Statistical tools for risk prediction models

Mortensen, Rikke Nørmark

DOI (link to publication from Publisher):
10.5278/vbn.phd.med.00120

Publication date:
2018

Document Version
Publisher's PDF, also known as Version of record

STATISTICAL TOOLS FOR RISK PREDICTION MODELS

BY
RIKKE NØRMARK MORTENSEN

DISSERTATION SUBMITTED 2018

AALBORG UNIVERSITY
DENMARK

Statistical tools for risk prediction models

Ph.D. Dissertation

Rikke Nørmark Mortensen

Dissertation submitted September, 2018

Dissertation submitted: September 30, 2018

PhD supervisors: Professor Thomas Alexander Gerds
Department of Biostatistics
University of Copenhagen

Professor Christian Torp-Pedersen
Unit of Epidemiology and Biostatistics
Aalborg University Hospital

PhD committee: Professor Martin Bøgsted (Chairman)
Aalborg University

Associate Professor Georg Heinze
Medical University of Vienna

Professor Jacob von Bornemann Hjelmborg
University of Southern Denmark

PhD Series: Faculty of Medicine, Aalborg University

Institut: Department of Clinical Medicine

ISSN (online): 2246-1302
ISBN (online): 978-87-7210-335-8

Published by:
Aalborg University Press
Langagervej 2
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk

Printed in Denmark by Rosendahls, 2018

Preface

This dissertation i based on studies carried out during my employment at Unit of Epidemiology and Biostatistics, Aalborg University Hospital.

This work was made possible due to a number of persons. First and foremost I would like to thank my two supervisors: Thomas Alexander Gerds for his outstanding professional competence and valuable inputs, for being very inspirational and always very helpful. Christian Torp-Pedersen for taking the fight with arduous reviewers, for always encouraging me to continue, and for providing a very inspirational working environment.

My sincerest thanks to all my colleagues at Unit of Epidemiology and Biostatistics, Aalborg University Hospital for providing a pleasant and inspiring work environment. A special thanks to Lone Brændgaard Hansen for the past five years we have been colleagues.

Finally, my warmest thanks to my family and my boyfriend Gerald for their support and patience.

Rikke Nørmark Mortensen September 30, 2018

List of papers

Paper I

Rikke Nørmark Mortensen, Thomas Alexander Gerds, Jørgen Lykke Jeppesen, and Christian Torp-Pedersen (2017). Office blood pressure or ambulatory blood pressure for the prediction of cardiovascular events. *European Heart Journal* 38(44): pp 3296-3304.

Paper II

Rikke Nørmark Mortensen, Tianxi Cai, and Thomas Alexander Gerds. Uncertainty of bootstrap cross-validation estimates of prediction performance in censored survival data. *(Manuscript)*

Paper III

Rikke Nørmark Mortensen, Michael Mørk Petersen, Michala Skovlund Sørensen, and Thomas Alexander Gerds. On risk predictions from the logistic recalibration method. *(Manuscript)*

Abstract

Accurate prediction of a patient's risk of a future fatal- or non-fatal event is important to secure an accurate treatment of the patient. For this reason, development of risk prediction models is a task of great interest in medical research. Depending on the research question at hand, the motivation for developing a prediction model may have different purposes; for example, it may be of interest to evaluate the predictive value of a new biomarker over established risk factors, or the aim may be to develop a prediction model that can assist in clinical practice to inform patients on the probability of a prognostic outcome. A statistical prediction model uses combinations of common, clinical risk factors to predict the risks of future events in individual patients. To secure accurate risk predictions, validation of the prediction models is needed. For this task, various statistical methods have been suggested in the literature, including the Brier score and AUC. The main aim of this thesis is to make contribution to the research area of prediction models in medical applications.

The first research paper in the thesis is a study on a large data set regarding the predictive value of various blood pressure measurements on the risk of future cardiovascular events. This paper illustrates how the accuracy of statistical prediction models can be used to assess the predictive value of a biomarker.

The second study in the thesis considers internal validation of statistical prediction models in the setting of right-censored survival data. In this study, we first review a well-known leave-one-out bootstrap estimator of the prediction error in a binary outcome setting. Then, we introduce an extension of this estimator to estimation of the time-dependent Brier score and the time-dependent AUC in right-censored survival data. We derive the influence function for these estimates and show how one can get the corresponding standard errors.

In the last study we first review a logistic re-calibration approach that provide updated risk estimates when a prediction model is applied in a setting different from the one in which the model was developed. We then demonstrate how to construct confidence intervals of the updated risk estimates.

Abstract

Resumé

En nøjagtig prædiktion af en patients risiko for en given fremtidige fatal eller ikke fatal hændelse er nødvendig for at sikre den korrekte behandling af patienten. Af denne grund er interessen for at udvikle statistiske prædiktionsmodeller stor, blandt forskere indenfor den medicinske verden. Alt efter det givne forskningsspørgsmål, kan interessen for at udvikle en prædiktionsmodel være forskellig; der kan for eksempel være interesse i at evaluere den prædiktive værdi af en ny biomarkør. Der kunne også være interesse for at udvikle en model, der skal bruges i klinisk praksis til at informere patienter om deres risiko for en fremtidige hændelse hos den enkelte patient. I en statistisk prædiktionsmodel kombineres forskellige risikofaktorer for at prædiktere risikoen for en fremtidig hændelse. Fr at sikre nøjagtige risiko prædiktioner er det nødvendigt at validere en statistisk prædiktionsmodel. Til dette formål er en række statistiske metoder blevet foreslået i litteraturen, bl.a. Brier scoren og AUC. Formålet med denne afhandling er at bidrage til det biostatistiske forskningsområdet indenfor prædiktionsmodeller.

