ij}, \omega_j) = [S_0(t)]^{\exp(\boldsymbol{\beta}^T \boldsymbol{x}_{ij} + \varpi_j)}. \tag{5.2}$$

To make predictions, one uses the estimates of $\boldsymbol{\beta}^T$ and $S_0(t)$ and the estimate of the log frailty ϖ_j. One approach to obtain the estimate of ϖ_j is empirical Bayes estimation [133, 142]. Briefly, the empirical Bayes estimates are the means of the posterior distribution of the frailty distribution, given the estimated model parameters and the data.

Similar to the random-intercept logistic model discussed in Chapter 4, the frailty model can be used to predict the survival probability using three different approaches depending on how the frailties are used in the predictions. These are conditional predictions obtained by either plugging in the estimated log-frailties $\hat{\varpi}_j$ or specifying the frailty at their mean value 1 (or the log-frailty at 0), and marginal predictions obtained by integrating out the frailty term from the conditional frailty model. The resulting marginal survival function takes the following form for marginal predictions:

$$S(t|\boldsymbol{x}_{ij}) = \int S(t|\boldsymbol{x}_{ij}, \omega) f(\omega|\theta) d\omega.$$

For convenience, these three approaches to prediction are denoted by $S(t|\omega)$, $S(t|1)$, and $S(t)$, respectively. This chapter only discusses the extension of the validation measures for use with $S(t|\omega)$. However, the validation measures for $S(t|1)$ and $S(t)$ can be derived in an analogous way to those derived for $S(t|\omega)$.

5.2.3 Approaches for the calculation of the validation measures

As in Chapter 4, an 'overall' and a 'pooled cluster-specific' measures are considered to calculate validation measures for clustered survival data. Briefly, the 'overall' measure can be estimated by comparing the subjects within as well as between clusters, and the resulting estimate assesses the overall predictive ability of the model. For the 'pooled

cluster-specific' measure, one calculates the validation measure for each cluster, and these estimates are pooled across clusters using the random effects summary statistic methods described in Chapter 4. The 'pooled cluster-specific' measure does not compare subjects across clusters and thus assess the predictive ability of the predictors whose values vary within a cluster. The detailed calculations of validation measures for each these approaches are considered in the following sections of this chapter.

5.2.4 Estimation: Overall measures

5.2.4.1 Harrell's C-index

Harrell's C-index is an estimator of concordance probability and is based on the idea that, for a randomly selected pair of subjects, a survival model should predict a lower survival probability for the subject who fails earlier than that for the subject who fails later. The overall C-index is the proportion of all usable pairs in which predictions and outcomes are concordant (see Section 3.3.3.1, Chapter 3). This definition is adapted here for use with clustered data in the following way. A randomly selected pair of subjects i and k from clusters j and l respectively, with survival times t_{ij} and t_{kl} is said to be a usable pair if $t_{ij} \neq t_{kl}$. For censored data, a pair is usable if the shorter time corresponds to an event. With corresponding predicted survival probabilities $S(t|\boldsymbol{x}_{ij}, \omega_j)$ and $S(t|\boldsymbol{x}_{kl}, \omega_l)$, a usable pair is said to be concordant if either $S(t|\boldsymbol{x}_{ij}, \omega_j) < S(t|\boldsymbol{x}_{kl}, \omega_l)$ and $t_{ij} < t_{kl}$ or $S(t|\boldsymbol{x}_{ij}, \omega_i) > S(t|\boldsymbol{x}_{kl}, \omega_l)$ and $t_{ij} > t_{kl}$. Otherwise, the pair is said to be discordant. The concordance probability for clustered survival data can be defined as

$$C_{re} = \Pr\Big[S(t|\boldsymbol{x}_{ij}, \omega_i) < S(t|\boldsymbol{x}_{kl}, \omega_l)|t_{ij} < t_{kl}\Big],$$

or equivalently

$$C_{re} = \Pr\Big[(\boldsymbol{\beta}^T\boldsymbol{x}_{ij} + \varpi_j) > (\boldsymbol{\beta}^T\boldsymbol{x}_{kl} + \varpi_l)|t_{ij} < t_{kl}\Big].$$

This applies to all possible pairs (i, k) in the data, where the pairs can be formed by taking subjects from the same cluster or from different clusters. If subjects are from different clusters, their frailty values contribute in determining whether the pair

is concordant, even if both subjects have the same predictor values. However, the frailties do not contribute in determining a concordant pair if both subjects in the pair are from the same cluster, as they share the same frailty value.

Comparing all possible pairs (i,k), in which at least one subject of a pair had an event, with observed data $\{(t_{ij},\delta_{ij},\boldsymbol{x}_{ij}),(t_{kl},\delta_{kl},\boldsymbol{x}_{kl})\}$, the C-index then can be calculated for the frailty model as

$$\hat{C}_{re}=\frac{\sum_{j=1}^{J}\sum_{l=1}^{J}\sum_{i=1}^{n_j}\sum_{k=1}^{n_l}\Big[I\Big((\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_{ij}+\hat{\varpi}_j)>(\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_{kl}+\hat{\varpi}_l)\ \&\ t_{ij}<t_{kl}\ \&\ \delta_{ij}=1\Big)\Big]}{\sum_{j=1}^{J}\sum_{l=1}^{J}\sum_{i=1}^{n_j}\sum_{k=1}^{n_l}\Big[I(t_{ij}<t_{kl}\ \&\ \delta_{ij}=1)\Big]}, \quad (5.3)$$

where $I(\cdot)$ is the indicator function, $\hat{\boldsymbol{\beta}}^T$ is the estimate of $\boldsymbol{\beta}^T$, and $\hat{\varpi}_j$ is the empirical Bayes estimate of the log frailty ϖ_j.

Confidence interval for C_{re}

The method discussed by Pencina and D'Agostino [75] for the C-index for independent survival data is adapted to derive a confidence interval for C_{re} for clustered data. Let us define

$$\begin{aligned} c_{ijkl} &= 1 \text{ if the pair } (i,k) \text{ from clusters } (j,l) \text{ is concordant} \\ &= 0 \text{ if discordant.} \end{aligned} \quad (5.4)$$

Further let c_{ij} be the number of subjects in the dataset that are concordant with the ith subject from the jth cluster, then applying the above definition

$$c_{ij}=\sum_{l=1}^{J}\sum_{k=1}^{n_l}c_{ijlk}. \quad (5.5)$$

Considering the entire sample, the unconditional probability of concordance

$$\pi_c = \Pr\Big[(\boldsymbol{\beta}^T \boldsymbol{x}_{ij} + \varpi_j) > (\boldsymbol{\beta}^T \boldsymbol{x}_{kl} + \varpi_l) \;\&\; t_{ij} < t_{kl}\Big]$$

can be estimated as

$$\hat{\pi}_c = \frac{1}{N(N-1)} \sum_{j=1}^{J} \sum_{i=1}^{n_j} c_{ij}. \tag{5.6}$$

Similarly, if we let d_{ij} be the corresponding number of subjects that are discordant with the ith subject from the jth cluster, then the estimated unconditional probability of discordant is

$$\hat{\pi}_d = \frac{1}{N(N-1)} \sum_{j=1}^{J} \sum_{i=1}^{n_j} d_{ij}. \tag{5.7}$$

As discussed by Pencina and D'Agostino [75], $\hat{\pi}_c$ and $\hat{\pi}_d$ are unbiased estimates of π_c and π_d, respectively. Note that $\pi_c + \pi_d = 1$ if there are no ties. Using the relationship between Harrell's C-index and the modified Kendall's τ_m [143] developed by Pencina and D'Agostino [75], $\hat{C}_{re}$ defined in equation (5.3) can be written as

$$\hat{C}_{re} = \frac{\hat{\pi}_c}{\hat{\pi}_c + \hat{\pi}_d} = \frac{1}{2}(\hat{\tau}_m + 1), \tag{5.8}$$

where $\hat{\tau}_m = \dfrac{\hat{\pi}_c - \hat{\pi}_d}{\hat{\pi}_c + \hat{\pi}_d}$ is an estimate of τ_m.

Using these estimates, the following expression can be written:

$$\sqrt{N}(\hat{C}_{re} - C_{re}) = \sqrt{N}\left(\frac{\hat{\pi}_c}{\hat{\pi}_c + \hat{\pi}_d} - \frac{\pi_c}{\pi_c + \pi_d}\right) = \frac{\sqrt{N}(\pi_d \hat{\pi}_c - \pi_c \hat{\pi}_d)}{(\hat{\pi}_c + \hat{\pi}_d)(\pi_c + \pi_d)},$$

which is asymptotically equivalent to

$$\rho = \frac{\sqrt{N}(\pi_d \hat{\pi}_c - \pi_c \hat{\pi}_d)}{(\pi_c + \pi_d)^2}.$$

By virtue of the central limit theorem, the above statistic is approximately normally distributed for large N [75, 143]. Since $\hat{\pi}_c$ and $\hat{\pi}_d$ are unbiased estimates of π_c and π_d, respectively, it can be shown that $E[\rho] = 0$. Therefore, $\hat{C}_{re}$ is an asymptotically unbiased and normal estimator of C_{re} [75]. Using the result of Pencina and D'Agostino [75], the variance expression for ρ can be written as

$$\text{var}(\rho) = \frac{4}{(\pi_c + \pi_d)^4}(\pi_d^2\pi_{cc} - 2\pi_c\pi_d\pi_{cd} + \pi_c^2\pi_{dd}),$$

where, for three subjects (i, k, r) from clusters (j, l, s) respectively,

$$\begin{aligned}
\pi_{cc} &= \Pr[i \text{ is concordant with both } k \text{ and } r], \\
\pi_{dd} &= \Pr[i \text{ is discordant with both } k \text{ and } r], \\
\pi_{cd} &= \Pr[i \text{ is concordant with } k \text{ but discordant with } r], \\
\pi_{dc} &= \Pr[i \text{ is discordant with } k \text{ but concordant with } r].
\end{aligned}$$

The last two probabilities are equal, since k and r can be interchanged.

These probabilities can be estimated from data in the following way. The term π_{cc} is interpreted as the probability that a given subject from a given cluster is concordant with two other randomly selected subjects from any clusters. For subject i from cluster j, the possible number of pairs of subjects that are concordant with i can be calculated as $\frac{c_{ij}!}{(c_{ij}-2)!} = c_{ij}(c_{ij}-1)$, where c_{ij} can be calculated using equation (5.5). Summing over all subjects and clusters and dividing by all possible number of ordered triples, $\frac{N!}{(N-3)!} = N(N-1)(N-2)$, the following estimate for π_{cc} can be obtained:

$$\hat{\pi}_{cc} = \frac{1}{N(N-1)(N-2)} \sum_{j=1}^{J} \sum_{i=1}^{n_j} c_{ij}(c_{ij}-1).$$

Similarly, π_{dd} and π_{cd} can be estimated as

$$\begin{aligned}\hat{\pi}_{dd} &= \frac{1}{N(N-1)(N-2)}\sum_{j=1}^{J}\sum_{i=1}^{n_j} d_{ij}(d_{ij}-1) \text{ and} \\ \hat{\pi}_{cd} &= \frac{1}{N(N-1)(N-2)}\sum_{j=1}^{J}\sum_{i=1}^{n_j} c_{ij}d_{ij},\end{aligned}$$

respectively. Therefore, the estimate of $\text{var}(\rho)$ can be written as

$$\widehat{\text{var}}(\rho) = \frac{4}{(\hat{\pi}_c+\hat{\pi}_d)^4}(\hat{\pi}_d^2\hat{\pi}_{cc} - 2\hat{\pi}_c\hat{\pi}_d\hat{\pi}_{cd} + \hat{\pi}_c^2\hat{\pi}_{dd}).$$

Finally, the confidence interval for C_{re} can be constructed as:

$$\hat{C}_{re} \pm z_{\alpha/2}\sqrt{\frac{\widehat{\text{var}}(\rho)}{N}},$$

where $z_{\alpha/2}$ denotes the $\alpha/2$ percentile of the standard normal distribution.

5.2.4.2 Gonen and Heller's $K(\beta)$

Gonen and Heller's $K(\boldsymbol{\beta})$ [48] is also an estimator of concordance probability under the Cox PH model (see, Chapter 3). In this chapter, the method of Gonen and Heller [48] is adapted to derive a concordance probability estimator $K_{re}(\boldsymbol{\beta}|\omega)$ for the PH frailty model. $K_{re}(\boldsymbol{\beta}|\omega)$ is a function of the regression parameters, the predictor distribution, and the frailty parameter.

For a pair of subjects (i,k) from clusters (j,l) respectively with corresponding prognostic indices (log hazards) $\{\boldsymbol{\beta}^T\boldsymbol{x}_{ij}+\varpi_j, \boldsymbol{\beta}^T\boldsymbol{x}_{kl}+\varpi_l\}$, the concordance probability

$$K_{re}(\boldsymbol{\beta}|\omega) = \Pr\Big[t_{kl>t_{ij}}|(\boldsymbol{\beta}^T\boldsymbol{x}_{ij}+\varpi_j) > (\boldsymbol{\beta}^T\boldsymbol{x}_{kl}+\varpi_l)\Big]$$

can be calculated for the PH frailty model as

$$\begin{aligned} K_{re}(\boldsymbol{\beta}|\omega) &= \Pr\Big[T(\boldsymbol{\beta}^T\boldsymbol{x}_{kl}+\varpi_l) > T(\boldsymbol{\beta}^T\boldsymbol{x}_{ij}+\varpi_j)\Big] \\ &= \int_0^\infty S(t|\boldsymbol{x}_{kl},\omega_l)dS(t|\boldsymbol{x}_{ij},\omega_j) \\ &= \frac{1}{1+\exp[\boldsymbol{\beta}^T(\boldsymbol{x}_{kl}-\boldsymbol{x}_{ij})+(\varpi_l-\varpi_j)]}, \end{aligned}$$

where $T(\boldsymbol{\beta}^T\boldsymbol{x}_{ij}+\varpi_j)$ represents the survival time that corresponds to $\boldsymbol{\beta}^T\boldsymbol{x}_{ij}+\varpi_j$. If one considers all possible pairs (i,k), $K_{re}(\boldsymbol{\beta}|\omega)$ can be estimated as

$$\begin{aligned} K_{re}(\hat{\boldsymbol{\beta}}|\hat{\omega}) &= \frac{1}{N(N-1)}\sum_{j=1}^{J}\sum_{l=1}^{J}\sum_{i=1}^{n_j}\sum_{k=1}^{n_l}\left[\frac{I\Big((\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_{ij}+\hat{\varpi}_j) > (\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_{kl}+\hat{\varpi}_l)\Big)}{1+\exp[\hat{\boldsymbol{\beta}}^T(\boldsymbol{x}_{kl}-\boldsymbol{x}_{ij})+(\hat{\varpi}_l-\hat{\varpi}_j)]}\right] \\ &= \frac{1}{N(N-1)}\sum_{j=1}^{J}\sum_{l=1}^{J}\sum_{i=1}^{n_j}\sum_{k=1}^{n_l}\left[\frac{I\Big([\hat{\boldsymbol{\beta}}^T(\boldsymbol{x}_{kl}-\boldsymbol{x}_{ij})+(\hat{\varpi}_l-\hat{\varpi}_j)] < 0\Big)}{1+\exp[\hat{\boldsymbol{\beta}}^T(\boldsymbol{x}_{kl}-\boldsymbol{x}_{ij})+(\hat{\varpi}_l-\hat{\varpi}_j)]}\right] \\ &= \frac{1}{N(N-1)}\sum_{j=1}^{J}\sum_{l=1}^{J}\sum_{i=1}^{n_j}\sum_{k=1}^{n_l}\left[\frac{I\Big((\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_{klij}+\hat{\varpi}_{lj}) < 0\Big)}{1+\exp[\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_{klij}+\hat{\varpi}_{lj}]}\right] \end{aligned} \tag{5.9}$$

where $\boldsymbol{x}_{klij}$ and $\hat{\varpi}_{lj}$ represent the differences $\boldsymbol{x}_{kl}-\boldsymbol{x}_{ij}$ and $\hat{\varpi}_l-\hat{\varpi}_j$, respectively. $K_{re}(\hat{\boldsymbol{\beta}}|\hat{\omega})$ is a conditional concordance probability estimator of the PH frailty model as $\hat{\boldsymbol{\beta}}$ is conditional on the frailty ω.

Asymptotic variance of $K_{re}(\boldsymbol{\beta}|\omega)$

The method of Gonen and Heller [48] for the standard $K(\hat{\boldsymbol{\beta}})$ is adapted here to derive an asymptotic variance expression for $K_{re}(\hat{\boldsymbol{\beta}}|\hat{\omega})$. The estimator $K_{re}(\hat{\boldsymbol{\beta}}|\hat{\omega})$ is a non-smooth function of $\hat{\boldsymbol{\beta}}$ and $\hat{\varpi}$. As a result, $K_{re}(\hat{\boldsymbol{\beta}}|\hat{\omega})$ is a nondifferentiable statistic for which it is difficult to obtain a local linear approximation to $K_{re}(\hat{\boldsymbol{\beta}}|\hat{\omega})$, from which the asymptotic distribution of $K_{re}(\hat{\boldsymbol{\beta}}|\hat{\omega})$ and the corresponding asymptotic variance can be derived. To address this problem, a smooth approximation to the above estimator can be obtained following Gonen and Heller [48] who used kernel smoothing technique

[144] as

$$\tilde{K}_{re}(\hat{\boldsymbol{\beta}}|\hat{\omega}) = \frac{1}{N(N-1)} \sum_{j=1}^{J} \sum_{l=1}^{J} \sum_{i=1}^{n_j} \sum_{k=1}^{n_l} \left[\frac{\Phi\Big(-(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_{ijkl} + \hat{\varpi}_{jl})/h \Big)}{1 + \exp[\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_{klij} + \hat{\varpi}_{lj}]} \right], \tag{5.10}$$

where h is the bandwidth which controls the amount of smoothing, and Φ is the standard normal cumulative distribution. Note that for $N \to \infty$, $h \to 0$ and therefore $\Phi(u/h) \to I(u > 0)$. As suggested by Gonen and Heller [48], we choose h so that $Nh^4 \to 0$ as N gets large. Based on this condition, it can be shown that the asymptotic distributions of the non-smoothed estimator $K_{re}(\hat{\boldsymbol{\beta}}|\hat{\omega})$ and the smoothed estimator $\tilde{K}_{re}(\hat{\boldsymbol{\beta}}|\hat{\omega})$ are equal, and the variance of $K_{re}(\hat{\boldsymbol{\beta}}|\hat{\omega})$ can be calculated using a linearisation argument based on the Taylor series expansion for smoothed $\tilde{K}_{re}(\hat{\boldsymbol{\beta}}|\hat{\omega})$. Following Gonen and Heller [48] the bandwidth used in the above approximation is chosen as $h = 0.5\hat{\sigma}N^{-1/3}$, where $\hat{\sigma}$ is the estimated standard deviation of the predicted prognostic index $\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_{ij} + \hat{\varpi}_j$. The term $N^{-1/3}$ confirms the asymptotic condition $Nh^4 \to 0$ required for the asymptotic equivalence of the smooth and non-smooth concordance probability estimator.

The asymptotic variance of $\tilde{K}_{re}(\hat{\boldsymbol{\beta}}|\hat{\omega})$ can be obtained by calculating its first-order Taylor series expansion. Using the results of Gonen and Heller [48] the variance expression can be written as:

$$\text{var}[\tilde{K}_{re}(\hat{\boldsymbol{\beta}}|\hat{\omega})] \approx \text{var}[\tilde{K}_{re}(\boldsymbol{\beta}_0|\omega)] + \left[\frac{\partial \tilde{K}_{re}(\boldsymbol{\beta}|\omega)}{\partial \boldsymbol{\beta}} \right]^{\mathrm{T}} \Bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \text{var}(\hat{\boldsymbol{\beta}}|\hat{\omega}) \left[\frac{\partial \tilde{K}_{re}(\boldsymbol{\beta}|\omega)}{\partial \boldsymbol{\beta}} \right] \Bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \tag{5.11}$$

which can be estimated by plugging in the estimates of the various components of this expansion. The variance of $\hat{\boldsymbol{\beta}}|\hat{\omega}$ can be computed from the inverse of the Fisher information matrix. The asymptotic variance of $\tilde{K}_{re}(\hat{\boldsymbol{\beta}}_0|\hat{\omega})$ can be estimated from data based on the U-statistic formulation. The U-statistic is a class of statistics in statistical estimation theory that produces minimum variance unbiased estimator, for more details

see Hoeffding [145] and Lee [146]. The resulting estimate is:

$$\widehat{\text{var}}[\tilde{K}_{re}(\boldsymbol{\beta}_0|\omega)] = \frac{1}{\{N(N-1)\}^2}\sum_{j=1}^{J}\sum_{l=1}^{J}\sum_{i=1}^{n_j}\sum_{k=1}^{n_l}\sum_{r\neq k}^{n_l}[\upsilon_{ijkl} - \tilde{K}(\hat{\boldsymbol{\beta}}|\hat{\omega})][\upsilon_{ijrl} - \tilde{K}(\hat{\boldsymbol{\beta}}|\hat{\omega})],$$

where $\upsilon_{ijkl} = \Phi\Big(-(\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_{ijkl}+\hat{\varpi}_{jl})/h\Big)\Big[1+\exp[\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_{klij}+\hat{\varpi}_{lj}]\Big]^{-1}$. The partial derivative vector $\dfrac{\partial\tilde{K}_{re}(\boldsymbol{\beta}|\omega)}{\partial\boldsymbol{\beta}}$ can be estimated at $\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}$ and is given by

$$\begin{aligned}\left.\frac{\partial\tilde{K}_{re}(\boldsymbol{\beta}|\omega)}{\partial\boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = \sum_{j=1}^{J}\sum_{l=1}^{J}\sum_{i=1}^{n_j}\sum_{k=1}^{n_l}\sum_{r\neq k}^{n_l}\Bigg[&\frac{\phi\Big(-(\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_{klij}+\hat{\varpi}_{lj})/h\Big)}{1+\exp[\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_{klij}+\hat{\varpi}_{lj}]}[-\boldsymbol{x}_{klij}/h]\\ &+\frac{\Phi\Big(-(\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_{klij}+\hat{\varpi}_{lj})/h\Big)}{\Big(1+\exp[\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_{klij}+\hat{\varpi}_{lj}]\Big)^2}\exp[\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_{klij}+\hat{\varpi}_{lj}][-\boldsymbol{x}_{klij}]\Bigg],\end{aligned}$$

where ϕ is the normal density. For notational convenience, K_{re} instead of $\tilde{K}_{re}(\hat{\boldsymbol{\beta}}|\hat{\omega})$ is used in the rest of the chapter.