Den første artikel i denne afhandling er et studie på et stor datasæt omhandlende den prædiktive værdi af forskellige blodtryksmålinger på den fremtidige risiko for at udvikle kardiovaskulære komplikationer. I denne artikel er det demonstret hvordan nøjagtigheden af en statistisk prædiktionsmodel kan benyttes, til at bestemme den prædiktive værdi af en given biomarkør.

Den anden artikel i denne afhandling omhandler "internal validation" af statistisk prædiktionsmodeller i højre-censureret overlevelses data. I denne artikel bliver der først gjort rede for leave-one-out boostrap estimatoren for binære outcome studier. Derefter foreslås en videreudvikling af denne estimator til estimation af Brier scoren og AUC i højre-censureret overlevelses data. I artiklen udvikles influence funktionen og det vises hvordan denne kan bruges til at estimere standard error for de to bootstrap estimater.

I den sidste artikel redegøres først for en opdateringsmetode "logistic recalibration" der kan benyttes til at opdatere en prædiktionsmodel fra en population til en anden. I dette studie viser vi hvordan man kan konstruere konfidensintervaller for de opdaterede risiko estimater.

Contents

Chapter 1

Introduction

This thesis is about prediction models that predict the risk of a future event given a set of baseline risk factors. Prediction models of this type are often encountered in clinical settings where they are used to aid the clinical decision-making [78]. Examples of these include the Framingham risk score for cardiovascular events [93], the CHA_2DS_2VASc score for predicting the risk of stroke, and the Gail model for predicting the risk of breast cancer [60]. In recent years there has also been an increasing interest in detection new biomarkers that can improve decision-making in clinical settings [65, 76]. A prediction model can broadly be defined as either a statistical model or a data mining algorithm that has been trained in a given data set with the purpose of predicting the outcome for future observations [11, 75].

Within the field of medical statistics there has historically been a tradition for statistical modeling with the aim of explaining a possible (causal) association between the risk factors and the the likelihood of the event, i.e., etiological studies. In particular, the Cox proportional hazards model has been used in a countless number of studies to obtain hazard ratios indicating the relative effect of a variable on the hazard rate of the event of interest. Although the mathematical models used in prediction studies often are the same as those used in etiological studies the purpose of the modeling processes is quite different. While etiological studies focus on finding causal associations in the average patient, prediction studies focus on developing the model that best predicts the outcome for individual persons. These quite different purposes have consequence for the way the selection of risk factors is approached and the way the models are validated [11]. Etiological studies are often concerned with terms like *confounding* and *effect modification* and all variables that are thought to be associated with the outcome are usually included in the model. Whereas, in prediction studies a variable is thought to be of value *if it improves our ability to predict the outcome*. Thus, a variable that

is causally related to the outcome may not be of interest in a prediction study if the magnitude of its effect is too small too improve the model's ability to predict the outcome to a clinically meaningful extent.

The process of developing a prediction model has three basic steps: 1) The first step is to develop the model. In this step a model is fitted to a given data set (the learning data), this step includes common modelling strategies such as variable selection and parameter estimation. 2) The second step is to perform internal validation. This step includes assessing the accuracy of the estimated event probabilities in the learning data. Measures of prediction accuracy and different strategies for doing internal validation will be discussed in more details later. 3) The last step is to do external validation, that is, to test the accuracy of the prediction model in a data set (the validation data) that is independent of the learning data. It is quite expected that the model will perform worse in an independent validation data than in its own (learning) data. The reason for this is simply that the model parameters are optimized in the learning data.

This thesis will give a broad treatment of the topic risk prediction models in medical applications, with a special focus on validation of risk prediction models. Throughout the thesis it will be assumed that the risk factors are measured at some baseline date, and the event can occur at any point in time hereafter. The task in prediction studies is then to predict the probability that the event will occur prior to some pre-specified prediction horizon $t > 0$. In medical applications it is quite often the case that the learning data and/or the validation data contain right-censored event times. That is, some subjects are lost to follow-up prior to the prediction horizon t and the event status at time t is thus unknown for these subjects. Another issue often encountered in medical application is the existence of competing risks. In analyses of survival data it is well-recognized that simply ignoring censored event times and/or competing risks may seriously bias the results. Various modeling strategies have been developed to properly analyze this type of data [2]. Analogously, estimates of prediction accuracy measures that properly account for censored data and competing risks have been developed in the past decades [33, 70].

The outline of this thesis is as follows: The aims of the thesis are given in chapter 2. Chapter 3 presents an overview of risk prediction in medical applications, introducing some of the most popular methods for validating statistical risk prediction models. The mathematical background used in the three papers is presented in chapter 4. Chapter 5 gives a summary of the three papers. A final discussion of relevant themes in the thesis is presented in chapter 6. Lastly, a discussion of the perspectives for future work is given in chapter 7.

Chapter 2

Aims of the thesis

Within the field of medical statistics there is a still increasing interest in identifying combinations of biomarkers that can provide an accurate prognosis for a patient. The overall aim of this thesis is to make contributions to this research area by emphasizing the accuracy of risk prediction models and by suggesting statistical tools that can be used in the analysis of such models.

The specific aims of the thesis are:

- To explain that significant hazard ratios do not necessarily translate into improved risk predictions in particular when there are competing risks.
- To study asymptotic inference and to provide confidence intervals for the internally validated performance of statistical risk prediction models (and differences thereof) when the outcome is a right-censored time-to-event.
- To provide asymptotic inference and to provide confidence intervals for risks predicted by a statistical model after re-calibration to a new population.

Chapter 3

The performance of risk prediction models

This chapter gives an overview of methods used to validate risk prediction models. The focus will be on the the measures most relevant to this thesis, however references to other related measures will be given. Throughout $\tilde{T}$ will denote the time to event, $Z \in \mathbb{R}^p$ will be a set of risk factors, and $t > 0$ will be a fixed prediction horizon. Further, $F(u|Z) = \mathrm{P}_n(\tilde{T} \leq u|Z)$ will be the conditional distribution function of $\tilde{T}$ given Z and $S(u|Z) = 1 - F(u|Z)$ will denote the corresponding survival function.

Consider a data set $\mathcal{L}_n$ consisting of data for $i = 1, \ldots, n$ i.i.d. subjects. A prediction modeling strategy R_t is a map on the set of datasets into the set of prediction models. Estimating the internal parameters of R_t in $\mathcal{L}_n$ yields a prediction model $R_{t,n} = R_t(\mathcal{L}_n)$. A risk prediction model is a mapping $R_{t,n} : \mathbb{R}^p \to [0, 1]$ that for subject i at time t provides an absolute risk estimate $R_{t,n}(Z_i)$. In medical applications popular modeling strategies include the logistic regression model and the Cox proportional hazards model. However, one may also consider more flexible modeling strategies such as penalized regression models or a random survival forest. An overview of these methods can be found in e.g. [13, 44, 78].

1 Time-dependent prediction performance

To ensure valid risk estimates, it is desirable to assess the accuracy of a trained prediction model. For this task various performance measures have been suggested. Performance measures are usually categorized according to whether they measure *discrimination* or *calibration*. Discrimination refer to

the models ability to discriminate high-risk patients from low-risk patients, while calibration refer to the agreement between the estimated risk probabilities and the actual event status. An important class of measures that falls somewhat outside of these two categories is the class of scoring rules that is given as the expected value of a suitable loss function [49]. This class scores the estimated probabilities directly to the event status, and may thus be considered as measures of the overall performance of the prediction model. In medical application the most important scoring rule is the Brier score that is characterize by a quadratic loss function. Details on the Brier score will be given in section 1.1 in this chapter.

The measures considered are time-depended because they measure the performance of the model at a fixed prediction horizon t. In applied settings it may be of interest to estimate the value of a performance measure at more than one time point. For example if one is interested in the models ability to predict the short-term risk and the long-term risk of a given event. In this case it is very likely that the value of the performance measure for prediction horizon, say 1 year is different from the value of the performance measure for prediction horizon 10 year.

Regardless whether one is interested in a model's overall performance, discrimination ability, or calibration it is important to distinguish between *internal validation* and *external validation*. Internal validation is the assessment of the model performance in the data in which the model was developed (the learning data), and external validation is the assessment in a data (the validation data) that is independent of the learning data. Internal validation will be discussed section 3 in this chapter. It is usually the case that the model performs a lot better in the learning data than in the validation data. Th reason for this is simply that the model parameters are optimized in the learning data. In this section and section 2 it will assumed that the model is developed in a data $\mathcal{L}_n$ and the model performance is assessed in an independent validation data $\mathcal{V}_m$ consisting of m i.i.d. observations. As a starting point it is assumed that the event time $\tilde{T}$ is not censored, and the outcome in the analysis is given by the event indicator $Y(t) = \mathcal{I}_{\{\tilde{T} \leq t\}}$; here $\mathcal{I}_{\{\cdot\}}$ denotes the indicator function. In section 2 extensions to right-censored survival data is discussed.

1.1 Brier score

The Brier score was originally introduced as a measure to verify a weather forecast [12]. Subsequently the Brier score has become a popular measure of the mean squared error of a prediction model. A time-dependent version of the Brier score is given as the expected squared difference between the

observed event status at time t and the estimated event probability [36, 40]:

$$Brier_t(R_{t,n}) = \mathbb{E}[\{Y_j(t) - R_{t,n}(Z_j)\}^2].$$

The Brier score can be interpret as the loss incurred by assigning probability $R_{t,n}(Z_j)$ of event at time t for a subject with event status $Y_j(t)$. An empirical estimate of the Brier score is given by:

$$\widehat{Brier}_t(R_{t,n}, \mathcal{V}_m) = \frac{1}{m}\sum_{j=1}^{m}\{Y_j(t) - R_{t,n}(Z_j)\}^2.$$

1.2 Discrimination

A good prediction model must accurately discriminate between subjects with event before time t and subjects with no event before time t. That is, subjects with event prior to time t must be assigned a probability of event that is higher than subjects without event at time t.

A popular method for describing the discrimination ability of the model is to summaries the correct classification rates defined by the sensitivity $\mathrm{P}_n\big(R_{t,n}(Z_i) > c | Y_i(t) = 1\big)$ and the specificity $\mathrm{P}_n\big(R_{t,n}(Z_i) \leq c | Y_i(t) = 0\big)$ [46, 47]. Here $c \in [0,1]$ is threshold that is used to determine whether subject i is classified as a *case* (i.e. event before time t) or a *control* (i.e. event after time t). It is common to depict the correct classification rates for the full spectrum of cutoff points by considering the Receiver Operating Characteristic (ROC) curve. This curve is a plot of the true positive rate (TPR) against the false positive rate (FPR) for consecutive cutoff points. Here, TPR is just another name for sensitivity and FPR is defined as 1-specificity, thus $FPR_t(R_{t,n}, c) = \mathrm{P}_n\big(R_{t,n}(Z_i) > c | Y_i(t) = 0\big)$. The ROC curve is given by the graph:

$$ROC_t\big(R_{t,n}, \cdot\big) = \big\{\big(FPR_t(R_{t,n}, c), TPR_t\big(R_{t,n}, c\big)\big) : c \in [0,1]\big\}.$$