5.2.4.3 Royston and Sauerbrei's D

The D statistic quantifies the separation between subjects with low and high predicted risks, as predicted by the model. The D statistic for the frailty model, D_{re}, can be obtained by transforming the prognostic index $\hat{\eta}_{re} = \hat{\boldsymbol{\beta}}^T\boldsymbol{x}_{ij} + \hat{\varpi}_j$ to a normal order statistic z in a similar way to that described for the standard Cox model and then fitting a PH frailty model with z as the only predictor. The resulting model takes the following form:

$$h(t|z,\omega_j) = \omega_j h_0(t)\exp(\beta_z z),$$

where $D_{re} = \hat{\beta}_z$ and has the same interpretation to the standard D statistic. One can also obtain D_{re} by fitting a standard Cox model with z, since the frailties are already included in z.

5.2.4.4 Calibration slope

The calibration slope for the PH frailty model, denoted by CS_{re}, can be obtained by fitting a PH frailty model (or a standard Cox model) with the estimated prognostic index $\hat{\eta}_{re}$ obtained for validation sample as the only predictor:

$$h(t|\hat{\eta}_{re}, \omega_j) = \omega_j h_0(t) \exp(\beta_{\eta_{re}} \hat{\eta}_{re}),$$

where $CS_{re}(= \hat{\beta}_{\eta_{re}})$ is the coefficient of $\hat{\eta}_{re}$ in the above frailty model, and has the same interpretation to the standard calibration slope.

5.2.4.5 Brier score

The Brier score for the frailty risk model can be calculated by comparing the predicted survival probabilities $\hat{S}(t|\boldsymbol{x}_{ij}, \hat{\omega})$ with the observed outcomes at time t over the study period and averaging over the N subjects. Let $Y_{ij}(t)$ be the observed outcome that takes value 1 if the ith subject from the jth cluster is alive at t, and 0 if not. If a subject is alive at time t, the predicted survival probability should ideally be close to 1, otherwise it should be close to 0. The Brier score can be estimated as

$$BS_{re}(t) = \frac{1}{N} \sum_{j=1}^{J} \sum_{i=1}^{n_j} \Big(Y_{ij}(t) - \hat{S}(t_{ij}|\boldsymbol{x}_{ij}, \hat{\omega}_j) \Big)^2 W(t, \hat{G}). \tag{5.12}$$

The weights $W(t, \hat{G})$ are to compensate for earlier censoring and are given by

$$W(t, \hat{G}) = \frac{1\{t_{ij} \leq t\}\delta_{ij}}{\hat{G}(t_{ij})} + \frac{1\{t_{ij} > t\}}{\hat{G}(t)},$$

where $\hat{G}(t)$ is the Kaplan-Meier estimate of the probability of being uncensored at time t. The corresponding integrated Brier score (IBS) is the cumulative Brier score over the interval $[0, \tau]$ and can be calculated for the PH frailty model as

$$IBS_{re}(\tau) = \int_0^{\tau} BS_{re}(t) dW(t), \tag{5.13}$$

where $W(t)$ is a function to weight the contribution of the Brier score at individual time point, and τ is chosen as a time before the last event time.

5.2.5 Estimation: Pooled cluster-specific measures

The estimation of the pooled cluster-specific measure for the frailty model is similar to that discussed for the random-intercept logistic model. For example, the pooled cluster-specific Harrell's C-index is calculated as follows.

Let $\hat{c}_j$ be the estimate of the C-index for the jth cluster with its estimated variance s_j^2 $(j = 1, \ldots, J)$, and τ^2 be the between cluster variance. Then the pooled C-index can be obtained as

$$\hat{C}_w = \bar{w}^{-1} \sum_{j=1}^{J} \hat{c}_j \hat{w}_j$$

where $\hat{w}_j = 1/(s_j^2 + \hat{\tau}^2)$, $\bar{w} = \sum_{j=1}^{J} \hat{w}_j$, and τ^2 can be estimated using the method of DerSimonian and Laird [113], which was described in Section 4.3.5, Chapter 4.

Similarly, pooled estimates for $K(\beta)$, D-statistic, calibration slope, and integrated Brier score (IBS) can be obtained in a similar manner and the resulting measures are denoted by K_w, D_w, CS_w, and IBS_w, respectively. Since analytical standard errors are not available for IBS, bootstrap based standard errors obtained (from 200 bootstrap samples) for each of the clusters are used to calculate the pooled estimate of IBS_w. All these 'pooled' validation measures have the same interpretation to the analogous versions of the 'pooled' validation measures for clustered binary data.

5.3 Application to child mortality data

In this section, the validation measures described above are illustrated using data on mortality of children under the age of five in Bangladesh. The following sections describe the data and present the analysis and results.

5.3.1 Child mortality data

Data on the mortality of children under five in Bangladesh were obtained from the database of the Bangladesh Demographic and Health Survey (BDHS) [147, 148]. The BDHS is a nationally representative survey that has been carried out once every two years since 1993 as part of the world-wide Demographic and Health Survey (DHS) programme, which has been carried out mostly in developing countries. This survey collects information on the reproductive history of women and their socio-demographic and health status, immunisation, and child mortality. Although the women were interviewed at one time point, they were asked to give information on predictors at the time of the event or before the event occurred as approximately. Therefore, although this is a survey design, it can be considered as a retrospective cohort study. The strategic objective of this survey is to improve the collection and use of data by host countries for programme monitoring and evaluation and for policy development decisions. Here the aim is to develop a risk model using data on mortality of children under five. The model may be useful to the country's health care providers to offer advice to women who are planning pregnancy and to identify high risk groups. Data used for this analysis were extracted from the 2004 and 2007 BDHS databases. Data collected in 2004 were used to develop the risk model and a 'temporal validation' of the model was conducted using data collected in 2007.

In each of the both 2004 and 2007 surveys, a total of 361 clusters were selected according to the country's geographical locations. For more details, see BDHS reports [147, 148]. Figure 5.1 shows the geographical location of urban and rural clusters across the country. From each cluster, 30 households, on average, were selected using an equal probability systematic sampling scheme. All married women age 10-49 in the selected households were interviewed to collect information on the survival history for each birth along with relevant background information. For this analysis, only singleton births that occurred in the 5 years preceding the interview were selected. Clustering was considered at only the cluster/geographical-location level, and one birth per household was randomly selected to avoid clustering of children at the household level. The risk model was developed using the data on 6,776 singleton births (with 440 events/deaths) collected in 2004, and the model was validated using data on 6,052 singleton births

(with 325 events/deaths) collected in 2007. In both datasets, survival times of more than 90% of the children were reported to be censored.

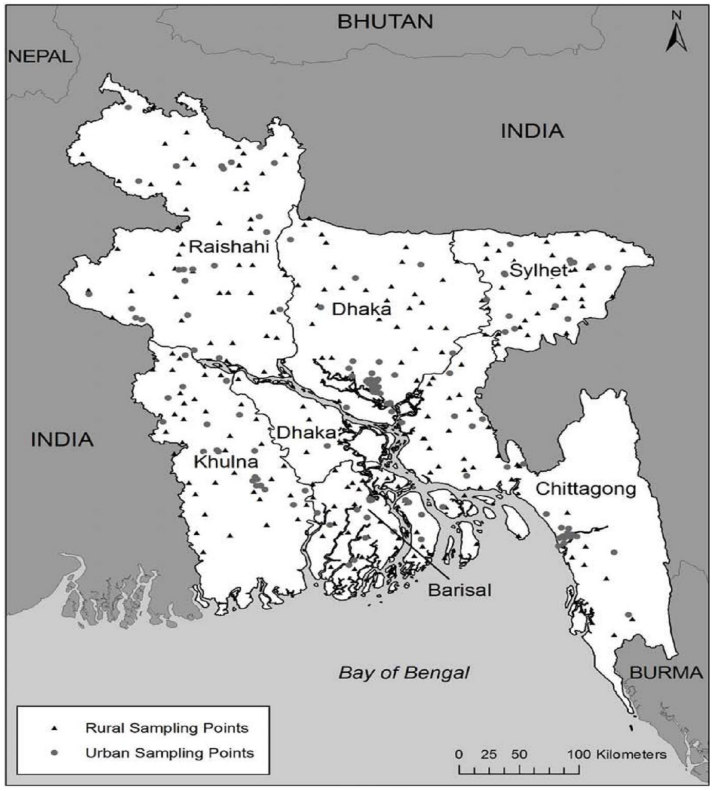

Figure 5.1: Map of Bangladesh indicating the distribution of the urban and rural sampling points (a total of 361 clusters), visited in the 2007 BDHS survey. Source: 2007 BDHS report.

The distribution of the clusters by the number of births per cluster (cluster size) and by the number of deaths per cluster is presented in Figure 5.2. For both the development and validation data, the distributions of the clusters by the number of births per cluster were approximately the same, with the median number of births per cluster reported as 20 (IQR: 16 to 25) and 18 (IQR: 14 to 23) during the period of 1999-2004 and 2002-2007, respectively. Similarly, the distribution of the clusters by the number of deaths per cluster for both datasets were similar, with both average deaths per cluster reported as 1.2 during the period of 1999-2007. From the total of 361 clusters, 132 clusters in the development data and 179 clusters in the validation

data had no deaths. The reason for similarities between these two datasets may be due to the fact that both were sampled from the same population of clusters, where clusters represent the geographical location.

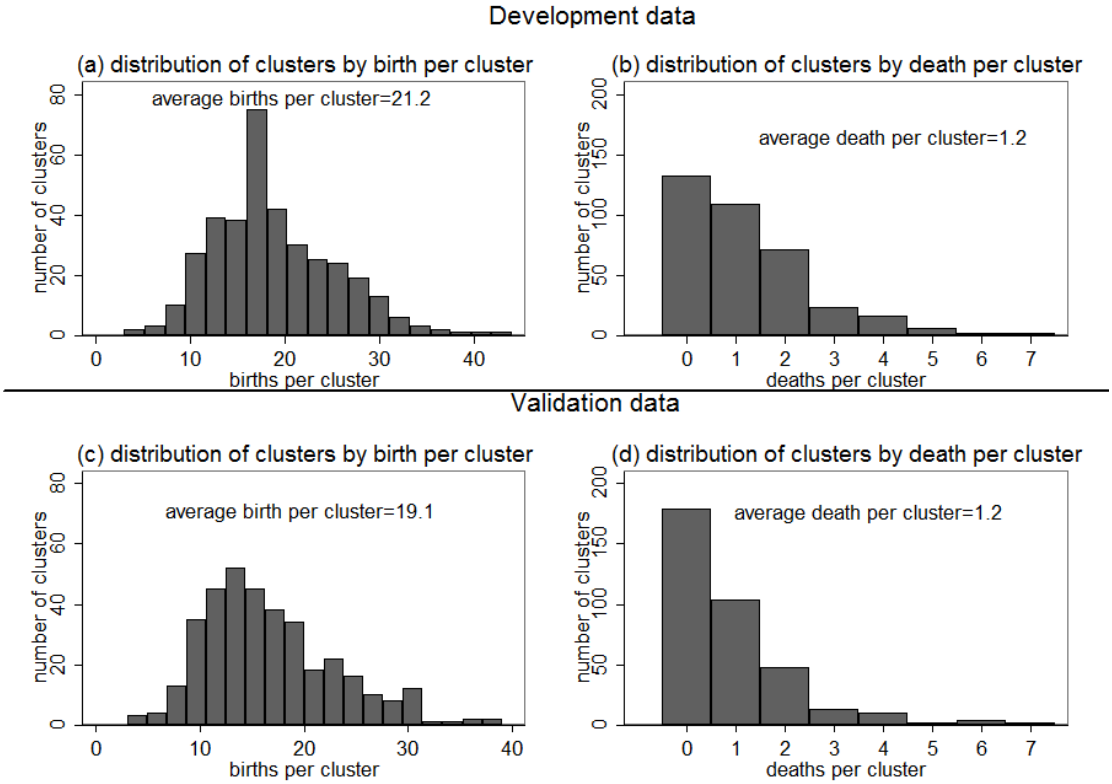

Figure 5.2: Distribution of clusters in both the development and validation data by the number of births per cluster (a,c) and by the number of deaths per cluster (b,d).

The outcome time-to-event (death/survived) was measured in days and was calculated for the births in the 5 years preceding the interview by subtracting the date of birth from the date of death or from the date of interview. The median follow-up times for the children in the development and validation datasets were 870 days and 900 days, respectively. The predictors included in the risk model were maternal age, mother's education, household's socio-economic status, child's birth order, and birth spacing which included both preceding and subsequent birth intervals. These predictors were found to be significantly associated with child mortality in previous studies in this area [149–155].

Maternal age was measured as the age (in years) of the mother at the birth of the child and had a non-linear relationship with outcome. A log transformation was unable to make the relationship linear and therefore maternal age was categorised, as in the BDHS report, as 14-19, 20-29, and 30+ years. Mother's education was categorized based on the number years of schooling the mother had attained: no-education (0 year of schooling), primary (5 years), secondary (10 years), and higher (11+ years). Household socio-economic status (poorest/poorer/middle/richer/ richest) was determined by calculating a wealth index for each household using a principal component analysis of the assets owned (yes/no) by the household. The 1st quintile of the index was referred to as the 'poorest' and the 5th quintile as the 'richest'.

The predictors based on child's birth order and birth spacing were categorized in a similar way to that described in previous studies [149, 154, 155]. Child's birth order was categorised as first birth, order 2-4, and order 5+. The preceding birth interval was categorised as short ($\leq$20 months), medium (21-36 months), long (37+ months), following the first birth. Similarly, the subsequent birth interval was categorised as short, medium, long, following the last birth. Since the information on the first birth is similar in both 'child's birth order' and 'preceding birth interval', these two predictors were combined together to create a single predictor defined as 'birth-order/preceding-birth-interval'. The categories of this combined predictor was defined as first birth, order 2-4/short, order 2-4/medium, order 2-4/long, order 5+/short, order 5+/medium, order 5+/long.

5.3.2 Analysis and results

5.3.2.1 Model development

Using the Cox PH model with shared gamma frailty parameters, a prognostic model of child mortality was developed using the development data. The model parameters ($\boldsymbol{\beta}$, θ) were estimated using penalised likelihood estimation [133]. The Stata package `stcox` using the `shared` option was used to fit the model. The results are presented in Table 5.1. All the predictors in the model were found to be statistically significant at the 5% level of significance. The predictor subsequent-birth-interval showed the strongest association with the outcome. The frailty parameter θ is estimated as 0.11,

which corresponds to the intra-cluster correlation calculated as $\theta/(2+\theta) = 0.05$. This suggests that the effect of clustering is weak; there is a low variation between the clusters in the risk of failures.

Table 5.1: Estimates of the PH frailty model in the development data

Variables	HR	95% CI	P-value
Maternal Age			<0.01
14-19 yrs	1.00	-	
20-29 yrs	1.01	[0.77, 1.32]	
30+ yrs	1.72	[1.15, 2.56]	
Mother's education			<0.05
no education	1.00	-	
primary	0.70	[0.55, 0.89]	
secondary	0.75	[0.56, 1.01]	
higher	0.57	[0.31, 1.04]	
Birth order/preceding birth interval			<0.01
first birth	1.91	[1.39, 2.62]	
2-4/short	1.77	[1.15, 2.72]	
2-4/medium	1.41	[1.04, 1.93]	
2-4/long	1.00	-	
5+/short	1.74	[1.03, 2.92]	
5+/medium	1.28	[0.85, 1.94]	
5+/long	1.06	[0.68, 1.64]	
Subsequent birth interval			<0.001
short	1.00	-	
medium	0.29	[0.22, 0.39]	
long	0.21	[0.13, 0.33]	
last birth	0.12	[0.09, 0.16]	
Socio-economic status			<0.05
poorest	1.00	-	
poorer	0.74	[0.55, 0.99]	
middle	1.02	[0.77, 1.34]	
richer	0.68	[0.48, 0.94]	
richest	0.74	[0.53, 1.04]	
Variance parameter θ (SE)	0.11 (0.07)	-	-

5.3.2.2 Model Validation

The model was then used to predict the survival probability in the validation data, and the predictive performance of the model was assessed using the validation measures described in Section 5.2. To calculate the validation measures, the frailties, the survival probabilities $S(t|\omega)$, and the associated prognostic indices were estimated in

the validation data, using the estimated model parameters from the development data. User written Stata code was used to calculate the validation measures (Appendix B: Figure B.3).

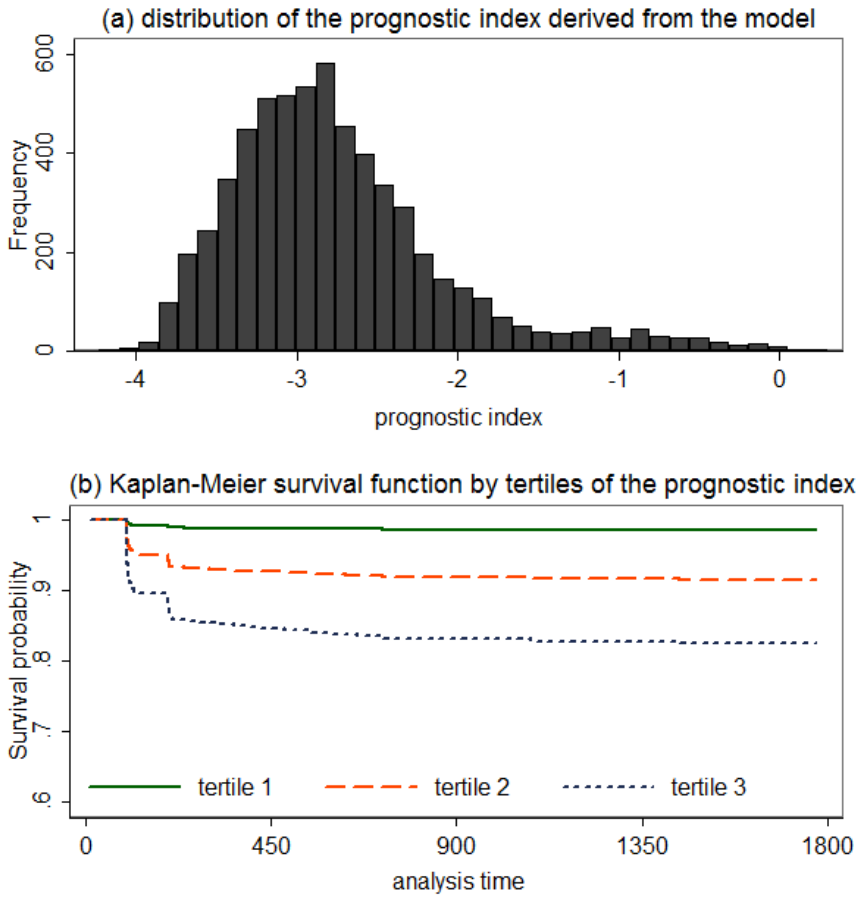

Figure 5.3: (a) Distribution of the predicted prognostic index $\hat{\eta}_{re} = \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_{ij} + \hat{\varpi}_j$ (b) Kaplan-Meier survival function at the tertiles of the predicted prognostic index.

Some of the validation measures are based on a normality assumption of the predicted prognostic index. Therefore, the distribution of the predicted prognostic index is presented in Figure 5.3(a); there appears to be some skewness towards the right, however it is perhaps not unreasonable to consider the distribution of the predicted prognostic index as normal. Furthermore, to examine the spread in survival predic-

tions as predicted by the model in the validation data, Kaplan-Meier survival function at the tertiles of the predicted prognostic index is presented in Figure 5.3(b). This plot suggests that the model had good ability to separate (discriminate) children with low risk of mortality from those with high risk.

The estimates of the validation measures are presented in Table 5.2. The concordance statistic C_{re} was estimated as 0.750 (95% CI:0.723 to 0.785), which suggests that the model has reasonably good ability to discriminate between low and high risk children. However, K_{re} suggests relatively a lower ability for discrimination, with an estimate of 0.686 (95% CI: 0.669 to 0.701). The possible reason for this difference is similar to that for the standard C-index that C_{re} may be affected by the high degree of censoring in the child mortality data. The estimate of the D statistic (D_{re}) also suggests moderate discrimination between low and high risk children. Similar to the standard D and K statistics, both D_{re} and K_{re} may not be affected by the censoring in this dataset. The calibration slope (CS_{re}) suggests that the model has good overall calibration. The extent of inaccuracy in the individual survival prediction (IBS_{re}) was estimated to be 0.07, suggesting reasonably lower inaccuracy in survival predictions. Additionally, the Brier score calculated at each time point was plotted against the observed time points in Figure 5.4, to see the model's predictive accuracy over the entire follow-up period. Two additional plots of the Brier score for the null model and the model with all fixed predictors only (frailties set to one) were obtained, to examine the difference in predictive accuracy between these three models. The survival prediction error (the Brier score) was lower when the predictions were made by the model with all the fixed predictors along with the frailties compared to those obtained from the model with only fixed predictors and the null model.