The ROC curve provides a useful tool to visualize the performance of a model. If the prediction model discriminate perfect between cases and controls, that is, if $R_{t,n}(Z_i) > R_{t,n}(Z_j)$ for all pair of subjects (i,j) in $\mathcal{V}_m$ with $Y_i(t) = 1$ and $Y_j(t) = 0$, then $TPR_t(R_{t,n}, c) = 1$ for all values of c for which $FPR_t(R_{t,n}, c) > 0$. In contrast, if the model is useless for discriminating between cases and controls, then the distribution of predicted risks will be identical among cases and controls and the ROC curve will lie on the 45° line. Figure 3.1 shows the ROC curve for two different prediction models that predict the same outcome. The plot shows that model A is a better model than model B because the curve for model A lies above the curve for model B for all values of c. A measure that is closely related to the ROC curve is the area

under the ROC curve (AUC) [42, 66] which is defined by:

$$AUC_t(R_{t,n}) = \int_0^1 ROC_t(R_{t,n}, c)\mathrm{d}c.$$

The scale of the AUC ranges from 0.5 (no discrimination ability) to 1 (perfect discrimination ability). It can be shown [66] that the AUC can be interpret as

$$AUC_t(R_{t,n}) = \mathrm{P}_n\big(R_{t,n}(Z_i) > R_{t,n}(Z_j) | Y_i(t) = 1, Y_j(t) = 0\big)$$

In the binary outcome setting considered in this section, this definition of the AUC is identical to the measure called Hareell's c-statistic [45]. An estimate of the AUC is given by

$$\widehat{AUC}_t(R_{t,n}, \mathcal{V}_m) = \frac{\sum_{i=1}^m \sum_{j=1}^m \mathcal{I}_{\{R_{t,n}(Z_i) > R_{t,n}(Z_j)\}} \mathcal{I}_{\{Y_i(t)=1, Y_j(t)=0\}}}{\sum_{i=1}^m \mathcal{I}_{\{Y_i(t)=1\}} \sum_{j=1}^m \mathcal{I}_{\{Y_j(t)=0\}}}$$

Various other measures of discrimination has recently been proposed [80] including the Net Reclassification Improvement (NRI) and integrated discrimination improvement (IDI) [64]. However, as pointed out by [48] these measures are not proper performance measures [39, 59] and must thus be used with care.

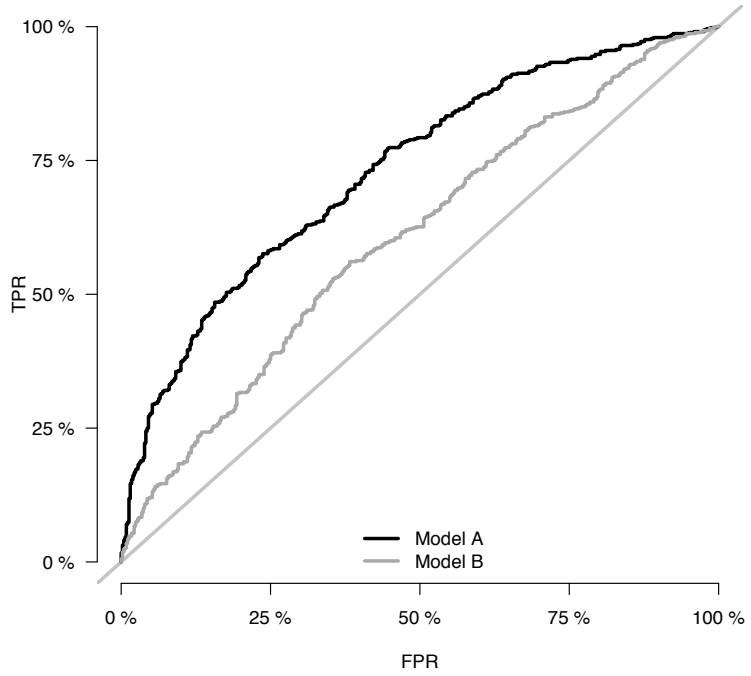

Fig. 3.1: Example of ROC curve for two prediction models

1.3 Calibration

Calibration of a prediction model refer to the extend of bias in the predicted event probabilities, that is, how close the estimated risk are to the true event probabilities [45]. A prediction model is said to be well-calibrated if the model provides risk estimates that coincide with the expected proportion of events [44]. That is, if the model assigns a 25% risk of event then the event should occur in approximately 25 out of 100 patients who all received an estimated event probability of 25%. In a more formal manner the concept of a well-calibrated model can be defined as [32]:

$$\mathbb{E}[Y(t)|R_{t,n}(Z) = p] = p, \quad \text{for } p \in [0,1]$$

Calibration of a prediction model is in particular important if the aim is to develop a model that is intended to inform patients on their future risk of some given event.

One method to assess the calibration of a prediction model is to use the Hosmer-Lemeshow test [50]. This test is computed by dividing the observations in g groups e.g. percentiles. For group $i = 1, \ldots, g$, let m_i be number of observations in the group, let O_i be the observed number of events in the group, and let $\bar{\pi}_i$ be the average estimated event probability in group i. The Hosmer-Lemeshow test is then calculated as the Pearson chi-square statistic from the table of observed and expected frequencies

$$\chi_{HL} = \sum_{i=1}^{g} \frac{(O_i - m_i\bar{\pi}_i)^2}{m_i\bar{\pi}_i(1-\bar{\pi}_i)}$$

A main limitation of this test is that it is based on a arbitrate grouping of the observations. Another popular method to assess calibration is to consider the calibration plot in which the observed event status is plotted against the predicted risks estimates. To assist the interpretation, the plot is often equipped with a calibration curve that can be obtained from a smoothing technique such as the loess curve or using a kernel [32, 58]. For a calibrated model the calibration curve will lie on the 45° line.