The 'pooled' estimate of the cluster-specific C-indices, C_w, indicates a good discrimination between low and high risk children belonging to the same cluster, whereas K_w indicates relatively less discrimination, and D_w suggested poor discrimination between these two groups (Table 5.2). However, these 'pooled' estimates are smaller than their corresponding 'overall' estimates, although the frailty effects in these data were not that strong. Similarly the 'pooled' estimate of the cluster-specific calibration slopes (CS_w)

Table 5.2: Estimates of the validation measures based on the validation data

		Overall measures	
Measures' name	Notations	Estimate	95% CI
Harrell's C-index	C_{re}	0.750	[0.723, 0.785]
Gonen and Heller's $K(\beta)$	K_{re}	0.686	[0.669, 0.701]
D-statistics	D_{re}	1.52	[1.29, 1.74]
Calibration slope	CS_{re}	1.01	[0.91, 1.09]
Integrated Brier score	IBS_{re}	0.07	[-]
		Pooled cluster-specific measures	
Harrell's C-index	C_w	0.701	[0.671,0.742]
Gonen and Heller's $K(\beta)$	K_w	0.649	[0.626, 0.679]
D-statistics	D_w	0.89	[0.65, 1.13]
Calibration slope	CS_w	0.64	[0.48, 0.81]
Integrated Brier score	IBS_w	0.06	[0.05, 0.07]

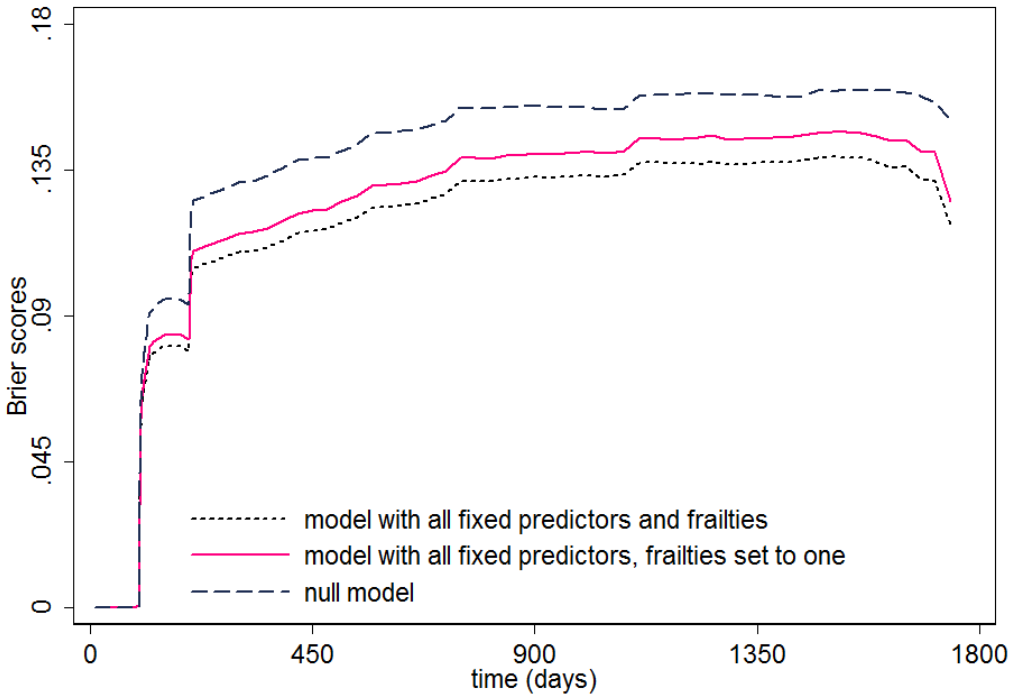

Figure 5.4: Brier scores over the entire follow-up period. The results are obtained for the predictions from the model with all fixed predictors and the frailties, the model with all fixed predictors only, and the null model.

suggests worse calibration than the 'overall' calibration. This difference may be caused by the cluster size of the child mortality data, where clusters are reasonably small and several of the clusters have no events. These clusters were dropped due to the lack of events to calculate the measure and the pooled estimate was based on the reduced

number of clusters. The reason is similar to those described for 'pooled cluster-specific' measures for clustered binary data in pages 98-99, Section 4.5.3.2, Chapter 4. However, the estimate of the 'pooled' Brier score ($IBS_w = 0.06$) was close to that of the 'overall' Brier score.

Since the cluster-specific estimates are useful in detecting outlying clusters (regions), these estimates with their level of uncertainty are plotted against the rank order of the clusters in Figure 5.5. Each horizontal solid line indicates the 'pooled estimate' of the respective measure. Figure 5.5 shows the estimates for 76 clusters only. It was not possible to estimate the validation measures for the rest of the clusters as the required number of events to enable the calculation of the measures were not observed in these clusters. Note that to enable the calculation, some of the measures, for example, C_w requires at least one event, and the other measures, for example, D_w requires two events (see, Section 5.2). This plot made a comparison between clusters (regions) in terms of the model performance. The results shows that the predictive ability of the model for some of the clusters were significantly worse (better) than the average performance. This heterogeneity in the model predictive performance between the clusters (regions) may be caused by unobserved cluster (region) level characteristics. Therefore, it may be important to identify the factor which explains this heterogeneity.

In summary, this illustration of the validation measures using the child mortality data showed that both the 'overall' and 'pooled cluster-specific' measures have a meaningful interpretation in a clustered survival data setting. However, the validation measures appeared to provide different conclusions regarding the model's predictive performance. For example, the C-index suggested strong discrimination, whereas $K(\beta)$ suggested moderate discrimination. Furthermore, while the 'overall' estimates of both the D-statistic and calibration slope indicated a reasonably good predictive performance of the model, their 'pooled' estimate indicated very poor performance. These dissimilarities may be caused by high degree of censoring in the child mortality data and/or by the small clusters. Therefore, in the next section, a simulation study is conducted to evaluate the performance of the measures in a range of conditions based on a clustered survival data setting.

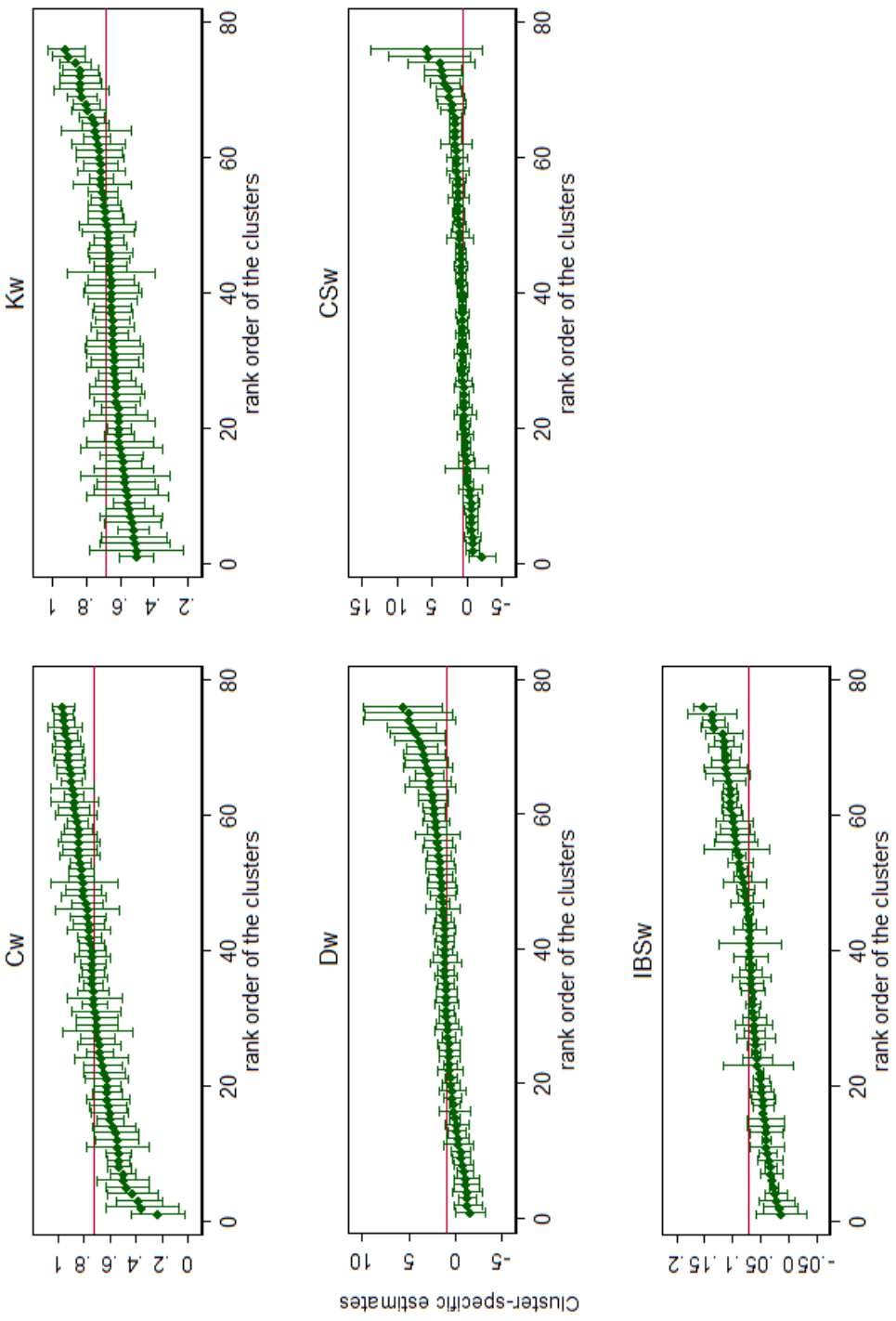

Figure 5.5: Cluster-specific estimates of the validation measures with their level of uncertainty against the rank order of the clusters. The horizontal solid line indicates the pooled estimate of the measure.

5.4 Simulation study

A simulation study was conducted to evaluate the bias, rMSE, and coverage of the estimate of the validation measures. Both development and validation data were simulated. Models were developed using the simulated development data and then evaluated using the corresponding simulated validation data. The properties of the measures were assessed in various clustered survival data scenarios, constructed by varying the number of clusters and their size, the intra-cluster correlation between patients within a cluster, and the degree of censoring in the validation data. In practice, censoring is common in survival data, and some of the validation measures for standard survival models were found to be considerably affected by censoring (see, Chapter 3). In addition, number of clusters and their sizes and the intra-cluster correlation may influence their performance. The validation measures for clustered binary data were found to be affected by small clusters and the level of ICC. The simulation studies would help in identifying which factors affected the performance of the validation measures.

5.4.1 Simulation design

5.4.1.1 True model

To simulate clustered survival data, a PH frailty model with Weibull baseline hazard with shape parameter $\gamma = 1.1$ and scale parameter $\mu = 1$ was chosen as a true model. The frailty distribution was chosen as Gamma with mean 1 and variance θ. For a sample of N subjects with J clusters, the predictor value x_{ij} for the ith subject in the jth cluster was generated from the standard normal distribution ($i = 1, \ldots, n_j; j = 1, \ldots, J$). The clustered survival data were then generated as follows:

(i) the frailty value ω_j for the jth cluster was generated from a Gamma distribution with mean 1 and variance θ.

(ii) the survival times t^*_{ij} were generated as

$$t^*_{ij} = \left(\frac{-\log(v_{ij})}{\mu \exp(\beta x_{ij}) \omega_j} \right)^{1/\gamma}$$

where $v_{ij} \sim U(0, 1)$.

(iii) to generate random right-censored data, a pseudo-random Weibull distributed censoring times c_{ij} were also generated in a similar manner to that used to generate the survival times, but the term $\exp(\beta x_{ij})$ was replaced by a scalar λ. Different choices of λ were used to produce different degrees of censoring.

(iv) the observed survival times were calculated as $t_{ij} = \min(t^*_{ij}, c_{ij})$, and the censoring indicator as $\delta_{ij} = 1$ if $t^*_{ij} \leq c_{ij}$.

The value of the frailty parameter θ was varied to generate data with different levels of clustering, and the regression coefficient β was set to 1.35 for all simulation settings, indicating strong predictor. The value of β was chosen arbitrarily, but the aim was to deal with a model with strong prognostic ability.

5.4.1.2 Simulation scenarios

To create scenarios with no, moderate, and high levels of clustering, the values of the frailty parameter θ were set to 0, 0.58, and 0.98, respectively. For each value of θ, development datasets each with 50 clusters of size 30 were generated without considering any censoring. For each development dataset, validation data from various scenarios were generated, to mimic scenarios with large number of small clusters and small number of large clusters. The validation data were also simulated to have low, moderate, and high degrees of censoring. The validation scenarios considered were: (a) 10 clusters of sizes 10, 30, and 50, and (b) 50 clusters of sizes 10 and 30. For each of these scenarios, four different degrees of censoring: 0%, 20%, 50%, and 80% were considered. This resulted in 20 validation scenarios for each of the three values of the frailty parameter θ, thus 60 altogether. For each of the development and validation scenarios, 500 datasets were generated. Similar to the simulation design discussed in Chapters 3 and 4, this specification (500 simulations) was also determined following Burton et al. [78] and provided very low Monte Carlo error for the validation measures for clustered survival data. The levels of clustering in the development and the corresponding validation datasets were set to equal, by generating both datasets from the same value of θ. This would represent a scenario where clusters in both the development and validation datasets are from the same population of clusters.

5.4.2 Strategies for evaluating the measures

5.4.2.1 Model fitting and calculation of the validation measures

A Cox PH hazard model with shared gamma frailty was fitted to each of the development datasets. The model parameters were estimated using penalised likelihood estimation [133]. These estimates were used to obtain the empirical Bayes estimates of the frailties and the predicted survival probability $\hat{S}(t|\hat{\omega})$ in the corresponding simulated validation datasets. Then the point estimates and confidence intervals for the validation measures were calculated for each of the validation datasets.

5.4.2.2 Assessing the properties of the measures

The effects of censoring, number of clusters and their size, and intra-cluster correlation were investigated by assessing the empirical bias, empirical rMSE, and coverage of nominal 90% confidence intervals for the validation measures. The true values of the validation measures were obtained empirically by averaging over 100 simulations of very large uncensored datasets (N=100,000 with J=500 clusters). In each simulated dataset, the 'overall' validation measures were calculated using the true regression parameter (β) and the frailty values. Since the true value of the frailty parameter θ was varied to create the scenarios with different levels of clustering, true values of the 'overall' measures were calculated as above for each value of θ. However, similar to the measures for clustered binary data, the true value of the rank-based 'pooled' measures (C_w, K_w, and D_w) and the calibration slope CS_w were calculated using the true value of the regression parameter (β) only, because the frailties do not contribute to the calculation of these measures (see Section 5.2.5). The IBS for each cluster however includes the frailties. Thus the true value of the 'pooled' IBS (IBS_w) was calculated using both the true value of the regression parameter (β) and the frailties.

The bias in the estimate of the validation measure was calculated as the mean of the difference between the estimate and the true value, over 500 simulations. Similarly, rMSE was calculated as the square root of the mean of the squared difference between the estimate and the true value. Coverage was calculated as the proportion of simulations where the estimated confidence interval contained the true value. Analytical as

well as bootstrap CIs (with 200 bootstrap samples) were used to calculate coverage. The bias and rMSE were rescaled to a percentage in a similar way to that discussed in Chapter 3 and Chapter 4.

5.4.3 Results

5.4.3.1 The Overall validation measures

The empirical (sampling) distribution of the 'overall' estimates of the validation measures, by different degrees of censoring, is summarised using box plots. Figure 5.6 shows the results for the simulation scenario with 10 clusters of size 100 where clustering was high ($\theta = 0.98$). The horizontal dashed lines show the true values of the measures. The inter-quartile range for each of the validation measures increases with the degree of censoring. The medians for CS_{re}, K_{re}, and D_{re} are approximately close to the true value, suggesting correct inference regarding the model's predictive performance. However, the medians for C_{re} and IBS_{re} increased with increasing degree of censoring, which indicates that misleading conclusions could be drawn regarding the model's predictive performance in the presence of censoring. C_{re} performed adequately for up to 20% censoring, however IBS_{re} was affected even with 20% censoring. Similar results were observed for the other simulation scenarios (not shown). These results are analogous to those for the standard validation measures for independent survival outcomes.

The relative percentage of bias induced by censoring was plotted against the degree of censoring in Figure 5.7. For the uncensored survival simulations, all the validation measures under investigation were approximately unbiased, particularly when cluster sizes were large and there was no clustering. For the censored survival simulations, C_{re} and IBS_{re} showed bias, which increased with increasing degree of censoring. This was the case for all simulation scenarios. The effect of censoring was also observed for D_{re} and CS_{re} particularly when the cluster sizes were small and there was some degree of clustering. Bias in all these cases suggest the possibility of reaching misleading conclusions regarding the model's predictive performance. However, K_{re} was unbiased in the presence of censoring, which was the case for all simulation scenarios. In addition, the validation measures, in general, were affected by non-zero intra-cluster correlation ($\theta > 0$) when clusters were small.

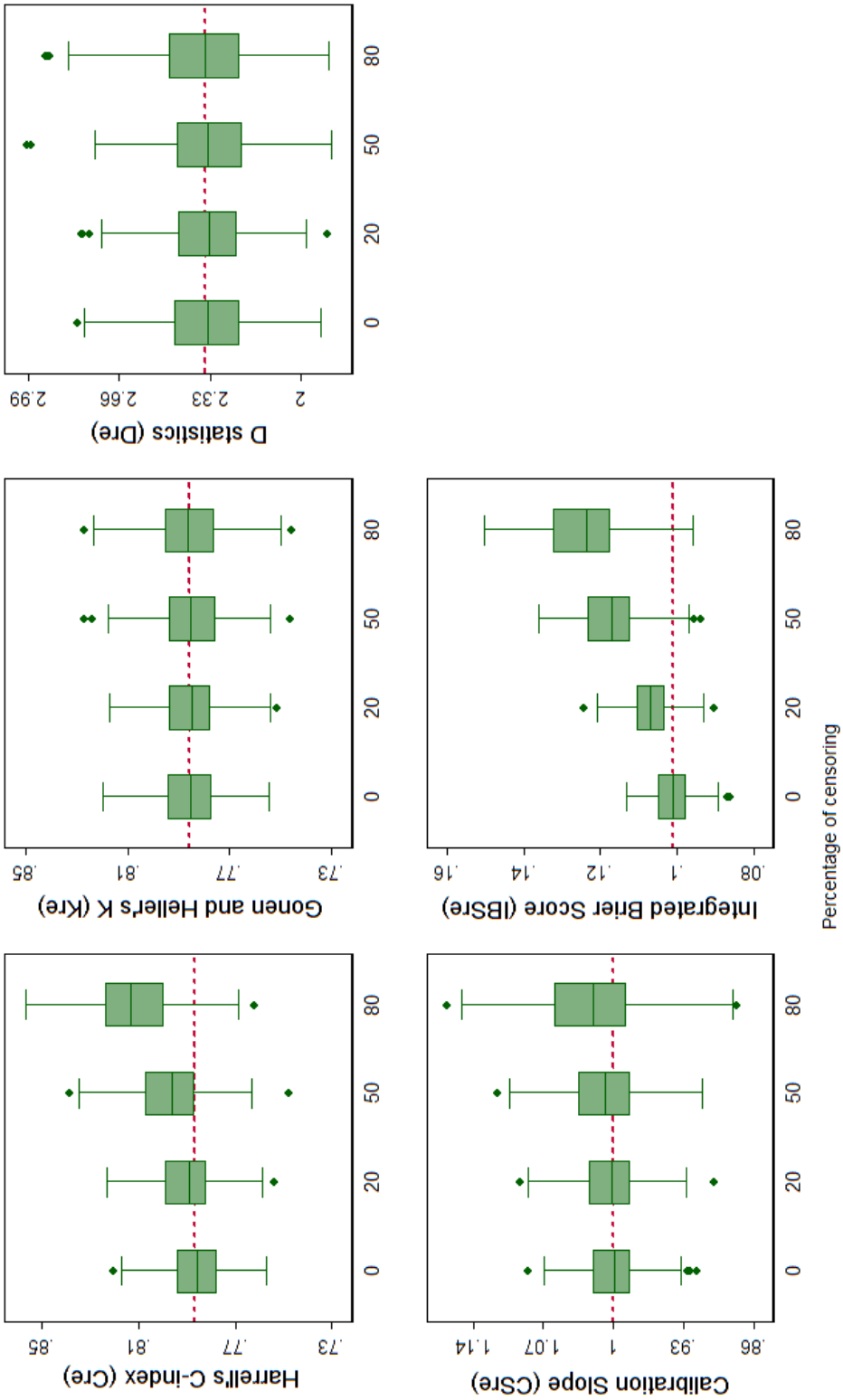

Figure 5.6: Empirical (sampling) distribution of the validation measures by degree of censoring, summarised using box plots. The results are from the simulations with 10 clusters of size 100 under $\theta = 0.98$. The horizontal (dashed) lines indicate the true values of the measures.

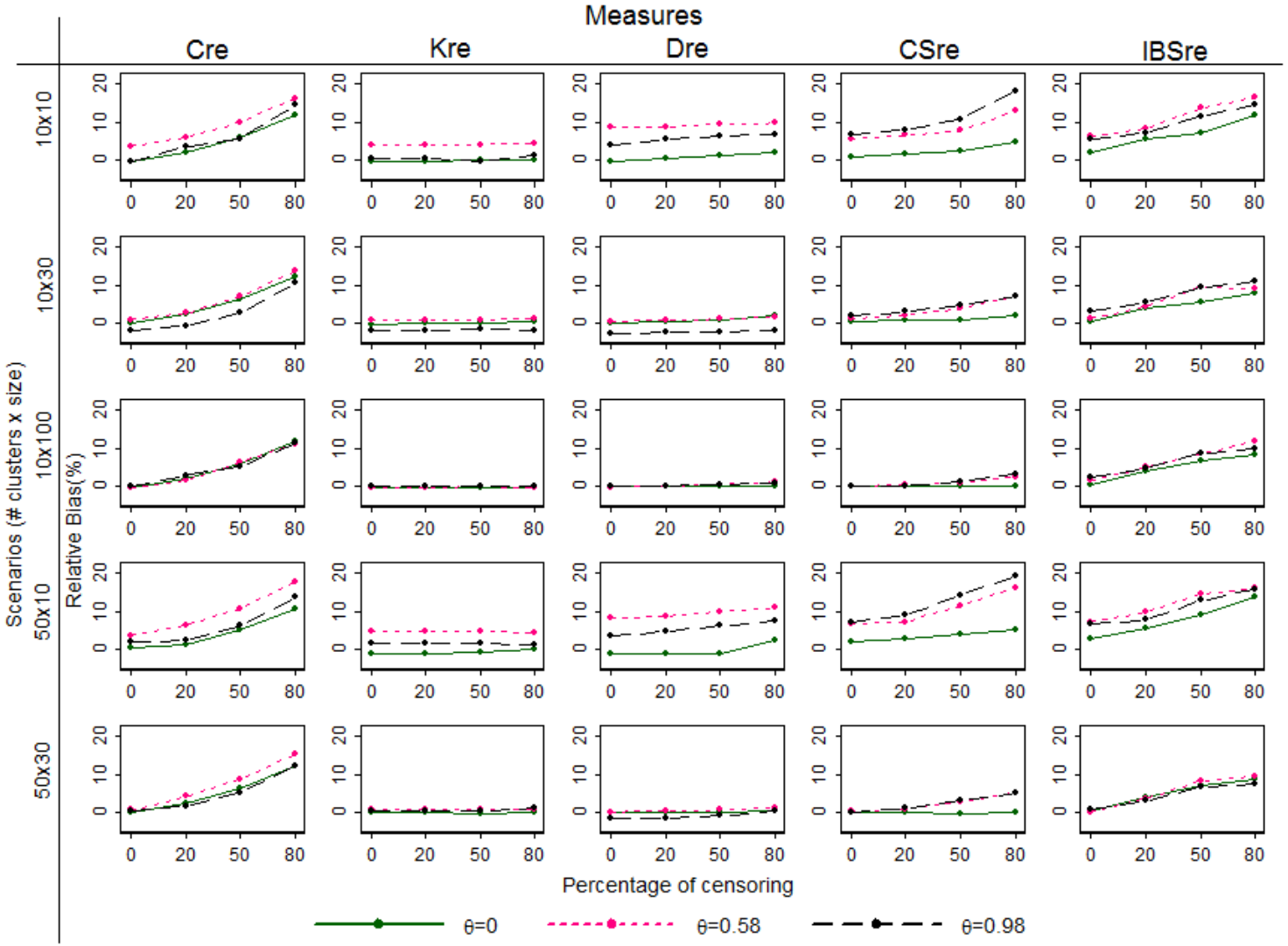

Figure 5.7: Relative bias (%) in the 'overall' estimate of the validation measures for degrees of censoring. The results are from the different simulation scenarios based on the number of clusters, their sizes, and the frailty parameter θ.

The reasons for bias in C_{re} and IBS_{re} induced by censoring are similar to those discussed for the standard C-index and IBS for independent survival data (see Section 3.5.2, Chapter 3). Briefly, these measures are calculated by comparing the predicted survival probability with the observed survival status of the two categories of subjects: those who developed the event and those who did not. For the dataset with large amount of censoring, there are far fewer comparisons than the value what we would obtain if the actual survival times were available, which may result in bias. The possible reason for bias caused by the non-zero intra-cluster correlation when the clusters are small is that the empirical Bayes estimates for the frailties were poorly estimated for small clusters. This reason is analogous to that discussed for validation measures for clustered binary outcomes.

In general, the relative rMSE (%) of the validation measures increased with increasing degree of censoring (Figure 5.8). The increase was sharp for C_{re} and IBS_{re} as their point estimates were biased. However, for the measures whose point estimates were unbiased in the presence of censoring, for example, K_{re}, there was also a steady increase, because of increasing empirical standard error. For all validation measures, the rMSE was low for the scenario with large number of large clusters while it was high for the scenario with small number of small clusters. The rMSE of the measures was also affected by the non-zero intra cluster correlation as their bias affected. Amongst the validation measures, rMSE was lowest for K_{re}, followed by D_{re} and CS_{re}, and it was the highest for C_{re} and IBS_{re}.

Coverage of nominal 90% confidence intervals for the 'overall' validation measures were calculated based on bootstrap standard errors. Table 5.3 presents the results for all simulation scenarios with high clustering. For the uncensored survival simulations with large clusters, coverage for all the validation measures was close to the nominal 90% value. When the clusters were small, the measures showed somewhat poor coverage as the point estimates were biased. For the censored survival simulations, the validation measures which were unbiased had good coverage. Similar results were observed for the other simulation scenarios (not shown). The results regarding coverage for the validation measures were similar to that discussed above when analytical standard errors were used (not shown).

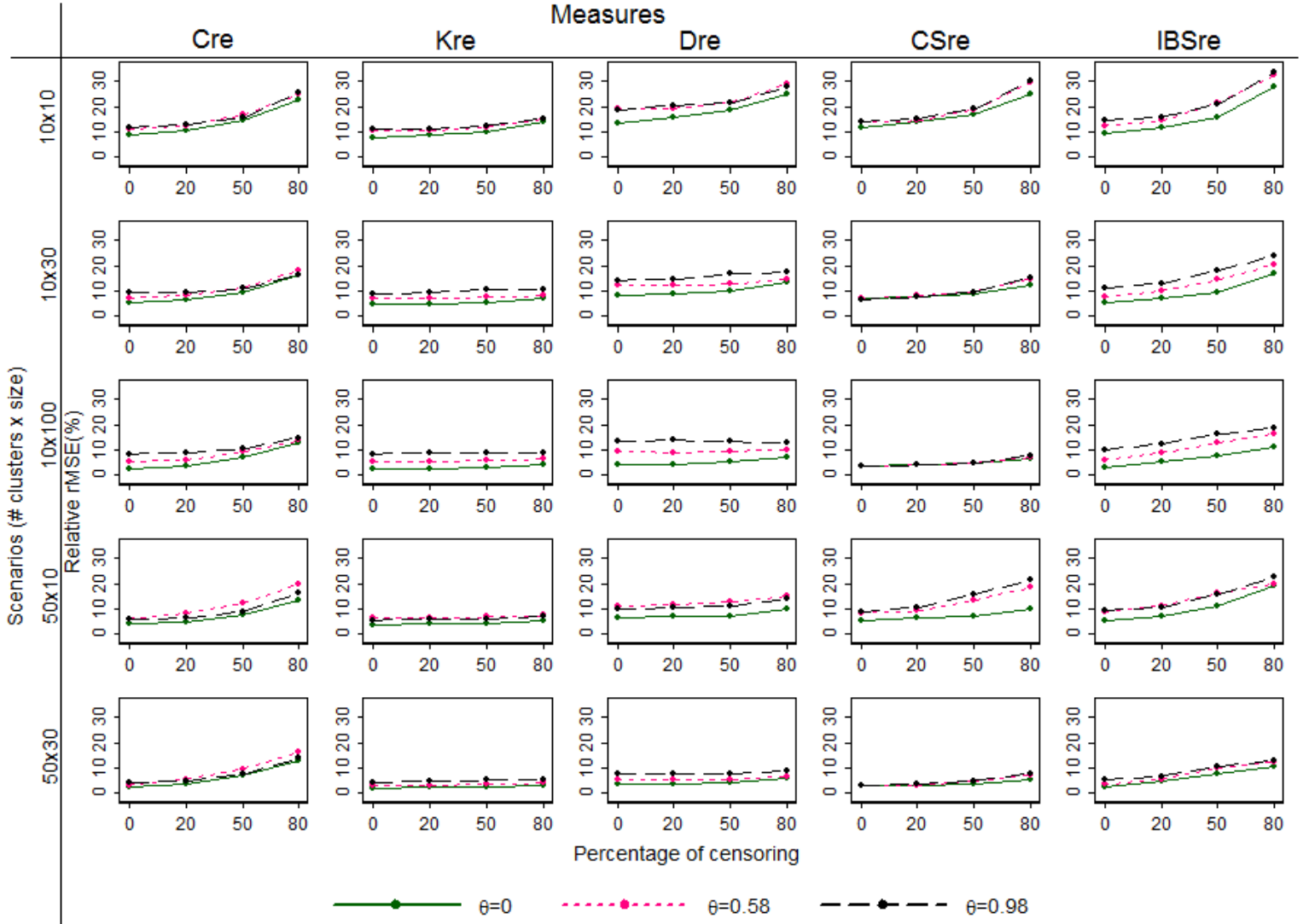

Figure 5.8: Relative rMSE (%) of the 'overall' estimates of the validation measures for different degrees of censoring. The results are from the different simulation scenarios.

Table 5.3: Estimated coverage of nominal 90% confidence intervals for the 'overall' measures. The confidence intervals were calculated based on bootstrap standard errors. The results are from all simulation scenarios where the level of clustering was high ($\theta = 0.98$). Maximum Monte Carlo Standard Error=2.4%.

		Overall measures				
Cluster × size	Censoring	C_{re}	K_{re}	D_{re}	CS_{re}	IBS_{re}
10×10	0	88	87	80	79	81
	20	80	86	78	75	77
	50	63	86	72	68	65
	80	30	84	65	56	32
10×30	0	85	88	83	88	85
	20	80	88	84	87	82
	50	50	87	84	80	53
	80	25	88	86	75	31
10×100	0	91	90	89	91	88
	20	77	91	89	89	82
	50	51	89	88	87	55
	80	28	88	89	84	25
50×10	0	87	86	84	82	85
	20	83	86	80	75	79
	50	50	84	68	55	54
	80	25	86	54	41	29
50×30	0	91	90	87	89	89
	20	70	91	88	89	80
	50	51	90	88	85	43
	80	30	89	87	80	23

The agreement between the validation measures was also investigated by using scatter plot matrix and the Pearson correlation coefficient, r. Figure 5.9 shows the results for the survival simulations with 50 clusters of size 30 where the level of clustering was high. The measures closely agree with each other for the uncensored survival simulations, but the agreement becomes weaker, particularly between C_{re} and IBS_{re}, with higher degrees of censoring. For example, the correlation between C_{re} and IBS_{re} was -0.99 for 0% censoring and reduced to -0.52 for 80% censoring. Similar results were observed for all other simulation scenarios (not shown). This finding is analogous to those for the standard validation measures for independent survival data (Section 4.5.1.3, Chapter 3).

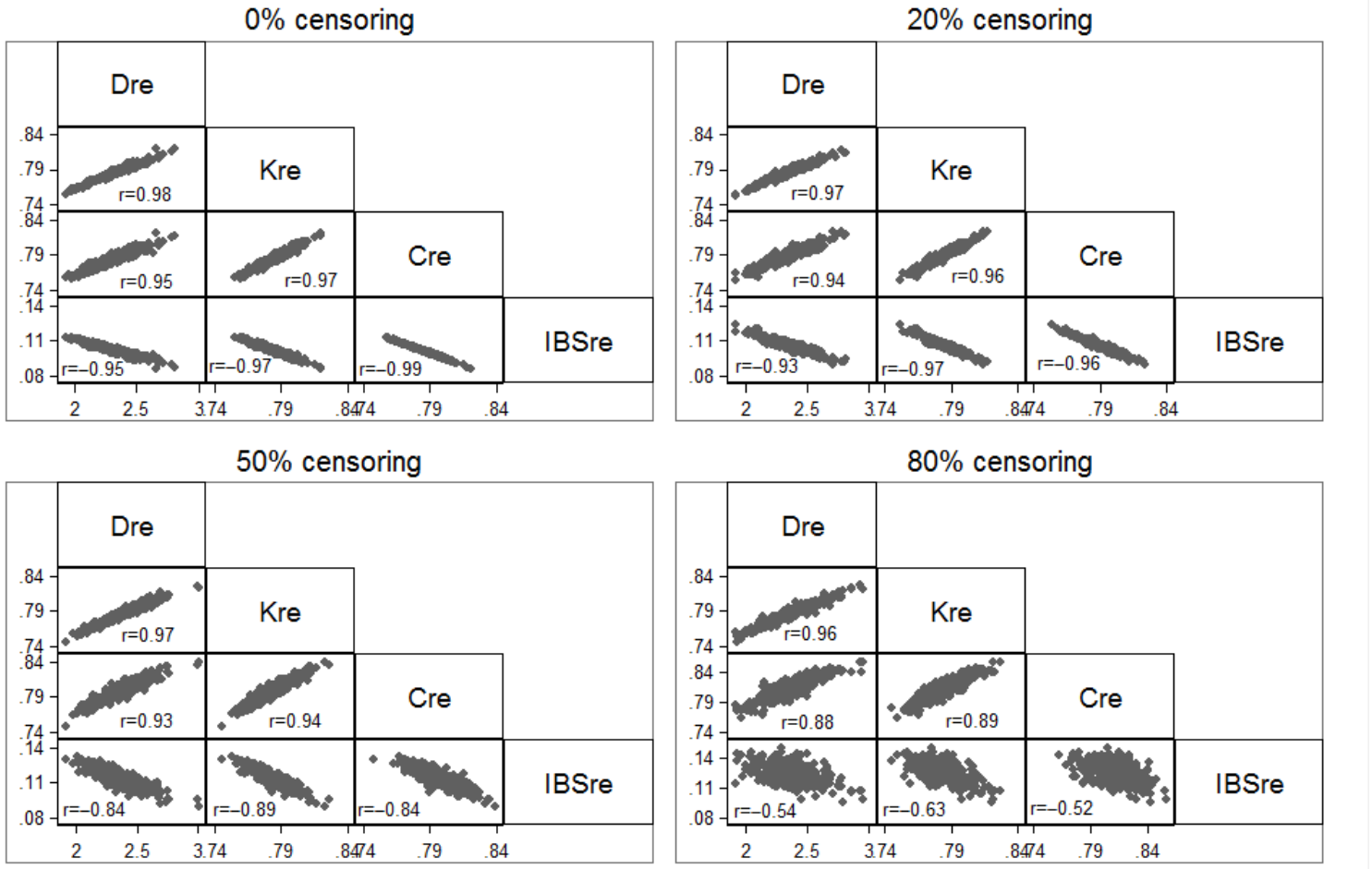

Figure 5.9: Agreement between the validation measures for different degrees of censoring. The results are from the simulations with 50 clusters of size 30 under $\theta = 0.98$. The r values indicate the estimated Pearson correlation coefficients between the measures.

5.4.3.2 Pooled cluster-specific validation measures

The relative bias and rMSE for the 'pooled' estimates of the cluster-specific validation measures were plotted against the degree of censoring in Figures 5.10 and 5.11, respectively. In general, the pooled estimates of the cluster-specific validation measures were affected by the cluster size, with small sizes producing greater bias and higher rMSE. Similar to the 'overall' estimates, K_w, D_w, and CS_w were not affected by censoring when the clusters were large. However, C_w and IBS_w were affected by censoring, even for large clusters. The bias in IBS_w increased with censoring up to 50 percent and then decreased. The probable reason for the decrease is that a very small number of events was observed for each cluster for the simulations with 80 percent censoring, resulting in low values of the Brier score and hence in IBS_w. For example, for the clusters of size 10, the simulations with 80 percent censoring provide just two events on average, and therefore the IBS_w underestimated the true value. Amongst the 'pooled' cluster-specific validation measures, only IBS_w includes frailties and appeared to be affected by the level of clustering when the clusters were small. This is because the frailties were not well estimated for these clusters.

The reason for bias in the 'pooled cluster-specific' validation measures when the cluster sizes are small is similar to those discussed for the measures for clustered binary data (Section 4.5.3.2, Chapter 4). With the simulations with 80 percent censoring, approximately 20% small clusters did not have the required number of events to calculate the measures and were ignored. The pooled estimate was based on the available clusters, which resulted in bias.

Coverage of nominal 90 percent confidence intervals for the 'pooled' cluster specific validation measures were calculated based on analytical standard errors. Table 5.4 presents the results for all simulation scenarios under a high level of clustering ($\theta = 0.98$). For the uncensored survival simulations with large clusters, coverage for all the validation measures was close to the nominal 90% value. For the censored survival simulations, the unbiased validation measures had good coverage. Of these, K_w had best coverage performance.

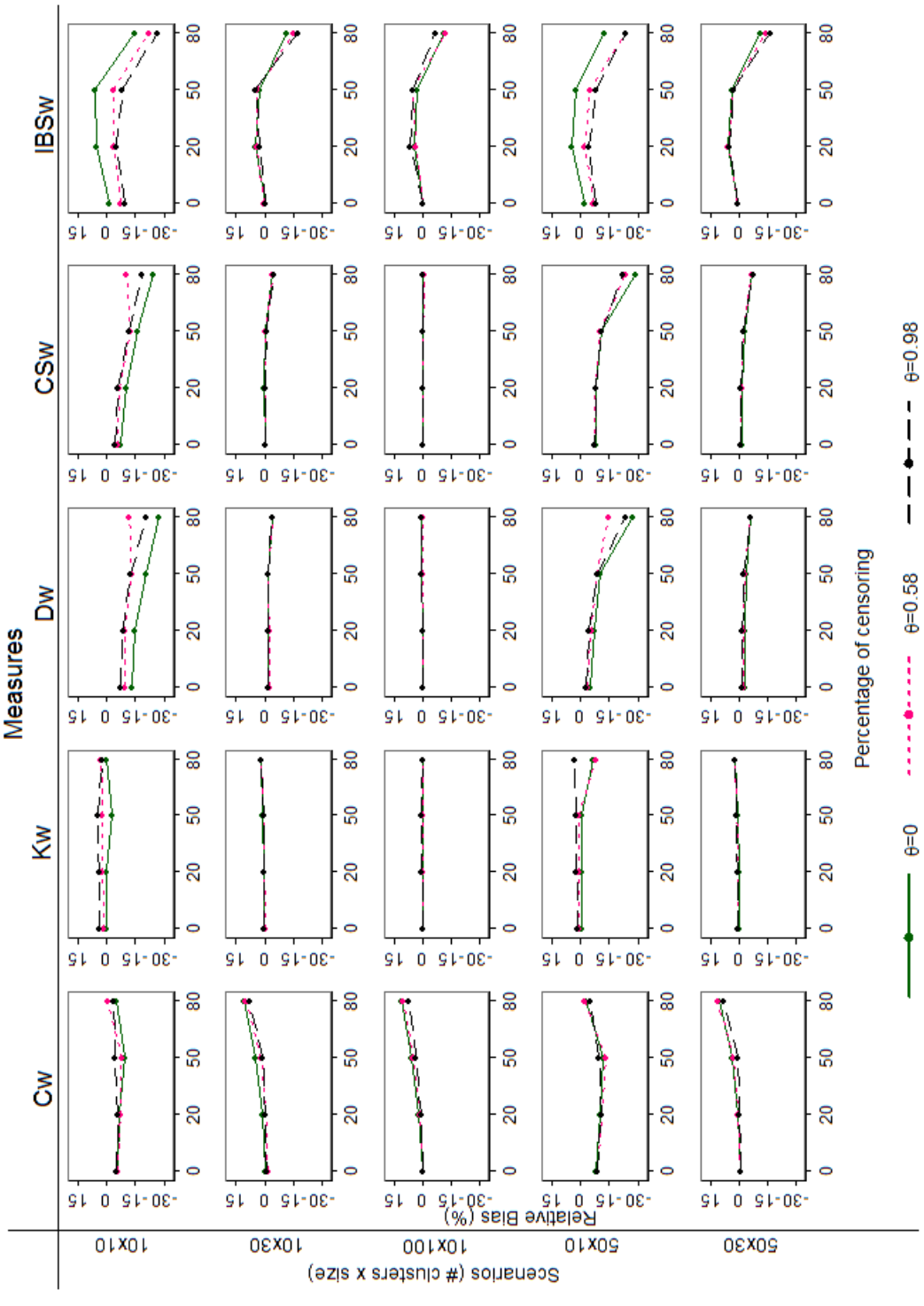

Figure 5.10: Relative bias (%) in the 'pooled estimates' of cluster-specific validation measures for different degrees of censoring. The results are from the different simulation scenarios.