2 Extensions to censored survival data

The performance measures introduced in the previous section apply to the case of binary outcome data and uncensored survival outcome. However in medical applications the analysis is often complicated by right-censored survival data and possible competing risks. In this section some comments regarding estimation of performance measures in this situation will be given. The focus will be on the Brier score and the AUC as these measures are most

important in relation to the papers in this thesis. Let C be a right-censoring time and denote by $G(u|Z) = \mathrm{P}_n(C > u|Z)$ the conditional survival function. In this section it is assumed that the data $\mathcal{V}_m$ contains survival data for m i.i.d. observations. That is, for each subject j the observation consists of the triplet $X_j = (T_j, \Delta_j, Z_j)$; here $T_j = \min(\tilde{T}_j, C_j)$ and $\Delta_j = \mathcal{I}_{\{\tilde{T}_j \leq C_j\}}$. Throughout it will be assumed that the event time and the censoring time is conditional independent given the covariates Z, that is, it is assumed that:

$$\tilde{T} \perp\!\!\!\perp C \mid Z \quad \text{(conditional independence)}. \tag{A1}$$

It is well known that ignoring censored event times in analyses will induce loss of information and result in biased estimates. For this reason the estimates of the Brier score and the AUC (and other measures) must be modified in order to proper account for the loss of information induced by censored data. One of the first attempts to account for censoring was initiated by Korn and Simon [56] who introduced a model-based estimator of the explained variation, which is a measure closely related to the Brier score [41]. This model-based estimation method relies on a correctly specified model of the survival function of event times. Model-based estimators have also been suggested to estimate the AUC [14, 77]. In the papers in this thesis we have focused on estimation of accuracy measures based on the inverse probability of censoring weighting (IPCW) technique. The main idea of this technique is to weight each observation in the data with its probability of not being censored. The method is discussed in more details in chapter 4. An IPCW estimate of the Brier score were introduced in [40] and was further studied in [36]. Let $\hat{G}_m$ be an estimate of G. The IPCW estimate of the Brier score is given by

$$\widehat{Brier}_t^{ipcw}(R_{t,n}, \mathcal{V}_m) = \frac{1}{m}\sum_{j=1}^{m}\{\mathcal{I}_{\{T_j \leq t\}} - R_{t,n}(Z_j)\}^2 W_t(X_j, \hat{G}_m), \tag{3.1}$$

where the weights are defined by

$$W_t(X_j, \hat{G}_m) = \frac{\mathcal{I}_{\{T_j \leq t\}}\Delta_j}{\hat{G}_m(T_j - |Z_j)} + \frac{\mathcal{I}_{\{T_j > t\}}}{\hat{G}_m(t|Z_j)}.$$

Gerds and Schumacher [36] showed that the estimate (3.1) is a consistent estimate of the Brier score provided that $G(t|Z) > \epsilon > 0$ and $\hat{G}_m$ is correctly specified. Various authors [7, 51, 85] have likewise suggested an IPCW estimate of the AUC:

$$\widehat{AUC}_t^{ipcw}(R_{t,n}, \mathcal{V}_m) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{m}\mathcal{I}_{\{R_{t,n}(Z_i) > R_{t,n}(Z_j)\}}\mathcal{W}_t(X_i, X_j, \hat{G}_m)}{\sum_{i=1}^{m}\left[\mathcal{I}_{\{T_i \leq t\}}\Delta_i\hat{G}_m^{-1}(T_i - |Z_i)\right]\sum_{j=1}^{m}\left[\mathcal{I}_{\{T_j > t\}}\hat{G}_m^{-1}(t|Z_j)\right]}$$

where the weights are defined by

$$\mathcal{W}_t(X_i, X_j, \hat{G}_m) = \frac{\mathcal{I}_{\{T_i \leq t\}} \Delta_i \mathcal{I}_{\{T_j > t\}}}{\hat{G}_m(T_i - |Z_i) \hat{G}_m(t|Z_j)}.$$

In the previous section it was noted that the definition of the AUC coincide with the definition of Harrells c-index. In the case of survival data the definition of the two discrimination measures does not coincide [34], and one may choose to use the c-index as a measure of discrimination. However, as pointed out by Blanche et al [8], the definition of the c-index for survival data does not use the prediction horizon twhen discriminating between cases and controls. This means that the measure is in general not proper for assessing the predictive value of the model at time t.

2.1 Competing risks

Competing risks are often encountered in applied settings where the event of interest is not all-cause mortality [1]. One example is the risk of lung cancer among people who smoke. These people are indeed at high risk of getting lung cancer, however, they might also die due to other reasons before they get lung cancer. In this case cancer-unrelated mortality is a competing risk of the event lung cancer. A naive approach would be to analysis such data by censoring people when they die to reasons not related to the event of interest. However, this approach would correspond to analyzing a population in which you cannot die before you have had the disease, which is not applicable in most realist settings.

Let $\eta \in \{1, 2\}$ be an event indicator of two competing events; here, without loss of generality, it is assumed that only two events can happen. Though, in applied settings it is of course possible that more competing events can happen. The data for subject j is given by the triplet $X_j = (T_j, \tilde{\eta}_j, Z_j)$, here $\tilde{\eta}_j = \eta_j \Delta_j$. Suppose $\eta = 1$ is the event of interest. The aim of a prediction study is then to predict the cumulative incidence function

$$F_1(t|Z) = \mathrm{P}_n(\tilde{T} \leq t, \eta = 1|Z)$$

Traditional regression models for the competing risks setup include the Fine-Gray model and cause-specific Cox regression, an overview of these models can be found in [35]. In what follows $R^1_{t,n}(Z_j)$ will denote a risk estimate of $F_1(t|Z_j)$ for subject j.