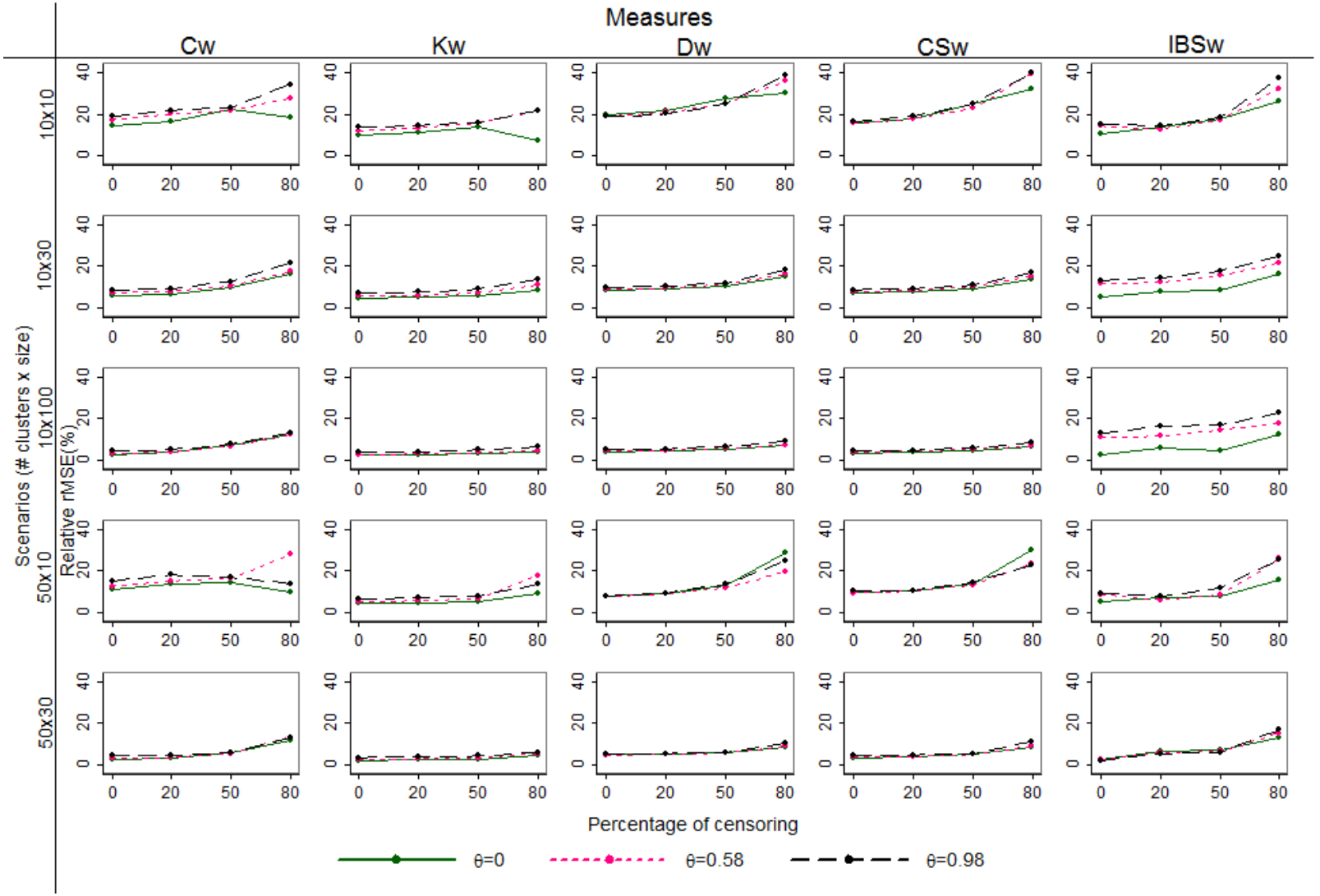

Figure 5.11: Relative rMSE (%) of the 'pooled estimates' of the cluster-specific validation measures for different degrees of censoring. The results are from the different simulation scenarios.

In general, the pooled cluster-specific validation measures had poor coverage for the censored survival simulations with small clusters, as their point estimates were biased. Similar results were observed for the other simulations scenarios (not shown). In addition, the validation measures had similar coverage performance to that discussed above when bootstrap standard errors were used (not shown).

Table 5.4: Coverage of 90% nominal confidence intervals for the 'pooled cluster-specific' measures. The confidence intervals were calculated based on analytical standard errors. The results are from the different simulation scenarios under $\theta = 0.98$. Maximum Monte Carlo Standard Error=2.5%.

		Pooled cluster-specific measures				
Cluster × size	Censoring	C_w	K_w	D_w	CS_w	IBS_w
10×10	0	90	87	87	78	88
	20	86	87	75	86	71
	50	71	83	70	75	72
	80	74	77	65	70	20
10×30	0	88	90	88	91	86
	20	84	90	89	91	84
	50	75	90	89	91	81
	80	63	84	89	88	35
10×100	0	91	90	89	90	89
	20	82	90	90	91	83
	50	55	89	90	90	84
	80	30	89	91	91	41
50×10	0	61	84	42	55	73
	20	54	86	32	49	74
	50	35	88	25	31	73
	80	72	78	10	14	19
50×30	0	91	89	87	89	88
	20	88	90	87	88	81
	50	69	89	84	86	84
	80	32	88	76	75	38

5.5 Conclusion

This chapter has discussed extensions of some of the standard validation measures for use with models for clustered survival data, using the same approach discussed for clustered binary data. This has lead to an 'overall measure' and a 'pooled cluster-specific measure', for each of the standard measures. Each of these approaches have three different definitions based on the model's conditional predictions using the frailties,

$S(t|\omega)$, or setting the frailties at their mean, $S(t|1)$, and the marginal predictions $S(t)$. This chapter has discussed the validation measures for use with $S(t|\omega)$. The validation measures for $S(t|1)$ and $S(t)$ can be derived in a similar manner to that discussed for $S(t|\omega)$.

The illustration of the validation measures using child mortality data from Bangladesh showed that the measures have meaningful interpretations in clustered survival settings. The statistical properties of the measures are also evaluated using simulation studies. The validation measures, in general, behaved similarly as their corresponding standard measures for independent survival data, particularly in the presence of censoring. The 'overall' K statistic (K_{re}) was not affected by censoring. The effect of censoring on the D statistic (D_{re}) was negligible except for the small clusters, which is to be expected since the distribution of the prognostic index was specified as normal. Based on central limit theorem, the prognostic index, in practice, is likely to be normally distributed as the number of predictors in the model increases. The effect of censoring on the calibration slope (CS_{re}) was also negligible in all simulation scenarios, except when the clusters were small and the intra-cluster correlation exists. The C-index (C_{re}) showed bias in the presence of censoring; the bias was acceptable for censoring up to 30%. The IBS (IBS_{re}) performed poorly even when there is small amount of censoring in the data. The 'overall' validation measures, in general, were affected by the non-zero intra-cluster correlation particularly when the clusters were small, possibly due to the fact that the frailties are poorly estimated for these clusters. The pattern of the effect of censoring on the 'pooled cluster-specific' measures were similar to the corresponding 'overall' measures. The 'pooled' measures, in general, were affected by the small clusters when the level of censoring was high. This is because this approach ignores some of the clusters due to lack of events to calculate the measures. These findings are similar to those with the validation measures for clustered binary outcome (Chapter 4).

Similar to the standard validation measures for independent survival outcome, these validation measures also differ in their flexibility regarding their assumptions and the form of the prognostic model. The C-indices (C_{re} and C_w) only require that the

prognostic model is able to rank the patients. However, the K statistics (K_{re} and K_w) require that the prognostic model was fitted using the proportional hazards (PH) frailty model, that is, the proportional hazards assumption given the frailty holds. The D statistics (D_{re} and D_w) assume that proportional hazards holds and that the prognostic index is normally distributed. Similarly, the calibration slopes (CS_{re} and CS_w) also assumes proportional hazards given the frailty. The predictive accuracy measure IBS (IBS_{re} and IBS_w) only requires that a survival function given the frailty can be calculated for all patients. In addition, all these measures have the same clinical interpretation as their corresponding standard measures.

A similar pattern of recommendations regarding the practical use of these measures for censored data can be made to those with the standard measures. The K statistic (K_{re} and K_w) and calibration slope (CS_{re} and CS_w) can be recommended for validating prognostic models developed with PH frailty model. The D statistic (D_{re} and D_w) can be recommended provided that the distribution of prognostic index is normal. The C-index (CS_{re} and CS_w) can be used when there is a relatively low amount of censoring, for example, not more than 30%. The IBS (IBS_{re} and IBS_w) cannot be recommended as they are affected by censoring. Generally, both the 'overall' and 'pooled cluster-specific' measures are recommended to use in practice. However, one needs to check whether the cluster sizes are sufficiently large (for example, greater than 30) before using the 'pooled' measures.

Similar to the analogous measures for clustered binary data, the validation measures based on the model's conditional predictions using frailty, $S(t|\omega)$, can be recommended for validating models using subjects from the same clusters as that of the development data. It would not be straightforward to use these methods when validation data involve subjects from new clusters. In this case, validation measures based on marginal predictions $S(t)$ or conditional predictions setting the frailties at their mean, $S(t|1)$, could be used. If validation data involve several clusters with moderate to high variability between the clusters, these methods may not produce optimal results. One alternative possibility is to investigate the characteristics of the new clusters to see whether they are similar to that of the existing clusters of the development data. Then

it may be reasonable to assume that clusters in both datasets were sampled from the same population of clusters. In this case, one could estimate frailties from validation data using the estimate of the frailty parameters from the development data and use them to make predictions. If this happens, one may consider this as a form of model re-calibration.

In summary, before choosing the validation measures, it is very important to check the characteristics of the validation data. For example, one needs to check whether the validation data involve the same or different clusters to those with the development data, the level clustering, cluster size, the level of censoring, and the distribution of the prognostic index.

Chapter 6

Summary and Conclusions

6.1 Summary of the research

Prognostic models play a vital role in the clinical management of patients by providing useful information regarding a patient's future health status. These models also have an important application in monitoring the performances of health institutions after adjusting for the case mix of patients. Therefore, it is essential for prognostic models to have the ability to make accurate predictions. One of the key requirements in the prognostic modelling process is the availability of useful and reliable validation measures to assess the predictive ability of these models. This research focuses on validation measures for prognostic models for binary and survival outcomes. The thesis starts with a motivation for this research in Chapter 1, followed by a description of the general procedure for validating a prognostic model and a literature review of some commonly used or proposed validation measures for binary and survival outcomes in Chapter 2.

The literature review on the validation measures for binary and survival outcomes suggests that validation measures for models for independent binary outcomes are well developed. Although a number of measures have been proposed in the last two decades, there is only limited guidance regarding their use in practice. A common feature of survival data is censoring and ideally the validation measures should not be affected

by censoring, however for some measures this may not be the case [52, 53]. This thesis reviews, in Chapter 3, a wide range of validation measures proposed for independent survival outcomes and evaluates their performances using an extensive simulation study in order to make practical recommendations for their use.

In risk prediction research, patients' health outcomes are often clustered within a larger unit, for example, outcomes measure on patients in a hospital, and are likely to be correlated. Ignoring this clustering may lead to incorrect predictions. Therefore, one needs to consider this clustering both in the process of model development and validation of its predictive ability. Random effects logistic and frailty models are often used to develop models for clustered binary and survival outcomes, respectively. However, only limited work has been done to develop validation measures to assess the predictive ability of these models. The rest of this thesis focuses on validation measures that could be used with random effects logistic and frailty models to make risk predictions for clustered binary (Chapter 4) and survival outcomes (Chapter 5), respectively.

6.2 Summary of the methods and results

6.2.1 Validation measures for independent survival outcomes

The investigation, in Chapter 3, focuses on validation measures for independent survival outcomes that have the potential of being routinely used in practice. The measures are selected on the basis of their ease of interpretation and communication, and their availability or ease of implementation in commonly used statistical software. The validation measures selected include the calibration slope [44] from the category of calibration measures; Graf et al's integrated Brier score (IBS) [55] from the category of predictive accuracy measures; Harrell's C-index [8], Gönen and Heller's K statistic [48] and Royston and Sauerbrei's D [49] from the discrimination measures; and Graf et al's R^2_{IBS} [55] and Schemper and Henderson's V [23] from the explained variation category. Using a simulation study based on two clinical datasets with contrasting characteristics, the performance of the validation measures are compared with respect to their robustness to the degree of censoring and sensitivity to the exclusion of important predictors from the model.

The results from simulation study suggest that the calibration slope (CS) and K statistic showed negligible bias induced by censoring, which is to be expected since both are derived from the Cox model. The performance of D statistic depended on the distribution of the prognostic index derived from the model. Provided that the prognostic index is normally distributed, the bias in D was negligible. By central limit theorem, the prognostic index, in practice, is likely to be normally distributed as the number of predictors in the model increases. The C-index, the most widely used measure, showed increasing bias with the increasing level of censoring, which may be expected as it depends on the censoring mechanism. The bias may be acceptable for censoring up to 30%. The measures of predictive accuracy and explained variation (IBS, R^2_{IBS}, and V) performed poorly in the presence of censoring, despite their use of weighting to alleviate the effect of censoring. The bias in all cases suggests that it is possible to reach misleading conclusions regarding a prognostic model's predictive performance using these measures in the presence of censoring. Censoring is a common feature in survival data and typically the degree of censoring will exceed 20% in most real clinical datasets. Thus validations measures for censored survival data need to be selected after careful consideration.

All the validation measures investigated, except the calibration slope, showed sensitivity to the omission of important predictors from a model. However, the ranked-based measure, the C-index, was less sensitive than the other measures, which may be expected as it does not incorporate the actual difference between predictions. The calibration slope showed only limited sensitivity to omission of important predictor since the developed risk model effectively re-calibrates itself to compensate for the omitted predictors.

The validation measures differ in their flexibility regarding their assumptions and the form of the prognostic model. Of the discrimination measures, the C-index only require that the prognostic model is able to rank the patients. In contrast, $K(\beta)$ requires that the prognostic model was fitted using the Cox proportional hazards model. The D statistic assumes that proportional hazards holds and that the prognostic index is normally distributed. The calibration slope also assumes proportional hazards, al-

though more general approaches are described by van Houwelingen [44]. The measures based on predictive accuracy, IBS, R^2_{IBS}, and V, only require that a survival function can be calculated for all patients.

Based on the findings of this simulation study, of the discrimination measures, $K(\beta)$ can be recommended for validating a prognostic model developed using the Cox proportional hazards model, since it is both robust to censoring and reasonably sensitive to the omission of important predictors. The D statistic can also be recommended provided that the distribution of the prognostic index derived from the model is approximately normal. It is more sensitive to predictor omission than $K(\beta)$ and can be calculated for models other than those fitted using the Cox model. The C-index was affected when data have high level of censoring and cannot be recommended for use with data with more than 30% censoring. The calibration slope can be recommended as a measure of calibration since it is not affected by censoring although it is less sensitive than the other measures to the omission of important predictors. In practice, one might additionally investigate calibration graphically by comparing observed and predicted survival curves for groups of patients. This approach also has the benefit of being easy to communicate. The measures of predictive accuracy (IBS) and explained variation (V and R^2_{IBS}) cannot be recommended for use with survival risk models due to their poor performance in the presence of censored data. However, these measures were all conservative with censored data so that high (or low for IBS) values would still be indicative of a good prognostic model.

In practice, it is very important to investigate the characteristics of the validation data before choosing the validation measures. In particular, one needs to check the level of censoring and the distribution of the prognostic index, assuming that the standard model assumptions such as proportional hazards hold. It is not clear that this is routinely done in practice.

6.2.2 Validation measures for clustered data

Chapter 4 shows extensions of some of the standard validation measures for use with models for clustered binary outcomes. These are the C-index [45] and D-statistic

[49] (both assess discrimination), the calibration slope [39, 42] (assesses calibration), and the Brier score [55] (assesses predictive accuracy). Two approaches, termed as the 'overall' and 'pooled cluster-specific' are proposed to calculate these measures for clustered data. Each approach can produce three different measures depending on how the random effects estimates are used in predictions from the model. For example, conditional predictions can be obtained by either using the random effects estimates in predictions or setting them at their mean value of zero. Marginal predictions can be obtained by integrating out the random effects.

The new validation measures are illustrated by developing a model that predicts in-hospital mortality following heart valve surgery in UK hospitals and validating its predictive performance. Both the 'overall' and 'pooled cluster-specific' measures are shown to have meaningful interpretation in a clustered data setting. Additionally, the separate cluster-specific estimates can be used to identify clusters where model performance is either good or poor compared to the average performance. It would be of great interest to investigate the factors which explain this heterogeneity. One possibility is the unobserved cluster level characteristics or mis-specification of the model. Simulation studies were conducted to evaluate the performance of the measures under a range of conditions related to clustered data. The 'overall' measures based on the conditional predictions using the estimates of the random effects showed reasonably good performance in a range of conditions, except for those where the clusters were small. This is because the empirical Bayes estimates of the random effects were poorly estimated for these clusters. These findings are similar to those obtained by Oirbeek and Lesaffre [131]. The validation measures based on the marginal predictions and the conditional predictions that set the random effects to be zero performed poorly in the presence of clustering, because they ignore the effect of clustering. In general, the 'pooled cluster-specific' measures had reasonably good performance when the clusters were large. They showed bias for small clusters, since this approach ignores information from clusters that have very few events.

The validation measures for clustered binary outcome also differ in their flexibility regarding their assumptions and the form of the prognostic model. Therefore one

needs to be careful about these before choosing the measures. Both the parametric C-index and D statistic require that the prognostic index derived from the model should be normally distributed. In contrast, the non-parametric C-index only requires that the prognostic model is able to rank the patients. The calibration slope (CS) assumes that the model is correctly specified. The Brier score only requires that a risk algorithm can be calculated for all patients. In practice, the non-parametric C-index, calibration slope, and Brier score are recommended since they are free from a distributional assumption of the prognostic index. The parametric C-index and D statistic can be used provided that the prognostic index is normally distributed.

In Chapter 5, the calibration slope, Harrell's C-index, K statistic, D statistic, and the Integrated Brier score (IBS) are extended for use with proportional hazards frailty model for clustered survival outcomes, using the same approach as that discussed for clustered binary outcomes. This chapter discusses the use of these measures only for model's conditional predictions that use empirical Bayes estimates of the frailties. Using this approach, it is straightforward to extend the measures for use with marginal predictions and conditional predictions that set the frailties to be one or log-frailties to be zero. An application of these validation measures is illustrated using child mortality data from Bangladesh. A simulation study was conducted to assess the effect of censoring on these measures under various clustered survival data scenarios. The validation measures behaved similarly as the corresponding standard measures for independent survival data, particularly in the presence of censoring. For example, the 'overall' K statistic (K_{re}) showed good performance against censoring in a range of conditions. The prognostic index was specified as normal throughout the simulations and thus the effect of censoring on the D statistic (D_{re}) was negligible when the clusters were large. Similar results were observed for the calibration slope. However, the C-index (C_{re}) was affected by censoring; the bias was acceptable for censoring up to 30%. Similar to the standard measures, IBS (IBS_{re}) had poor performance even when data have small amount of censoring. In general, the measures were affected by the non-zero intra-cluster correlation particularly when the clusters were small, possibly due to the poor estimation of the frailties. Similar to the analogous measures for clustered binary data, the 'pooled' measures had poor performance for the small clusters, probably due to ignoring the clusters that have few events.

Similar to the standard measures, the validation measures for clustered survival data differ in their flexibility regarding their assumptions and the form of the prognostic model. The C-index (C_{re} and C_w) only requires that the prognostic model is able to rank the patients. However, the K statistic (K_{re} and K_w) requires that the prognostic model was fitted using the proportional hazards (PH) frailty model. The D statistic (D_{re} and D_w) assumes that proportional hazards given the frailty holds and that the prognostic index is normally distributed. Similarly, the calibration slope (CS_{re} and CS_w) also assumes proportional hazards given the frailty. The predictive accuracy measure IBS (IBS_{re} and IBS_w) only requires that a survival function given the frailty can be calculated for all patients. One should be aware of these before choosing the measures.

A similar pattern of recommendations regarding the practical use of these measures for censored data can be made to those with the standard measures. For example, the K statistic (K_{re} and K_w) and calibration slope (CS_{re} and CS_w) can be recommended for validating prognostic model developed with PH frailty model. The D statistic (D_{re} and D_w) can be recommended provided that the distribution of prognostic index is normal. The C-index (C_{re} and C_w) cannot be recommended for censoring more than 30%. IBS (IBS_{re} and IBS_w) cannot be recommended.

In practice, both the 'overall' and 'pooled cluster-specific' measures are recommended to use when validating models for clustered data. However, one needs to investigate whether the clusters in the validation data are sufficiently large (for example, greater than 30) and each of these contains at least two events before using the 'pooled' measures.

An important issue that one should consider when validating model for clustered data is whether the validation data involve the same clusters as the development data or involve new clusters. If the clusters are the same for which the random effects are known, conditional predictions using the random effects and the validation measures based on this approach are recommended to assess the predictive ability of the model. It is not straightforward to use this approach for validating model using subjects from

new clusters, since the random effects are unknown. In such circumstances, one option would be to investigate the characteristics of the new clusters to see whether they match those of the clusters in the development data. For example, when predicting clinical outcomes in hospitals one could investigate the prevalence of the outcome, the geographical location, the experience of the clinicians, staff to patient ratios, and information on other relevant factors that could be obtained from routinely collected hospital data. If these important characteristics are similar for the development and validation hospitals, it may then be appropriate to assume that the development and validation hospitals come from the same population. Then the random effects could be estimated from the validation data using the information from the development data, provided that the number of patients in each hospital is not small, for example, not less than 30. When the random effects are estimated from the validation data and used in the predictions, this may be considered as a form of model re-calibration. One could also inspect the value of the between cluster variance in the development data to examine how closely it agrees with that in the validation data and infer whether it is reasonable to use predictions based on the random effects from the validation data. Thus the estimate of the between cluster variance for development data clusters will need to be published along with the risk algorithm by the model developers. If the number of clusters in both validation and development data are of reasonable size, one could use more formal method of comparison such as examining whether the confidence intervals for the between cluster variances from the two datasets overlap or use F-test (for models with normally distributed random effects). However, the equality in the level of clustering between both datasets may be unlikely in practice.

Alternatively, the marginal predictions or conditional predictions setting the random effects at their mean value and the validation measures based on these approach could be used. However, if the validation dataset involves several new clusters, and there is a moderate to high degree of variation between these clusters, then the validation measures based on these two approaches may not produce optimal results regarding the model predictive performance. However, they are conservative with the level of clustering so that high (low for Brier score) values would still imply a model with good predictive ability. However, any form of validation for clustered data would require expert statistical skills and thus may not be suitable to be done by clinicians

6.3 Conclusions

This research describes and evaluates a range of existing and new statistical measures for validating prognostic models for both independent survival outcomes and clustered binary and survival outcomes. In one part of this research (Chapter 3), recommendations for the practical use of some of the validation measures for standard survival models have been presented. In other parts (Chapters 4 and 5), this research extended the calibration slope (CS), C-index, D statistic, K statistic, and Brier score for use with models for clustered binary and survival outcomes. The use of these measures when making predictions in validation data that includes either the same or different clusters to those in the development data are also discussed.

An important point to note is that one needs to investigate the characteristics of the validation data before choosing the validation measures. In particular, one needs to check whether the clusters in the validation data are the same or different to those with the development data, the level clustering, cluster size, the level of censoring (for survival outcome), and the distribution of the prognostic index.

6.4 Possibilities for future research

A number of areas have been identified where further research is possible. These are now described as follows.

In Chapter 3, the investigation of validation measures for standard survival models is conducted based on the Cox proportional hazards model under a range of scenarios. Further investigation could be conducted based on other survival models such as lognormal and accelerated failure time (AFT) models to assess whether the measures perform well for these models. Based on the simulation results one could recommend whether these measures are generalisable to all of these survival models. In addition, further investigation may be required to see whether the measures are sensitive to model mis-specification (if a wrong model is fitted).

In Chapter 4 and 5, the 'overall' validation measures that use the random effects showed bias when the clusters are small. The bias is due to the use of the empirical Bayes estimates of the random effects that are not well estimated for small clusters. Therefore, methods for estimating random effects for small clusters would be another area of research. Simple alternatives are to consider empirical Bayes mode rather than empirical Bayes mean of the posterior distribution of the random effects or to fit clusters as fixed effects rather than random effects.

The validation measures for clustered data are estimated and assessed only under the Normal or Gamma distribution of the random or frailty effects for the logistic and Cox models, respectively. Therefore, estimation and assessment of the validation measures assuming other distributions of the random effects, for example, log-normal (random effects logistic models) and inverse Gaussian (frailty models), could be an area of further research. With this one could assess the sensitivity of the measures to the distribution of the random or frailty effects.

The simulation studies for clustered data were conducted by generating data from a true model based on random effects logistic or frailty models. A possible alternative to these models are marginal models. It may be interesting if one generates clustered data where the true model is marginal and assess the performance of the validation measures based on the random effects models. This would help us to assess the sensitivity of the measures to model mis-specification (a random effects or frailty model is fitted where true model is marginal).

In reality there is likely to be imbalance in the cluster sizes. A more detailed investigation could be conducted to assess whether the degree of imbalance in cluster sizes for validation data may affect the performance of the validation measures, both for binary and survival outcomes.

It may also be of interest to examine how the validation measures developed for clustered data respond to omission of important predictors.

External validation in prognostic modelling process is essential. Therefore, future research is required to identify the best approach to validate a model's predictive performance through an external validation exercise where the validation data include a number of new clusters

Further work could involve in devising approaches to investigate the performance of Hosmer-Lemeshow test for clustered binary data.

Appendix A

Additional Results for Chapter 3

Table A.1 describes the results of the breast cancer and sudden cardiac death simulations with different risk profiles, under administrative censoring mechanism. Table A.2 shows the estimates of the Cox PH model obtained from breast cancer data. Similarly, Table A.3 shows the Cox PH estimates obtained from sudden cardiac death data.

Table A.1: Relative bias (%) and 95% CIs are given by censoring proportions. The results are from the (a) breast cancer simulations (maximum Monte Carlo standard error (%)=0.88) and (b) sudden cardiac death simulations (maximum Monte Carlo standard error (%)=0.82), with different risks profile (low, medium, and high) and under administrative censoring mechanism.

			CS		IBS		D		$K(\beta)$		C-index		V		R^2_{IBS}	
	Scenarios	% cens	bias	CIs	bias	CIs	bias	CIs	bias	CIs	bias	CIs	bias	CIs	bias	CIs
		0	-0.4	[-0.9, 0.2]	-0.7	[-1.4, 0.0]	-0.5	[-1.1, 0.1]	-0.2	[-0.7, 0.3]	-0.8	[-1.4, -0.3]	-0.4	[-0.8, 0]	-0.3	[-0.6, 0]
	low	20	-0.1	[-0.7, 0.6]	2.1	[1.6, 2.7]	1.4	[0.8, 2.1]	-0.4	[-0.9, 0.1]	2.0	[1.3, 2.6]	-8.3	[-9.8, -7.1]	-3.3	[-6.7, -0.1]
		50	-0.2	[-0.9, 0.5]	10.4	[9.5, 11.5]	4.0	[3.2, 4.9]	-0.6	[-1.2, 0.0]	5.6	[4.7, 6.4]	-19.4	[-20.9, -17.8]	-7.3	[-12.1, -2.3]
		80	0.2	[-0.8, 1.3]	19.4	[17.8, 21.1]	7.8	[6.6, 9.1]	-0.4	[-1.2, 0.5]	11.5	[10.2, 12.7]	-39.2	[-40.8, -37.8]	-18.2	[-24.4, -12.5]
		0	-0.3	[-0.8, 0.3]	0.5	[-0.2, 1.2]	-0.3	[-0.8, 0.3]	0.1	[-0.5, 0.3]	-0.2	[-0.7, 0.3]	-0.3	[-0.7, 0.1]	0.1	[-0.3, 0.5]
(a)	medium	20	0.0	[-0.5, 0.5]	2.2	[1.6, 3.8]	0.4	[-0.2, 1.0]	0.1	[-0.3, 0.5]	1.7	[1.1, 2.2]	-10.2	[-11.7, -8.9]	-4.1	[-8.2, -0.1]
		50	0.2	[-0.5, 0.9]	11.5	[10.5, 12.5]	0.8	[-0.1, 1.6]	0.2	[-0.4, 0.7]	4.5	[3.7, 5.2]	-21.1	[-22.6,-19.5]	-8.0	[-12.9, -3.2]
		80	0.7	[-0.4, 1.8]	16.5	[14.8, 18.3]	0.8	[-0.3, 1.9]	0.3	[-0.5, 1.2]	7.7	[6.5, 8.8]	-31.9	[-33.2, -30.1]	-14.9	[-20.5, -8.6]
		0	0.2	[-0.4, 0.8]	-0.3	[-0.8, 0.2]	0.2	[-0.4, 0.8]	0.0	[-0.5, 0.5]	-0.2	[-0.8, 0.4]	0.3	[-0.1, 0.7]	-0.3	[-0.7, 0.1]
	high	20	-0.4	[-1.0, 0.3]	2.8	[1.8, 3.3]	1.2	[0.5, 1.8]	-0.5	[-1.0, 0.0]	1.3	[0.4, 1.9]	-8.2	[-10.4, -6.4]	-3.6	[-10.7, -0.6]
		50	0.1	[-0.7, 0.9]	10.5	[9.5, 11.1]	2.2	[1.5, 3.0]	-0.2	[-0.8, 0.4]	2.6	[1.6, 3.4]	-18.3	[-20.2, -16.5]	-10.8	[-15.0, -6.6]
		80	1.0	[-0.1, 2.2]	14.5	[13.6, 15.5]	2.8	[1.8, 3.9]	0.3	[-0.6, 1.3]	3.9	[2.8, 5.2]	-38.1	[-39.3, -36.8]	-16.5	[-21.5, -11.4]
		0	0.2	[-0.2, 0.7]	0.1	[-0.3, 0.5]	0.1	[-0.4, -0.5]	0.2	[-0.3, 0.7]	0.2	[-0.3, 0.7]	-0.3	[-0.7, 0.1]	-0.1	[-0.6, 0.5]
	low	20	-0.1	[-0.5, 0.4]	2.7	[2.2, 3.2]	3.5	[2.9, 3.9]	-0.1	[-0.5, 0.3]	3.1	[2.5, -3.6]	-9.8	[-12.7, -7.8]	-11.9	[-14.2, -12.0]
		50	-0.2	[-0.7, 0.3]	11.1	[10.2, 12.0]	8.7	[7.5, 9.9]	-0.1	[-0.6, 0.4]	9.1	[8.2, 10.1]	-24.1	[-28.1, -20.9]	-16.3	[-18.5, -14.1]
		80	0.1	[-0.7, 0.8]	18.6	[17.1, 20.1]	17.6	[16.5, 18.7]	0.5	[-0.1, 1.1]	17.7	[15.7, 19.6]	-43.7	[-47.3, -39.2]	-25.2	[-29.5, -21.0]
		0	0.0	[-0.4, 0.4]	-0.2	[-0.5, 0.1]	0.0	[-0.4, 0.4]	0.2	[-0.1, 0.5]	0.1	[-0.3, 0.5]	0.5	[-0.1, 1.0]	0.3	[-0.2, 0.8]
(b)	medium	20	0.1	[0.0, 0.2]	4.6	[4.1, 5.1]	2.7	[2.3, 3.1]	0.2	[-0.2, 0.5]	2.7	[2.3, 3.1]	-10.9	[-12.1, -10.0]	-7.8	[-8.9, -6.7]
		50	-0.1	[-0.4, 0.3]	10.2	[9.7, 11.0]	7.1	[6.5, 7.6]	0.2	[-0.2, 0.6]	7.5	[6.9, 8.0]	-17.7	[-19.6, -15.6]	-14.3	[-16.4, -11.8]
		80	0.1	[-0.4, 0.5]	17.7	[16.7, 18.8]	11.2	[10.3, 12.3]	0.2	[-0.4, 0.8]	13.9	[13.0, 14.7]	-35.2	[-38.7, -32.4]	-21.9	[-24.1, -21.4]
		0	-0.1	[-0.4, 0.3]	0.5	[-0.1, 1.1]	0.1	[-0.2, 0.4]	-0.1	[-0.3, 0.1]	0.5	[0.3, 0.9]	0.3	[-0.2, 0,8]	0.3	[-0.2, 0.8]
	high	20	0.1	[-0.4, 0.5]	3.3	[2.8, 3.8]	1.8	[1.5, 2.0]	0.3	[-0.1, 0.7]	2.4	[1.9, 2.8]	-12.5	[-14.8, -10.3]	-6.7	[-8.8, -4.6]
		50	-0.1	[-0.6, 0.3]	8.5	[7.4, 9.6]	4.1	[3.6, 4.5]	0.2	[-0.3, 0.7]	5.8	[5.2, 6.3]	-25.1	[-29.1, -21.0]	-10.4	[-12.7, -8.1]
		80	0.4	[-0.4, 1.1]	12.4	[11.4, 13.4]	7.4	[6.7, 8.2]	0.1	[-0.5, 0.7]	10.6	[9.7, 11.5]	-35.1	[-39.1, -31.3]	-20.2	[-23.6, -17.1]

Table A.2: The Cox model estimates for the breast cancer data

Predictors	Meas. scale	Mean(SD)/%	HR	95% CI	P-value
lymph nodes (lnod)	log	1.16(0.94)	1.61	[1.41, 1.83]	< 0.001
progesterone status (progst)	log	3.35(1.93)	0.82	[0.76, 0.89]	< 0.001
hormone	no	64.1	1.00		< 0.01
	yes	35.9	0.68	[0.53, 0.87]	
menopausal status (menst)	pre	42.3	1.00		< 0.01
	post	57.7	1.33	[0.95, 1.88]	
age	≤ 45	22.3	1.00		0.068
	45-60	50.3	0.66	[0.46, 0.94]	
	> 60	27.4	0.65	[0.41, 1.02]	
tumour grade (tgrad)	1	11.8	1.00		0.102
	≥ 2	64.7	1.69	[1.03, 2.77]	
	3	23.5	1.74	[1.01, 3.00]	
tumour size (tsize)	log	3.27(0.46)	1.21	[0.93, 1.56]	0.151

-2loglikelihood= 3436.10; Likelihood Ratio=138.91 with d.f=9

Table A.3: The Cox model estimates for the sudden cardiac death data

Predictors	Meas. scale	Mean(SD)/%	HR	95% CI	P-value
runs-ventricular-tachycardia (runvent)	none	83.5	1.00		< 0.001
	1	10.4	2.40	[1.32, 4.33]	
	2^+	6.1	2.69	[1.36, 5.28]	
obstruction to blood flow (BF)	mmHg	30.8(35.2)	1.01	[1.00, 1.02]	< 0.001
blood pressure during exercise (BP)	normal	74.7	1.00		< 0.01
	abnormal	25.3	1.82	[1.14,2.89]	
thickness of heart muscle (HM)	mm	19.5(6.1)	1.04	[1.00,1.07]	< 0.05
age	years	37.8 (16.2)	0.98	[0.97, 1.00]	0.11

-2loglikelihood= 1032.70; Likelihood Ratio=36.03 with d.f=6

Appendix B

Stata code for validation measures

Figure B.1: Stata code for calculating D-statistic, Gonen and Heller' K, Integrated Brier score, and Schemper and Henderson' V measures for independent survival data.

```
set obs 500
qui gen index=_n
qui gen IBSx=.
qui gen IBS0=.
qui gen IRsq=.
qui gen V=.
qui gen D=.
qui gen K=.
local reps 500
local seed=6734535
set seed `seed'
forvalues j=1/`reps'{
preserve
***generating data***
qui set obs 500
local haz=0.2 // this gives 20% censoring on average
qui gen x=invnormal(runiform()) // x from standard normal
qui gen xb0=1.2*x // beta=1.2
qui gen t_fail=(-1/exp(xb0)*log(uniform()))^(1/0.45) //shape=0.45 and scale=1
qui gen t_cens=(-1/(`haz')*log(uniform()))^(1/0.45)
qui gen time_sm=min(t_fail,t_cens)
qui gen byte d=(t_fail<=t_cens)
```

```
***fitting Cox Model***
qui stset time_sm, f(d)
qui stcox x, nohr basesurv(bsurv) // model with covariates x
predict xb, xb
qui stcox, estimate basesurv(st0) // null model

***D statistics***
sort xb
gen z= invnorm(((_n-3/8)/(_N+1/4)))/sqrt(8/_pi)
qui stcox z, nohr
local D=_b[z]

***Gonen & Heller's K***
qui tempvar Phi
qui gen double Phi=.
qui gen wv=1
qui local i=1
qui while `i'<=_N {
qui local x=xb[`i']
qui tempvar phi Fhi
qui gen float phi=wv if xb<`x'
qui replace   phi= 0  if xb>`x'
qui gen float Fhi=phi/(1+exp(xb-`x'))
qui sum Fhi, meanonly
qui replace Phi=r(sum) if _n==`i'
qui drop phi Fhi
qui local i=`i'+1
}
qui sum Phi, meanonly
qui gen sumPhi=r(sum)
qui gen K=2*sumPhi/(_N*(_N-1))

***Graf et al.'s IBS and IRsq, and Schemper and Henderson's V***
qui sort time_sm
qui gen tm_d=time_sm if d==1
qui sum tm_d
local tau=r(max) //maximum event-time; we calculate IBS up-to this time point
qui gen yt=0 if time_sm<=`tau'
qui replace yt=1 if time_sm>`tau'
qui sts gen km=s //K-M survival probability to calculate weight function
set obs 501
qui replace _t=`tau' in 501
ipolate km _t, gen(ks2) epolate
qui summarize ks2 if _t==`tau'
local gt1=r(mean)
qui gen wt1=(1-km)/(1-`gt1') // calculate weight function to use in IBS
qui gen exb=exp(xb)
qui gen delta=1 if d==0
qui replace delta=0 if d==1
```

```
qui stset time_sm, f(delta) // censoring indicator reverse
qui sts gen gt=s   //calculate K-M estimate of not being censored
qui gen wt=1/gt //calculate weight to compensate earlier censoring
tempvar Mt Mt0 Bst Bst0
qui gen double Mt=.
qui gen double Mt0=.
qui gen double Bst=.
qui gen double Bst0=.
sort time_sm
qui gen sumd=sum(d) if d==1
qui gen wv=1
qui gen bsurvj=bsurv if d==1 // baseline survival probability at each event-time
qui gen st0j=st0 if d==1   // survival estimate of null model at each event-time
local i=1
while `i'<=_N {
sort time_sm
local time=time_sm[`i']
local st0=st0[`i']
local st0j=st0j[`i']
local bsurv=bsurv[`i']
local bsurvj=bsurvj[`i']
local wtg=wt[`i']
tempvar y st bs bs0 mt mt0 stj
qui gen float y=wv if time_sm>`time' //actual survival status at each time point
qui replace   y= 0  if time_sm<`time'
qui gen float st=(`bsurv')^exb //based on model with x at each observed time
qui gen float stj=(`bsurvj')^exb //based on model with x at each event-time

*Brier score********
qui gen  float bs=d*(1-y)*(0-st)^2*(wt)+y*(1-st)^2*(`wtg')//from model with x
sum bs, meanonly
qui replace Bst=r(mean) if _n==`i'
qui gen float bs0=d*(1-y)*(0-`st0')^2*(wt)+y*(1-`st0')^2*(`wtg')//from null model
sum bs0, meanonly
qui replace Bst0=r(mean) if _n==`i'

*Mtx and Mt0 part of V*******
qui gen float mt=y*(1-stj)+(1-y)*stj+(1-d)*(1-y)*((1-stj)*(stj/st)
                +stj*(1-stj/st)) if d==1 // based on model with x
sum mt, meanonly
qui replace Mt=r(mean) if _n==`i'
qui gen float mt0=y*(1-`st0j')+(1-y)*`st0j'+(1-d)*(1-y)*((1-`st0j')*(`st0j'/`st0')
                 +`st0j'*(1-`st0j'/`st0')) if d==1 // based on null model
sum mt0, meanonly
qui replace Mt0=r(mean) if _n==`i'
drop y st bs bs0 mt mt0 stj
local i=`i'+1
}
***Integrated BS***
```

```
qui integ Bst wt1 if yt==0, trapezoid gen(BS_x)
local IBSx=r(integral) //integrated Brier score based on the model with x
qui integ Bst0 wt1 if yt==0, trapezoid gen(BS_0)
local IBS0=r(integral) //integrated Brier score based on the null model
local IRsq=1-‘IBSx’/‘IBS0’ // R-square

***Dx and D0 part of V***
qui gen w=sumd/gt if d==1
qui replace Mt0=Mt0*w
qui replace Mt=Mt*w
qui sum w if yt==0
local sumw=r(sum)
qui sum Mt0 if yt==0
local Mt00=r(sum)
local D0=‘Mt00’/‘sumw’
qui sum Mt if yt==0
local Mtx=r(sum)
local Dx=‘Mtx’/‘sumw’
local V=1-‘Dx’/‘D0’

**line Bst Bst0 time_sm if yt==0 //to draw graph for BS over the entire follow-period
restore
qui replace IBSx=‘IBSx’ if index==‘j’
qui replace IBS0=‘IBS0’ if index==‘j’
qui replace IRsq=‘IRsq’ if index==‘j’
qui replace V=‘V’ if index==‘j’
qui replace D=‘D’ if index==‘j’
}
```