In the case of competing risks the Brier score is defined by [74]:

$$Brier_t(R^1_{t,n}) = \mathbb{E}[\{\mathcal{I}_{\{\tilde{T}_j \leq t, \eta_j = 1\}} - R^1_{t,n}(Z_j)\}^2] \tag{3.2}$$

This definition is a straight forward extension of the Brier score defined in section 1. Shoop et al. [74] suggested an IPCW estimate of the Brier score

(3.2) that is much similar to the estimate in (3.1). One extension of the AUC to competing risks is given by [6, 95]:

$$AUC_t(R^1_{t,n}) = \mathrm{P}_n(R^1_{t,n}(Z_i) > R^1_{t,n}(Z_j)|\tilde{T}_i \leq t, \eta_i = 1, \{\tilde{T}_j > t\} \cup \{\tilde{T}_j \leq t \wedge \eta_j \neq 1\}) \tag{3.3}$$

This definition defines a *case* as a subject with event of interest prior to the prediction horizon t and a *control* as any subject who either is event-free prior to time t or who experienced the competing event prior to time t. This definition of the AUC is used in paper I in this thesis. Another definition of the AUC for competing risks is:

$$AUC^*_t(R^1_{t,n}) = \mathrm{P}_n(R^1_{t,n}(Z_i) > R^1_{t,n}(Z_j)|\tilde{T}_i \leq t, \eta_i = 1, \tilde{T}_j > t)$$

This defines a case as about, however, a control in this definition is a subject without any event prior to time t. Dependent on the research question at hand, one may choose any of the two definitions. IPCW estimates for these definitions of the AUC were suggested by Blanche et. al [6].

3 Internal validation

In the previous sections it was assumed that the risk prediction model was developed in a learning data $\mathcal{L}_n$, and the performance measures were estimated in a validation data $\mathcal{V}_m$. This mode of validating a prediction model is called external validation. Another mode of validation is internal validation which means to estimate the accuracy of the prediction model in the learning data. In this section a generic notation for the performance measures will be used in order to accommodate all measures discussed in the previous sections. In what follows $Q_t(R_{t,n}, \mathcal{L}_n)$ will denote a performance measure, e.g. the Brier score, for the prediction model $R_{t,n}$ estimated in data $\mathcal{L}_n$.

Estimating the performance measure in the same data as the model was developed is called the apparent performance:

$$Q^{App}_t(R_t) = Q_t(R_{t,n}, \mathcal{L}_n)$$

As parameters of a trained model are optimized in the learning data, the apparent performance will usually provide overoptimistic results regarding accuracy of risk predictions in patients outside the learning data [25]. The degree of overoptimism will increase with the complexity of the prediction model as complex models will utilize information in the learning data to a greater extent than more parsimonious models. In general, a prediction model is said to be overfitted if the model specification is too complex to capture the true underlying data structure.

Overfitted models are likely not to perform well in patients outside of the learning data, and thus it is of interest to detect a model's degree of

overfitting. For this task it is common to use a cross-validation algorithm. The usual scheme in a cross-validated algorithm is to repeatedly split the data $\mathcal{L}_n$ into a training part and a validation part. For each split, the model is fitted in the training part and the performance measure is estimated in the validation part. Accordingly, a cross-validation algorithm imitates a process in which the prediction model is applied to future patients and the prediction accuracy evaluated in these new patients. The final step in a cross-validation algorithm is to average the estimated accuracy measures over multiple splits of $\mathcal{L}_n$. This last step means that a cross-validated accuracy measure provides an (often biased) estimate of the expected model performance over all possible learning sets of size n:

$$\mathbb{E}_{\mathcal{L}_n}[Q_t(R_{t,n}, \mathcal{L}_n)].$$

A traditional cross-validated estimate is the K-fold cross-validated estimate that is obtained by random splitting the data $\mathcal{L}_n$ into K mutually exclusive subsamples $\mathcal{L}_1, \ldots, \mathcal{L}_K$ of approximately equal size. Let $R_{t,n-k} = R_t(\mathcal{L}_{-k})$ be the modelling strategy R_t trained on $\mathcal{L}_{-k} = \mathcal{L}_n \setminus \mathcal{L}_k$, for $k = 1, \ldots, K$. The idea is to train the model in $\mathcal{L}_{-k}$ and evaluate the performance measure in $\mathcal{L}_k$. The K-fold cross-validated estimate is then given by

$$Q^{CVK}(R_t) = \frac{1}{K} \sum_{k=1}^{K} Q(R_{t,n-k}, \mathcal{L}_k)$$

A popular choice is to set $K = 10$. The main advantages of this estimate is that it is computationally efficient, provided K is not "too large". This is in particular an advantages if it is time consuming to fit the prediction modelling strategy. However, the K-fold cross-validated estimate is usually subject to high Monte-Carlo variance [10, 57, 92]. Setting $K = n$ defines the classic leave-one-out cross-validated estimator [31, 81].