Figure B.2: Stata code for calculating validation measures for clustered binary data: C-index, D-statistic, Calibration slope, and Brier score.

```
qui set obs 500
local seed=1677445
qui gen index=_n
qui gen Cre=.
qui gen Dre=.
qui gen CSre=.
qui gen BSre=.
qui gen seCre=.
qui gen seDre=.
qui gen seCSre=.

set seed ‘seed’
local reps=500
forvalues j=1/‘reps’{
preserve
```

```
*******************************
*generate development data*
*******************************
qui set obs 10000
qui gen rnd=uniform()
qui xtile cluster=rnd, n(100)
qui gen u=invnormal(runiform())
qui bysort cluster:replace u=u[1]
qui gen x=invnormal(runiform())

local total sigmau+sigmae=1.4 //varying sigmau gives different ICC values
*local sigmau=0.0  // ICC 0%
*local sigmau=0.44 // ICC 5%
*local sigmau=0.68 // ICC 10%
local sigmau=0.88 //  ICC 20%
local sigmae=sqrt(1.4^2-‘sigmau’^2)
qui gen z=-1.8+‘sigmae’*x+‘sigmau’*u
qui gen p=1/(1+exp(-z))
qui gen y=(runiform()<p)

***Fitting random intercept logistic model*****
qui gllamm y x, i(cluster) link(logit) family(binomial) adapt nip(20)
qui estimates store gllamm
clear
*******************************************************
*generate validation data of 10 clusters of size 300*
*******************************************************
qui set obs 3000
qui gen rnd=uniform()
qui xtile cluster=rnd, n(10)
qui gen u=invnormal(runiform())
qui bysort cluster:replace u=u[1]
qui gen x=invnormal(runiform())

local total sigmau+sigmae=1.4 //total variability is fixed
*local sigmau=0.0  // ICC 0%
*local sigmau=0.44 // ICC 5%
*local sigmau=0.68 // ICC 10%
local sigmau=0.88 //  ICC 20%
local sigmae=sqrt(1.4^2-‘sigmau’^2)
qui gen z=-1.8+‘sigmae’*x+‘sigmau’*u
qui gen p=1/(1+exp(-z))
qui gen y=(runiform()<p)
*****************************************************************
*calculation of the C-index, D-statistic and Brier score measures*
*****************************************************************
qui estimates restore gllamm // estimates from development data are restored
qui gllapred eb, u fsample // empirical Bayes estimates from validation data
qui gllapred condpred, mu us(ebm) fsample // cond. pred. with random effect(u)
```

```
qui gllapred margpred, mu marginal fsample // marginal predictions
qui gen zeta1=0
qui gllapred condpred0, mu us(zeta) fsample // cond. pred. with u set to zero
qui gen xb=ln(condpred/(1-condpred))
qui gen xb0=ln(condpred0/(1-condpred0))
qui gen xbm=ln(margpred/(1-margpred))

**** nonparametric Overall C-index****
qui sum xb if y==1
local N1=r(N)
qui sum xb if y==0
local N0=r(N)
tempvar Phi
qui gen double Phi=.
qui gen wv=1
sort cluster y xb
local i=1
while `i'<=_N {
             tempvar phi Fhi
             local x=xb[`i']
if y[`i']==1 {
            sort cluster
            qui by cluster:gen float phi=wv if xb<`x' & y==0
            qui by cluster:replace   phi= 0.5*wv  if xb==`x' & y==0
            qui by cluster:replace   phi= 0  if xb>`x' & y==0
    qui by cluster:egen float Fhi=sum(phi)
     qui by cluster:replace Fhi=0 if _n!=1
     sum Fhi, meanonly
            qui replace Phi=r(sum) if _n==`i'
     }
        else {
             qui by cluster:gen float phi=wv if xb>`x' & y==1
             qui by cluster:replace   phi= 0.5*wv  if xb==`x' & y==1
             qui by cluster:replace   phi= 0  if xb<`x' & y==1
     qui by cluster:egen float Fhi=sum(phi)
     qui by cluster:replace Fhi=0 if _n!=1
     sum Fhi, meanonly
             qui replace Phi=r(sum) if _n==`i'
     }
             qui drop phi Fhi
             local i=`i'+1
           }
qui gen Phi1=Phi if y==1
qui by cluster:egen theta00=sum(Phi1)
qui by cluster:replace theta00=0 if  _n!=1
sum theta00, meanonly
qui gen theta22 =r(sum)
local Cre=theta22/(`N1'*`N0')
qui  drop theta00 theta22 Phi1
```

```
****SE of C-index****
qui by cluster:gen v10=Phi/(‘N0’) if y==1
qui by cluster:egen v10_s=sum(v10)
qui by cluster:gen v01=Phi/(‘N1’) if y==0
qui by cluster:egen v01_s=sum(v01)
sort cluster y
qui by cluster y:gen Ng=_N
qui by cluster:gen mi=Ng if y==1
qui by cluster:egen m=mean(mi)
qui by cluster: replace m=0 if m==.
qui by cluster:gen ni=Ng if y==0
qui by cluster:egen n=mean(ni)
qui by cluster: replace n=0 if n==.
qui by cluster:gen D10_1=(v10_s-m*‘Cre’)
qui by cluster:gen D10=D10_1^2 if _n==1
qui by cluster:replace D10=0 if _n!=1
qui by cluster:gen D01_1=(v01_s-n*‘Cre’)
qui by cluster:gen D01=D01_1^2 if _n==1
qui by cluster:replace D01=0 if _n!=1
qui tab cluster if y==1
local N10=r(r)
qui tab cluster if y==0
local N01=r(r)
qui egen S10=sum(D10)
qui replace S10=(S10*‘N10’)/((‘N10’-1)*‘N1’)
qui egen S01=sum(D01)
qui replace S01=(S01*‘N01’)/((‘N01’-1)*‘N0’)
qui by cluster:gen DD=D10_1*D01_1
qui by cluster:replace DD=0 if _n!=1
qui egen DD1=sum(DD)
qui tab cluster
local I=r(r)
qui gen S11=(DD1*‘I’)/(‘I’-1)
qui gen var_Cbcn=S10/‘N1’+S01/‘N0’+(2*S11)/(‘N1’*‘N0’)
local seCre=sqrt(var_Cbcn)
qui  drop v* S* D* mi ni  Ng

****D-statistics****
sort xb
qui gen zre= invnorm(((_n-3/8)/(_N+1/4)))/sqrt(8/_pi)
qui logit y zre
local Dre=_b[zre]
local seDre=_se[zre]

***CalibrationSlope***
qui logit y xb
local CSre=_b[xb]
local seCSre=_se[xb]
```

```
***Brier score***
qui gen bs=(y-condpred)^2
sum bs, meanonly
local BSre=r(mean)
*****************************
di `j'
restore
qui replace Cre=`Cre' if index==`j'
qui replace Dre=`Dre' if index==`j'
qui replace CSre=`CSre' if index==`j'
qui replace BSre=`BSre' if index==`j'
qui replace seCre=`seCre' if index==`j'
qui replace seDre=`seDre' if index==`j'
qui replace seCSre=`seCSre' if index==`j'
}
```

Figure B.3: Stata code for calculating validation measures for clustered survival data: Harrell's C-index, Gonen and Heller's K, D statistic, Integrated Brier score (IBS).

```
qui set obs 500
qui gen index=_n
qui gen Cre=.
qui gen Kre=.
qui gen Dre=.
qui gen CSre=.
qui gen seCre=.
qui gen seKre=.
qui gen seDre=.
qui gen seCSre=.
qui gen IBSx=.
qui gen IBS0=.
qui gen IRsq=.
local reps=500
forvalues j=1/`reps'{
preserve
**************************
*generate development data
**************************
set obs 1500
gen cluster=int(uniform()*50)+1
*local theta=0.0 // no corr
local theta=0.58 // moderate corr
*local theta=0.98 // high corr
local beta=1.35
local lambda=0.0 //censoring 0%
local gamma=1.1
```

```
gen mu=rgamma(1/‘theta’, ‘theta’)
qui bysort cluster:replace mu=mu[1]
qui gen x=invnormal(runiform())
qui gen xb0=‘beta’*x
qui gen t_fail=(-1/(mu*exp(xb0))*log(uniform()))^(1/‘gamma’)
*qui gen t_fail=(-1/(exp(xb0))*log(uniform()))^(1/‘gamma’) // if theta=0
qui gen t_cens=(-1/(mu*‘lambda’)*log(uniform()))^(1/‘gamma’)
*qui gen t_cens=(-1/(‘lambda’)*log(uniform()))^(1/‘gamma’) // if theta=0
qui gen time_sm=min(t_fail,t_cens)
qui gen byte d=(t_fail<=t_cens)
qui stset time_sm, f(d)

***Fit PH frailty model***
qui stcox x, nohr shared(cluster)
local sebeta=_se[x]
qui estimates store stcox
clear
****************************
*generate validation data
*************************
set obs 1500
gen cluster=int(uniform()*50)+1
*local theta=0.0 // no corr
local theta=0.58 // moderate corr
*local theta=0.98 // high corr
local beta=1.35
*local lambda=0.0 //censoring 0%
local lambda=0.20 //censoring 20%
*local lambda=0.99 //censoring 50%
*local lambda=5.0// censoring 80%

local gamma=1.1
gen mu=rgamma(1/‘theta’, ‘theta’)
qui bysort cluster:replace mu=mu[1]
qui gen x=invnormal(runiform())
qui gen xb0=‘beta’*x
qui gen t_fail=(-1/(mu*exp(xb0))*log(uniform()))^(1/‘gamma’)
*qui gen t_fail=(-1/(exp(xb0))*log(uniform()))^(1/‘gamma’) // if theta=0
qui gen t_cens=(-1/(mu*‘lambda’)*log(uniform()))^(1/‘gamma’)
*qui gen t_cens=(-1/(‘lambda’)*log(uniform()))^(1/‘gamma’) // if theta=0
qui gen time_sm=min(t_fail,t_cens)
qui gen byte d=(t_fail<=t_cens)
qui stset time_sm, f(d)
****************************************
*Calculation of validation measures
****************************************
qui estimates restore stcox
predict xbf, xb
qui stcox, estimate shared(cluster) offset(xbf)
```

```
predict xbu,effects
qui gen xb=xbf+xbu
qui sort cluster
qui egen sdx=sd(xb)
qui gen h=(0.5*sdx)/(_N)^(1/3)
tempvar Phi Psi Psi0 Chi Chi0
qui gen double Phi=.
qui gen double Psi=.
qui gen double Psi0=.
qui gen double Chi=.
qui gen double Chi0=.
qui gen double Chij=.
qui gen double Dhij=.
qui gen wv=1
qui gen xb_1=xb if d==1
qui gen time_1=time_sm if d==1
local i=1
while `i'<=_N {
local x=xb[`i']
local x1=xb_1[`i']
local time1=time_1[`i']
tempvar fhi Fhi chi chi0 fhi0 Fhi0 Csi0 chij dhij chij0 dhij0
sort cluster
qui by cluster:gen float fhi=normal(-(xb-`x')/h)/(1+exp(xb-`x'))
qui by cluster:egen float Fhi=sum(fhi)
qui by cluster:replace Fhi=0 if _n!=1
sum Fhi, meanonly
qui replace Phi=r(sum) if _n==`i'
qui by cluster:gen float chi=wv if xb<`x1'
            qui by cluster:replace   chi= 0  if xb>`x1'
qui by cluster:gen float chi0=wv if time_sm>`time1'
            qui by cluster:replace   chi0= 0  if time_sm<`time1'
qui by cluster:gen float fhi0=chi*chi0
qui by cluster:egen float Fhi0=sum(fhi0)
qui by cluster:replace Fhi0=0 if _n!=1
qui by cluster:egen float Csi0=sum(chi0)
qui by cluster:replace Csi0=0 if _n!=1
sum Fhi0, meanonly
qui replace Chi=r(sum) if _n==`i'
  sum Csi0, meanonly
  qui replace Chi0=r(sum) if _n==`i'
qui by cluster:egen chij=count(fhi0) if fhi0==1
qui by cluster: replace chij=0 if chij==.
qui by cluster: egen chij0=mean(chij)
qui by cluster:replace chij0=0 if _n!=1
sum chij0, meanonly
qui replace Chij=r(sum) if _n==`i'
qui by cluster: egen dhij=count(fhi0) if fhi0==0
qui by cluster: replace dhij=0 if dhij==.
```

```
qui by cluster: egen dhij0=mean(dhij)
qui by cluster:replace dhij0=0 if _n!=1
sum dhij0, meanonly
qui replace Dhij=r(sum) if _n==‘i’
drop fhi Fhi  chi chi0 fhi0 Fhi0 Csi0 chij dhij chij0 dhij0
       local i=‘i’+1
}

qui by cluster:egen Kn0=sum(Phi)
qui by cluster:replace Kn0=0 if  _n!=1
sum Kn0, meanonly
qui gen sumPhi=r(sum)
qui gen Knre=2*sumPhi/(_N*(_N-1))
local Kre=Knre
qui by cluster:egen C1=sum(Chi)
qui by cluster:replace C1=0 if  _n!=1
sum C1, meanonly
qui gen sumChi1=r(sum)
qui by cluster:egen C0=sum(Chi0)
qui by cluster:replace C0=0 if  _n!=1
sum C0, meanonly
qui gen sumChi0=r(sum)
qui gen  Cre=sumChi1/sumChi0
local Cre=Cre

***SE of C for clustered data***
qui by cluster:egen Chij0=sum(Chij)
qui by cluster:replace Chij0=0 if _n!=1
sum Chij0, meanonly
gen sumChij0=r(sum)
gen Pc=sumChij0/(_N*(_N-1))

qui by cluster:egen Dhij0=sum(Dhij)
qui by cluster:replace Dhij0=0 if _n!=1
sum Dhij0, meanonly
qui gen sumDhij0=r(sum)
gen Pd=sumDhij0/(_N*(_N-1))

qui by cluster:gen Ch=Chij*(Chij-1)
qui by cluster:gen Dh=Dhij*(Dhij-1)

qui by cluster:egen Ch0=sum(Ch)
qui by cluster:replace Ch0=0 if _n!=1

sum Ch0, meanonly
gen sumCh0=r(sum)
gen Pcc=sumCh0/(_N*(_N-1)*(_N-2))

qui by cluster:egen Dh0=sum(Dh)
```

```
qui by cluster:replace Dh0=0 if _n!=1

sum Dh0, meanonly
gen sumDh0=r(sum)
gen Pdd=sumDh0/(_N*(_N-1)*(_N-2))

qui by cluster:gen ChDh=Chij*Dhij
qui by cluster:egen ChDh0=sum(ChDh)
qui by cluster:replace ChDh0=0 if _n!=1

sum ChDh0, meanonly
gen sumChDh0=r(sum)
gen Pcd=sumChDh0/(_N*(_N-1)*(_N-2))
gen var_p=(4/(Pc+Pd)^4)*(Pd^2*Pcc-2*Pc*Pd*Pcd+Pc^2*Pdd)
local seCre=sqrt(var_p/_N)

****SE of Kre****
local i=1
while `i'<=_N {
local x=xb[`i']
local xi=x[`i']
tempvar phi dji phi0 Dji
sort cluster
qui by cluster:gen float phi=(normal(-(xb-`x')/h)/(1+exp(xb-`x'))-Knre)
            *(normal(-(xb[_n+1]-`x')/h)/(1+exp(xb[_n+1]-`x'))-Knre)
qui by cluster:gen float dji=(-(x-`xi')/h)*normalden(-(xb-`x')/h)*(1+exp(xb-`x'))^(-1)
            +normal(-(xb-`x')/h)*(-(x-`xi'))*exp(xb-`x')*(1+exp(xb-`x'))^(-2)
qui by cluster:egen float phi0=sum(phi)
qui by cluster:replace phi0=0 if _n!=1
sum phi0, meanonly
qui replace Psi=r(sum) if _n==`i'
qui by cluster:egen float Dji=sum(dji)
qui by cluster:replace Dji=0 if _n!=1
sum Dji, meanonly
qui replace Psi0=r(sum) if _n==`i'
drop  phi dji phi0 Dji
local i=`i'+1
}
qui by cluster: egen Knre0=sum(Psi)
qui by cluster:replace Knre0=0 if  _n!=1
sum Knre0, meanonly
qui gen sumPsi=r(sum)
qui gen var_Knre0=(4*sumPsi)/(_N*(_N-1))^2
qui by cluster: egen Dji0=sum(Psi0)
qui by cluster:replace Dji0=0 if  _n!=1
sum Dji0, meanonly
qui gen sumPsi0=r(sum)
qui gen Delji=(2*sumPsi0)/(_N*(_N-1))
qui gen var_Knre=var_Knre0+Delji*`sebeta'^2*Delji
```

```
local seKre=sqrt(var_Knre)
qui drop  Ch* Dh* Ph* Ps* sumCh* sumDh* Pd* Pc* var_p

****D and calibration Slope******
sort xb
qui gen z = invnorm(((_n-3/8)/(_N+1/4)))/sqrt(8/_pi)
qui stcox z, nohr
local Dre=_b[z]
local seDre=_se[z]
qui stcox xb, nohr  basesurv(bsurv)
local CSre=_b[xb]
local seCSre=_se[xb]

*** IBS*********
qui sort time_sm
qui gen tm_d=time_sm if d==1
qui sum tm_d
local tau=r(max) //maximum event-time; we calculate IBS up-to this time point
qui gen yt=0 if time_sm<=‘tau’
qui replace yt=1 if time_sm>‘tau’
qui sts gen km=s
set obs 1501
qui replace _t=‘tau’ in 1501
ipolate km _t, gen(ks2) epolate
qui summarize ks2 if _t==‘tau’
local gt1=r(mean)
qui gen wt1=(1-km)/(1-‘gt1’)
qui gen exb=exp(xb)
local gamma=1.1
qui gen st0=km
qui gen delta=1 if d==0
qui replace delta=0 if d==1
qui stset time_sm, f(delta)
qui sts gen gt=s
***************************
tempvar  Bst Bst0
qui gen double Bst=.
qui gen double Bst0=.
qui sort time_sm
qui gen wt=1/gt
qui sort time_sm
local i=1
while ‘i’<=_N {
local time=time_sm[‘i’]
local st0=st0[‘i’]
local wtg=wt[‘i’]
local bsurv=bsurv[‘i’]
tempvar bs bs0 y st
qui gen float y=wv if time_sm>‘time’
```

```
        qui replace   y= 0  if time_sm<‘time’
qui gen float st=(‘bsurv’)^exb
qui gen  float bs=d*(1-y)*(0-st)^2*(wt)+y*(1-st)^2*(‘wtg’)
*qui gen  float bs=(y-st)^2 // if there is no censoring
sum bs, meanonly
qui replace Bst=r(mean) if _n==‘i’
qui gen float bs0=d*(1-y)*(0-‘st0’)^2*(wt)+y*(1-‘st0’)^2*(‘wtg’)
*qui gen  float bs0=(y-‘st0’)^2 // if there is no censoring
sum bs0, meanonly
qui replace Bst0=r(mean) if _n==‘i’
drop  bs bs0 y st
local i=‘i’+1
}

qui integ Bst wt1 if yt==0, trapezoid gen(BS_x)
local IBSx=r(integral)
qui integ Bst0 wt1 if yt==0, trapezoid gen(BS_0)
local IBS0=r(integral)
local IRsq=1-‘IBSx’/‘IBS0’
line Bst Bst0 time_sm if yt==0

di ‘j’ // print number of simulations completed

restore
qui replace Cre=‘Cre’ if index==‘j’
qui replace Kre=‘Kre’ if index==‘j’
qui replace Dre=‘Dre’ if index==‘j’
qui replace CSre=‘CSre’ if index==‘j’
qui replace seCre=‘seCre’ if index==‘j’
qui replace seKre=‘seKre’ if index==‘j’
qui replace seDre=‘seDre’ if index==‘j’
qui replace seCSre=‘seCSre’ if index==‘j’
qui replace IBSx=‘IBSx’ if index==‘j’
qui replace IBS0=‘IBS0’ if index==‘j’
qui replace IRsq=‘IRsq’ if index==‘j’
}
```

Bibliography

[1] G. Ambler, R. Z. Omar, P. Royston, R. Kinsman, B. E. Keogh, and K. M. Taylor. Generic, simple risk stratification model for heart valve surgery. *Circulation*, 112:224–231, 2005.

[2] M. H. Galea, R. W. Blamey, C. E. Elston, and I. O. Ellis. The Nottingham prognostic index in primary breast cancer. *Breast Cancer Research and Treatment*, 22:207–219, 1992.

[3] K. G. M. Moons, P. Royston, Y. Vergouwe, D. E. Grobbee, and D. G. Altman. Prognosis and prognostic research: what, why and how? *BMJ*, 338:1317–1320, 2009.

[4] P. Royston, K. G. M. Moons, D. G. Altman, and Y. Vergouwe. Prognosis and prognostic research: developing a prognostic model. *BMJ*, 338:1373–1377, 2009.

[5] A. Abu-Hanna and P. J. F. Lucas. Prognostic models in medicine. *Methods of Information in Medicine*, 40:1–5, 2001.

[6] K. G. M. Moons, D. G. Altman, Y. Vergouwe, and P. Royston. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *BMJ*, 338: 1487–1490, 2009.

[7] D. G. Altman and P. Royston. What do you mean by validating a prognostic model? *Statistics in Medicine*, 19:453–473, 2000.

[8] F. E. Harrell Jr., K. L. Lee, and D. B. Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing error. *Statistics in Medicine*, 15(4):361–387, 1996.

[9] D. G. Altman. Prognostic models: A methodological framework and review of models for breast cancer. *Cancer Investigation*, 27:235–243, 2009.

[10] R. Z. Omar, G. Ambler, P. Royston, J. Elihaoo, and K. M. Taylor. Cardiac surgery risk modeling for mortality: a review of current practice and suggestions for improvement. *Ann Thorac Surg*, 77:2232–2237, 2004.

[11] G. M. Clark. Do we need prognostic factors for breast cancer? *Breast Cancer Research and Treatment*, 30(2):117–126, 1994.

[12] F. C. R. Cockburn, H. R. Gamsu, A. Greenough, A. Hopkins, (...), and A. R. Wilkinson. The CRIB (clinical risk index for babies) score: a tool for assessing initial neonatal risk and comparing performance of neonatal intensive care units. *Lancet*, 342:193–8, 1993.

[13] J. C. Wyatt and D. G. Altman. Prognostic models: clinically useful or quickly forgotten? *BMJ*, 311:1539–1541, 1995.

[14] D. G. Altman, Y. Vergouwe, P. Royston, and K. G. M. Moons. Prognosis and prognostic research: validating a prognostic model. *BMJ*, 338:1432–1435, 2009.

[15] Y. Vergouwe, E. W. Steyerberg, M. J. Eijkemans, and J. D. Habbema. Validity of prognostic models: when is a model clinically useful? *Seminars in Urologic Oncology*, 20(2): 96–107, 2002.