Another popular cross-validation algorithm is the bootstrap cross-validation approach [24, 28, 29]. This approach is based on the general theory of bootstrapping introduced by Efron [23, 26, 27]. In the bootstrap cross-validated algorithm the data is splitted by drawing bootstrap samples from $\mathcal{L}_n$. That is, let P_n be the empirical measure of the data $\mathcal{L}_n$. A bootstrap sample is a random sample $\mathcal{L}^*$ of size n from P_n. In other words a bootstrap sample is a sample $X_1^*, \ldots, X_n^*$ drawn with replacement from $\mathcal{L}_n$. The bootstrap cross-validated estimate is obtained by drawing B bootstrap samples $\mathcal{L}_1^*, \ldots, \mathcal{L}_B^*$, for some large number B. In each run of the algorithm the prediction strategy is trained in $\mathcal{L}_b^*$ yielding a trained prediction model $R_{t,b}^* = R_t(\mathcal{L}_b^*)$ and the performance measure is evaluated in people how are out-of-bag i.e. $\mathcal{L}_b^0 = \{X_i : X_i \notin \mathcal{L}_b^*\}$. Averaging over all bootstrap samples yield the boot-

strap cross-validated estimate:

$$Q^{bootCV}(R_t) = \frac{1}{B}\sum_{b=1}^{B} Q(R^*_{t,b}, \mathcal{L}^0_b)$$

In paper II of this thesis we considered a bootstrap cross-validated estimate that is defined a bit different, though comparable to the bootstrap cross-validated estimate as defined above.

Chapter 4

Mathematical background

This chapter discusses two specific mathematical techniques that played a central role in the analysis of the aims of the thesis. The first is the inverse probability of censoring weighting technique which we applied to deal with censoring, the second is the functional delta method which we applied to discuss the asymptotic inference of the estimators in papers II and III.

1 Inverse probability of censoring weighting

The method of inverse probability of censoring weighting (IPCW) is a general weighting scheme that can be used to overcome the problem of censored survival data in estimating problems. The method was originally introduced in connection with a general methodology for parameter estimation in non- and semi-parametric models with coarsed data, i.e. missing or censored data [71, 72]. The ideas behind this methodology is based on an inverse probability weighting technique introduced in the early 1950s [42] combined with general results from semi-parametric theory [5, 84]. The method has subsequently been used in various estimation problems, including estimates of performance measures for prediction models as discussed in chapter 3, and estimation of parameters in regression models [30, 35, 73].

The intuitive reasoning behind inverse probability weighting is as follows. Let $p(Z)$ be the probability that a person, with risk factors Z, has complete data; here complete data means that the outcome variable is non-missing, while the risk factors are assumed never to be missing. A person with complete data and risk factors Z can thus be assumed to represented $1/p(Z)$ of the intended study population (in which some subjects may have missing data). This suggest a weighting scheme that put weight $1/p(Z)$ on subjects with complete data and covariates Z. Specially, suppose $p(Z) = 0.5$, intuitively this means that subjects with covariates Z is only half as frequent

represented in the observed data as in the study population. Thus, in this case observed subjects with covariates Z are weighted $1/0.5 = 2$ to account for both themselves and for missing subjects with the same set of risk factors.

In this thesis we have used the IPCW method to estimate performance measures of prediction models for prediction horizon t in right-censored survival data [86]. See chapter 3 for examples. In what follows it assumed that event time and censoring time is conditional independence given the covariates, i.e. assumption (A1). The problem with right-censored event times can be summarized as follows. For person i the data is the triplet (T_i, Δ_i, Z_i) with $T_i = \min(\tilde{T}_i, C_i)$ and $\Delta_i = \mathcal{I}_{\{\tilde{T}_i \leq C_i\}}$. At prediction horizon t the contribution of this person to the estimate of the performance measure falls into one of the following three categories

- If $T_i \leq t$ and $\Delta_i = 1$ then person i has observed event prior to time t, and he contributes to the estimate as a case, cf. the definition in chapter 3.
- If $T_i > t$ then person i is known to have event after time t, and he will contribute to the estimate of the performance measure as a control, cf. the definition in chapter 3.
- If $T_i \leq t$ and $\Delta_i = 0$ then person i is censored before time t, and it is not known whether the event will occur before or after time t.

Ignoring observations from the last category in estimation of a performance measure will induce a loss of information and it will quite often result in a biased estimate [7, 40]. The method of IPCW accommodate this by weighting each subject that has "complete data" with the inverse of the probability that this subject is uncensored at time t. Subjects with complete data are in this case subjects that belongs to one of the two first categorizes above. Denote by G the conditional survival function of the censoring time C. If subject i has event prior to time t then the probability of being uncensored at time t is $G(\tilde{T}_i - |Z_i)$, and if subject i has event after time t then the probability is $G(t|Z_i)$ [5, 36]. This means that subjects that are censored before time t will only contribute to the IPCW estimate through an estimate of G.

The IPCW method depends on a working model $\mathcal{G}$ for G. Let $\hat{G}_m$ be an estimate that converges to some element $G^* \in \mathcal{G}$. If the working model is correctly specified, i.e. if $G \in \mathcal{G}$, then $\hat{G}_m \xrightarrow{m\to\infty} G^* = G$. As mentioned in chapter 3, if $\hat{G}_m$ is a consistent estimator of G then a number of studies have shown that the IPCW estimation method provides consistent estimators of various performance measures [6, 7, 34, 36, 51, 74, 85, 94]. For example, if it can be assumed that $G(t|Z) = G(t)$, i.e. the censoring time C is independent of the covariates Z, then the Kaplan-Meier estimator will provide a consistent estimate of G. However, this choice is not efficient because information from

covariates for subjects that are censored before time t will not contribute to the IPCW estimate [71]. In general it is recommended to use an estimator that ignores any prior knowledge of G such as a non- or semi-parametric regression model that include all covariates effecting the event time distribution [71, 86].