[16] S. E. Bleeker, H. A. Moll, E. W. Steyerberg, A. R. Donders, G. Derksen-Lubsen, D. E. Grobbee, and K. G. Moons. External validation is necessary in prediction research: a clinical example. *Journal of Clinical Epidemiology*, 56:826–32, 2003.

[17] J. Concato, A. R. Feinstein, and T. R. Holford. The risk of determining risk with multivariate models. *Annals of Internal Medicine*, 118(201):210, 1993.

[18] R. Simon and D. G. Altman. Statistical aspects of prognostic factor studies in oncology. *British Journal of Cancer*, 69(6):979–985, 1994.

[19] D. G. Altman and G. H. Lyman. Methodological challanges in the evaluation of prognostic factors in breast cancer. *Breast Cancer Research and Treatment*, 52:289–303, 1998.

[20] D. B. Toll, K. J. M. Janssen, Y. Vergouwe, and K. G. M. Moons. Validation, updating and impact of clinical prediction rules: A review. *Journal of Clinical Epidemiology*, 61: 1085–1094, 2008.

[21] C. Chatfield. Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society, A*, 158(3):419–466, 1995.

[22] J. Hilden, J. D. Habbema, and B. Bjerregaard. The measurement of performance in probabilistic diagnosis. III. Methods based on continuous functions of the diagnostic probabilities. *Methods of Information in Medicine*, 17(4):238–46, 1978.

[23] M. Schemper and R. Henderson. Predictive accuracy and explained variation in Cox regression. *Biometrics*, 56:249–255, 2000.

[24] E. W. Steyerberg, A. J. Vickers, N. R. Cook, T. Gerds, M. Gonen, N. Obuchowski, M. J. Pencina, and M. W. Kattan. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*, 21(1):128–138, 2009.

[25] P. Royston and D. G. Altman. Visualizing and assessing discrimnation in the logistic regression model. *Statistics in Medicine*, 29(24):2508–20, 2010.

[26] A. C. Justice, K. E. Covensky, and J. A. Berlin. Assessing the generalisability of prognostic information. *Annals of Internal Medicine*, 130(6):515–524, 1999.

[27] E. W. Steyerberg, S. E. Bleeker, H. A. Moll, and D. E. Grobbee. Internal and external validation of predictive models: A simulation study of bias and precision in small sample. *Journal of Clinical Epidemiology*, 56:441–447, 2003.

[28] F. E. Harrell Jr., K. L. Lee, R. M. Califf, D. B. Pryor, and R. A. Rosati. Regression modeling stretegies for improved prognostic prediction. *Statistics in Medicine*, 3:143–152, 1984.

[29] P. Peduzzi, J. Concato, A. R. Feinstein, and T. R. Holford. The importance of events per independent variable in proportional hazards regression analysis. II. Acuracy and precision of regression estimates. *Journal of Clinical Epidemiology*, 48:1503–1510, 1995.

[30] P. Peduzzi, J. Concato, E. Kemper, T. R. Holford, and A. R. Feinstein. A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, 49:1373–1379, 1996.

[31] B. Efron and B. Tibshirani. *An introduction to the Bootstrap. Monographs on statistics and applied probability.* New York: Chapman & Hall, 1993.

[32] D. R. Cox. A note on data-splitting for the evaluation of significance levels. *Biometrika*, 62:441–444, 1975.

[33] P. J. Verweij and H. C. van Houwelingen. Cross-validation in survival analysis. *Statistics in Medicine*, 12:2305–2314, 1993.

[34] M. Schumacher, N. Hollandar, and W. Sauerbrei. Resampling and cross-validation techniques: a tool to reduce bias caused by model building. *Statistics in Medicine*, 12: 2305–2315, 1997.

[35] E. W. Steyerberg, F. E. Harrell Jr, G. J Borsboom, M. J. Eijkemans, Y. Vergouwe, and J. D. Habbema. Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. *Journal of Clinical Epidemiology*, 54(8):774–81, 2001.

[36] B. Efron. Estimating the error rate of a prediction rule: some improvements in cross-validation. *Journal of the American Statistical Society*, 78:316–31, 1983.

[37] The Fibrinogen Studies Collaboration. Measures to assess the prognostic ability of the stratified Cox proportional hazards model. *Statistics in Medicine*, 28:389–411, 2008.

[38] E. W. Steyerberg. *Clinical Prediction Models: A Practical Approach to Development, Validation and Updating.* Springer, New York., 2009.

[39] M. E. Miller, C. D. Langefeld, W. M. Tierney, S. L. Hui, and C. J. McDonald. Validation of probabilistic predictions. *Medical Decision Making*, 13:49–58, 1993.

[40] F. E. Harrell Jr. *Regression Modeling Strategies.* Springer, 2001.

[41] D. W. Hosmer and S. Lemeshow. A goodness-of-fit test for the multiple logistic regression model. *Communications in Statistics*, A10:1043–1069, 1980.

[42] D. R. Cox. Two further applications of a model for binary regression. *Biometrika*, 45: 562–565, 1958.

[43] H. C. van Houwelingen and J. Thorogood. Construction, validation and updating of a prognostic model for kidney graft survival. *Statistics in Medicine*, 14:1999–2008, 1995.

[44] H. C. van Houwelingen. Validation, calibration, revision and combination of prognostic survival models. *Statistics in Medicine*, 19:3401–3415, 2000.

[45] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143:29–36, 1982.

[46] F. E. Harrell Jr., R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati. Evaluating the yield of medical tests. *Journal of the American Medical Association*, 247:2543–46, 1982.

[47] B. W. Brown, M. Hollander, and R. M. Korwar. Nonparametric tests of independence for censored data, with application to heart transplant studies. *Reliability and Biometry*, pages 327–354, 1974.

[48] M. Gonen and G. Heller. Concordance probability and discriminatory power in proportional hazards regression. *Biometrika*, 92(4):1799–1809, 2005.

[49] P. Royston and W. Sauerbrei. A new measure in prognostic separation in survival data. *Statistics in Medicine*, 23:723–748, 2004.

[50] G. Blom. *Statistical Estimates and Transformed Beta-Variables*. Wiley: New York, 1991.

[51] M. Mittlbock and M. Schemper. Explained varition for logistic regression. *Statistics in Medicine*, 15:1987–97, 1996.

[52] B. Choodari-Oskooei, P. Royston, and M. K. B. Parmar. A simulation study of predictive ability measures in a survival model I: Explained variation measures. *Statistics in Medicine*, 30:00–00, 2011.

[53] M. Schemper and J. Stare. Explained variation in survival analysis. *Statistics in Medicine*, 15:1999–2012, 1996.

[54] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. of Math. Statist.*, 22:79–86, 1951.

[55] E. Graf, C. Schmoor, W. Sauerbrei, and M. Schumacher. Assessment and comparison of prognostic classfication schemes for survival data. *Statistics in Medicine*, 18:2529–2545, 1999.

[56] M. Schemper. Predictive accuracy and explained variation. *Statistics in Medicine*, 22: 2299–2308, 2003.

[57] B. H. Margolin and R. J. Light. An analysis of variance for categorical data. *Journal of the American Statistical Association*, 69:755–764, 1974.

[58] S. J. Haberman. Analysis of dispersion of the multinomial responses. *Journal of the American Statistical Association*, 77:568–580, 1982.

[59] H. C. van Houwelingen and S. Le Cessie. Predictive value of statistical models. *Statistics in Medicine*, 9:1303–1325, 1990.

[60] M. Schemper. The explained variation in proportional hazards regression. *Biometrika*, 77:216–218, 1990.

[61] J. T. Kent and J. O'Quigley. Measures of dependence for censored survival data. *Biometrika*, 75:523–534, 1988.

[62] D. R. Cox and E. J. Snell. *The Analysis of Binary Data*. Chapman & Hall: London, 1989.

[63] E. L. Korn and R. Simon. Measures of explained variation for survival data. *Statistics in Medicine*, 9:487–503, 1990.

[64] L. Magee. R-square measures based on Wald and likelihood ratio joint significance tests. *The American Statistician*, 44(3):250–253, 1990.

[65] N. J. D. Nagelkerke. A note on a general definition of the coefficient of determination. *Biometrika*, 78:691–692, 1991.

[66] J. O'Quigley, R. Xu, and J. Stare. Explained randomness in proportional hazards models. *Statistics in Medicine*, 24:479–489, 2005.

[67] P. Royston. Explained variation for survival models. *The Stata Journal*, 6(1):1–14, 2006.

[68] T. Hielscher, M. Zucknick, W. Werft, and A. Benner. On the prognostic value of survival models with application to gene expression signatures. *Statistics in Medicine*, 29(7-8): 818–29, 2009.

[69] T. A. Gerds and M. Schumacher. Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biometrical Journal*, 48:1029–1040, 2006.

[70] G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3, 1950.

[71] C. Schmoor, M. Olschewski, and M. Schaumacher. Randomized and non-randomized patients in clinical trials: experiences with comprehensive cohort studies. *Statistics in Medicine*, 15:263–271, 1996.

[72] W. Sauerbrei, P. Royston, H. Bojar, C. Schmoor, and M. Schumacher. Modelling the effects of standard prognostic factors in node-positive breast cancer. *British Journal of Cancer*, 79:1752–1760, 1999.

[73] D. R. Cox. Regression models and life table. *Journal of the Royal Statistical Society, Series B*, 34:187–220, 1972.

[74] M. E. Miller, S. L. Hui, and W. M. Tierney. Validation techniques for logistic regression models. *Statistics in Medicine*, 10(8):1213–26, 1991.

[75] M. J. Pencina and R. B. D'Agostino. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Statistics in Medicine*, 23:2109–2123, 2004.

[76] T. A. Gerds and M. Schumacher. Efron-type measures of prediction error for survival analysis. *Biometrics*, 63:1283–1287, 2007.

[77] M. J. van der Laan and J. M. Robins. *Unified Methods for Censored Lognitudinal Data and Causality*. Springer, 2003.

[78] A. Burton, D. G. Altman, P. Royston, and R. L. Holder. The design of simulation studies in medical statistics. *Statistics in Medicine*, 25:4279–4292, 2006.

[79] S. L. Zeger, K. Y. Liang, and P. S. Albert. Models for longitudinal data: A generalized estimating equation approach. *Biometrics*, 44:1049–1060, 1988.

[80] J. M. Neuhaus, J. D. Kalbfleisch, and W. W. Hauck. A comparison of cluster-specific and population-averaged approachs for analyzing correlated binary data. *International Statistical Review*, 59:25–36, 1991.

[81] R. Stiratelli, N. M. Laird, and J. R. Ware. Random-effects models for serial observations with binary response. *Biometrics*, 40:961–71, 1984.

[82] G. Y. Wong and W. M. Mason. The hierarchical logistic regression model for multilevel analysis. *Journal of the American Statistical Association*, 80:513–24, 1985.

[83] K. Y. Liang and S. L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73,:13–22, 1986.

[84] P. J. Diggle, K. Y. Liang, and S. L. Zeger. *Analysis of Longitudinal Data*. Oxford University Press, Oxford, 1994.

[85] Y. Lee and J. A. Nelder. Conditional and marginal models: another view. *Statistical Science*, 19(2):219–238, 2004.

[86] A. Skrondal and S. Rabe-Hesketh. Prediction in multilevel generalized linear models. *Journal of the Royal Statistical Society, Series A*, 172(3):659–687, 2009.

[87] S. Chebon. Do we need frailty modeling in the development of prognostic models? Master's thesis, Hasselt University, 2011.

[88] D. W. Hosmer and S. Lemeshow. *Applied Logistic Regression*. Wiley-Interscience Publication, 2nd edition, 2000.

[89] D. D. Dorfman and E. Alf. Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals-rating-method data. *Journal of Mathematical Psychology*, 6:487–496., 1969.

[90] C. E. Metz. Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, 8:283–298, 1978.

[91] H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18:50–60, 1947.

[92] D. Bamber. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, 12:387–415, 1975.

[93] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. New York, NY: John Wiley &Sons, 1958.

[94] B. Efron. The efficiency of logistic regression compared to Normal linear discriminant analysis. *Journal of the American Statistical Association*, 70:892–898, 1975.

[95] J. B. Copas. Regression, prediction and shrinkage. *Journal of Royal Statistical Society, Series B*, 45:342–354, 1983.

[96] Q. Liu and D. A. Pierce. A note on Gauss-Hermite quadrature. *Biometrika*, 81:624–629, 1994.

[97] J. C. Pinheiro and D. M. Bates. Approximation to the log-likelihood function in the non-linear mixed-effects model. *Journal of Computational and Graphical Statistics*, 4: 12–35, 1995.

[98] E. Lesaffre and B. Spiessens. On the effect of the number of quadrature points in a logistic random-effects model: an example. *Journal of the Royal Statistical Society, Series C*, 50: 325–335, 2001.

[99] J. C. Pinheiro and E. C. Chao. Efficient laplacian and adaptive gaussian quadrature algorithms for multilevel generalized linear mixed models. *Journal of Computational and Graphical Statistics*, 15:58–81, 2006.

[100] N. E. Breslow and D. G. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88:9–25, 93.

[101] H. Goldstein. *Multilevel Statistical Models*. London:Arnold, 1995.

[102] H. Goldstein and J. Rasbash. Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A*, 159:505–513, 1996a.

[103] T. R. Ten Have and A. R. Localio. Empirical Bayes estimation of random effects parameters in mixed effects logistic regression models. *Biometrics*, 55(4):1022–9, 1999.

[104] B. P. Carlin and T. A. Louis. Empirical Bayes: past, present and future. *Journal of the American Statistical Association*, 95:1286–1289, 2000a.

[105] B. P. Carlin and T. A. Louis. *Bayes and Empirical Bayes methods for data analysis*. Chapman and Hall-CRC, 2nd edition, 2000b.

[106] H. Goldstein and D. J. Spiegelhalter. League tables and their limitations: statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society, Series A*, 159:385–409, 1996b.

[107] I. H. Longford and T. Lewis. Outliers in multilevel data (with discussion). *Journal of the Royal Statistical Society, Series A*, 161:121–160, 1998.

[108] C. E. Rose, D. B. Hall, B. D. Shiver, M. L. Clutter, and B. Borders. A multilevel appraoch to individual tree survival prediction. *Forest Science*, 52:31–43, 2006.

[109] R. D. Gibbons, D. Hedeker, S. C. Charless, and P. Frisch. A random-effects probit models for predicting medical malpractice claims. *Journal of the American Statistical Association*, 89:760–767, 1994.

[110] P. J. Farrell, B. MacGibbon, and T. J. Tomberlin. Bootstrap adjustments for impirical Bayes interval estimates of small-area proportions. *Candian Journal of Statistics*, 25: 75–89, 1997.

[111] D. Afshartous and J. de Leeuw. Prediction in multilevel models. *Journal of Educational and Behavioral Statistics*, 30:109–139, 2005.

[112] R. Van Oirbeek and E. Lesaffre. An application of Harrell's C-index to PH frailty models. *Statistics in Medicine*, 29:3160 – 3171, 2010.

[113] R. DerSimonian and N. Laird. Meta analysis in clinical trials. *Controlled Clinical Trials*, 7:177–188, 1986.

[114] N. A. Obuchowski. Nonparametric analysis of clustered ROC curve data. *Biometrics*, 53: 567–578, 1997.

[115] B. Rosner and D. Grove. Use of the Mann-Whitney U-test for clustered data. *Statistics in Medicine*, 18:1387–1400, 1999.

[116] M. L. T. Lee and B. Rosner. The average area under correlated receiver operating characteristic curves: A nonparametric approach based on generalized two-sample Wilcoxon statistics. *Journal of the Royal Statistical Society, Series C*, 50:337–344, 2001.

[117] M. L. T. Lee and H. G. Dehling. Generalized two-sample U-statistics for clustered data. *Statistica Neerlandica*, 59(3):313–323, 2005.

[118] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44:847–844, 1988.

[119] S. Wieand, M. H. Gail, B. R. James, and K. L. James. A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika*, 76:585–592, 1989.

[120] H. E. Rockette, N. Obuchowski, C. E. Metz, and G. Gur. Statistical issues in ROC curve analysis. *SPIE, Medical Imaging IV: PACS System Design and Evaluation*, 1234:111–119, 1990.

[121] J. N. K. Rao and A. J. Scott. A simple method for the analysis of clustered binary data. *Biometrics*, 48:577–585, 1992.

[122] R. G. Miller Jr. *Survival Analysis*. New York, NY: John Wiley & Sons, 1981.

[123] G. W. Oehlert. A note on the delta method. *American Statistician*, 46:27–29, 1992.

[124] E. F. Schisterman, D. Faraggi, B. Reiser, and M. Trevisan. Statistical inference for the area under the receiver operating characteristic curve in the presence of random measurement error. *American Journal of Epidemiology*, 154(2):174–179, 2001.

[125] R. DerSimonian and R. Kacker. Random-effects model for meta-analysis of clinical trials: an update. *Contemporary Clinical Trials*, 28:105–114, 2007.

[126] R. J. Hardy and S. G. Thompson. A likelihood approach to meta-analysis with random effects. *Statistics in Medicine*, 15:619–29, 1996.

[127] A. Donner and J. J. Koval. The estimation of intra-class correlation in the analysis of family data. *Biometrics*, 36:19–25, 1980.

[128] H. Chakraborty, J. Moore, W. A. Carlo, T. D. Hartwell, and L. L. Wright. A simulation based technique to estimate intracluster correlation for binary variable. *Contemporary Clinical Trials*, 30:71–80, 2009.

[129] *Stata Corporation. Stata Statistical Software, Release 11. College station, Tex: Stata Corporation; 2010.*

[130] M. J. Campbell, R. M. Jacques, J. Fotheringham, R. Maheswaran, and J. Nicholl. Developing a summary hospital mortality index: retrospective analysis in english hospitals over five years. *BMJ*, to appear, 2012.

[131] R. V. Oirbeek and E. Lesaffre. Assessing the predictive ability of a multilevel binary regression model. *Computational Statistics & Data Analysis*, to appear, 2012.

[132] P. Hougaard. Frailty Models for Survival Data. *Lifetimes Data Analysis*, 1:255–273, 1995.

[133] T. M. Therneau and P. M. Grambsch. *Modeling Survival data: Extending the Cox model.* Springer-Verlag: New York, 2000.

[134] L. Duchateau and P. Janssen. *The Frailty Model.* Springer Science, Business Media, New York, 2008.

[135] L. J. Wei and L. Weissfeld. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*, 84:1065–1073, 1989.

[136] D. Y. Lin. Cox regression analysis of multivariate failure time data: the marginal approach. *Statistics in Medicine*, 13:2233–2247, 1994.

[137] T. Cai, L. J. Wei, and M. Wilcox. Semiparametric regression analysis for clustered failure time data. *Biometrika*, 87:867–878, 2000.

[138] P. Hougaard. *Analysis of multivariate survival data.* Springer-Verlag, New York., 2000.

[139] D. G. Clayton. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65: 141–151, 1978.

[140] J. P. Klein. Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics*, 48:795–806, 1992.

[141] G. Nielsen, R. Gill, P. Anderson, and T. Sorensen. A counting process approach to maximize likelihood estimation in frailty models. *Scandinavian Journal of Statistics*, 19: 25–43, 1992.

[142] T. A. Louis. Using empirical bayes methods in biompharmaceutical research. *Statistics in Medicine*, 10:811–829, 1991.

[143] G. E. Noether. *Elements of Nonparametric Statistics.* Wiley: New York, 1967.

[144] M. P. Wand and M. C. Jones. *Kernel Smoothing.* Chapman & Hall: London: Monographs on Statistics and Applied Probability, 1995.

[145] W. Hoeffding. A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, 19:293–325, 1948.

[146] A. J. Lee. *U-Statistics: Theory and practice.* Marcel Dekker Inc., New York, 1990.

[147] NIPORT, Mitra-Associates, and Macro International. Bangladesh Demographic and Health survey 2004. Technical report, National Institute of Population Research and Training (NIPORT); Dhaka, Bangladesh, and Calverton, Maryland, USA, 2005.

[148] NIPORT, Mitra-Associates, and Macro International. Bangladesh Demographic and Health Survey 2007. Technical report, National Institute of Population Research and Training (NIPORT); Dhaka, Bangladesh, and Calverton, Maryland, USA, 2009.

[149] G. Guo and G. Rodriguez. Estimating a multivariate proportional hazards model for clustered data using the EM Algorithm, with an application to child survival in Guatemala. *Journal of the American Statistical Association*, 87:969–976, 1992.

[150] M. Koenig, J. Phillips, O. Campbell, and S. D'Souza. Birth intervals and childhood mortality in rural Bangladesh. *Demography*, 27(2):251–265, 1990.

[151] E. Zenger. Siblings' neonatal mortality risks and birth spacing in Bangladesh. *Demography*, 30 (3):477–488, 1993.

[152] J. Trussell and C. Hammerslough. A hazards-model analysis of the covariates of infant and child mortality in SriLanka. *Demography*, 20 (1):1–26, 1983.

[153] J. Miller, J Trussell, A. Pebley, and B Vaughan. Birth spacing and child mortality in Bangladesh and the Philippines. *Demography*, 29 (2):305–318, 1992.

[154] N. Sastry. Family-level clustering of childhood mortality risk in Northeast Brazil. *Population Studies*, 51:245–261, 1997.

[155] N. Sastry. A nested frailty model for survival data, with an application to the study of child survivalin Northeast Brazil. *Journal of the American Statistical Association*, 92: 426–435, 1997.