2 Functional delta-method

The functional delta-method is a generalization of the classic delta-method for estimators in $\mathbb{R}^d$ to functions defined on a normed vector space [38, 87]. In paper II and paper III we used this method to study asymptotic properties of statistical functionals. In this section an overview of the methodology behind the functional delta-method is presented. In this section we let $X_1, \ldots, X_n$ be an i.i.d. random sample on the real line from a distribution function F, and denote by F_n the empirical distribution function

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{I}_{\{X_i \leq x\}}.$$

A statistical functional is a statistic that can be written in the form $\varphi(F_n)$ for some functional φ. As shall be explain, if the functional φ is sufficiently smooth then asymptotic properties of $\varphi(F_n) - \varphi(F)$ can be obtained by combining the functional delta-method with general limit results from the theory of empirical processes. A comprehensive review of limit results for empirical processes of i.i.d. observations can be found in [88]. In what follows it is assumed that the functional φ is defined on a the Skorohod space $D[-\infty, \infty]$ of cadlag functions on the extended real line equipped with the uniform norm $\|z\|_\infty = \sup_{x \in [-\infty, \infty]} |z(x)|$ [87, 88].

We start with two well-known results from empirical processes theory regarding the asymptotic behavior of $F_n - F$ [88]

Theorem 1 (Clivenko-Cantelli)
If $X_1, X_2, \ldots$ are iid random variables with distribution function F, then

$$\|F_n - F\|_\infty \xrightarrow{a.s} 0$$

Theorem 2 (Donsker)
If $X_1, X_2, \cdots$ are iid random variables with distribution function F, then the sequence of empirical processes $\mathbb{G}_n = \sqrt{n}(F_n - F)$ converges in distribution in $D[-\infty, \infty]$ to a tight Gaussion process $\mathbb{G}_F$.

The basic idea of the functional delta-method is translate the above results for $F_n - F$ into asymptotic results for $\varphi(F_n) - \varphi(F)$ through a differential analysis

of φ. The idea of studying the asymptotic behavior of $\varphi(F_n) - \varphi(F)$ through a differential analyses of φ was initiated by von Mises [91] who proposed an expansion of statistical functionals similar to the Taylor expansion of smooth functions. Heuristically, we would like to obtain an expansion of the following form

$$\begin{aligned}\varphi(F_n) - \varphi(F) &= \frac{1}{\sqrt{n}}\varphi'_F(\mathbb{G}_n) + o_P(1) \\ &= \frac{1}{n}\sum_{i=1}^{n}\varphi'_F(\delta_{X_i} - F) + o_P(1), \end{aligned} \tag{4.1}$$

for some linear functional φ'_F. The existence of such an expansion depends on the differentiability properties of the functional φ.

The most basic type of differentiability for functions defined on normed vector spaces is the type of *Gâteaux differentiability* [38, 87]. Let $\mathcal{F}$ and $\mathcal{E}$ be normed vector spaces, a function $\psi : \mathcal{F} \to \mathcal{E}$ is said to be Gâteaux differentiable at $\theta \in \mathcal{F}$ if

$$\left\| \frac{\psi(\theta + \epsilon h) - \psi(\theta)}{\epsilon} - \psi'_\theta(h) \right\|_{\mathcal{E}} \to 0, \quad \text{as } \epsilon \to 0$$

for some fixed $h \in \mathcal{F}$. However, because h is not allowed to depend on ϵ, this form of differentiability is to weak for most statistical applications. Intuitively, for asymptotic statistics it is useful to "allow the direction h to depend on n". A stronger form of differentiability is the type of *Hadamard differentiability* [38, 87]. Suppose the function ψ is defined on $\mathcal{F}_\psi \subset \mathcal{F}$. The function ψ is said to be Hadamard differentiable at $\theta \in \mathcal{F}$ if there exist a continuous, linear map $\psi'_\theta : \mathcal{F} \to \mathcal{E}$ such that

$$\left\| \frac{\psi(\theta + \epsilon h_\epsilon) - \psi(\theta)}{\epsilon} - \psi'_\theta(h) \right\|_{\mathcal{E}} \to 0, \quad \text{as } \epsilon \to 0, \text{ every } h_\epsilon \to h.$$

Here the direction h_ϵ is allowed to change with ϵ and thus Hadamard differentiablity is more suited for statistical applications. By the following theorem, an expansion as the one in (4.1) can be obtained if the statistical functional φ is Hadamard differentiable. The theorem and its proof can be found in e.g. [38, Theorem 3] or [87, Theorem 20.8].

Theorem 3 (Functional delta-method)
Let $\mathcal{F}$ and $\mathcal{E}$ be normed linear spaces. Let $\psi : \mathcal{F}_\psi \subset \mathcal{F} \to \mathcal{E}$ be Hadamard differentiable at θ. Let $T_n : \Sigma_n \to \mathcal{F}_\psi$ be maps such that $r_n(T_n - \theta) \rightsquigarrow T$ for some sequence of numbers $r_n \to \infty$ and a random element T. Then $r_n(\psi(T_n) - \psi(\theta)) \rightsquigarrow \psi'(T)$. If ψ' is defined and continuous on the whole space $\mathcal{F}$, then we also have that $r_n(\psi(T_n) - \psi(\theta)) = \psi'(r_n(T_n - \theta)) + o_P(1)$.

Chapter 5

Summary of papers

1 Paper I

Rikke Nørmark Mortensen, Thomas Alexander Gerds, Jørgen Lykke Jeppesen, and Christian Torp-Pedersen (2017). Office blood pressure or ambulatory blood pressure for the prediction of cardiovascular events. *European Heart Journal* 38(44): pp 3296-3304.

Paper I is a study on a large data set regarding the predictive value of various blood pressure measurements on the risk of future cardiovascular events. The main aim of this paper was to use statistical risk prediction models to study the relative prognostic value of office blood pressure measurements and ambulatory blood pressure measurements on the 10-year risk of fatal and non-fatal cardiovascular events. We further considered the relative prognostic value of daytime blood pressure measurements and nighttime blood pressure measurements on the 10-year risk of cardiovascular events. For this task we used data from the International Database on Ambulatory blood pressure monitoring in relation to Cardiovascular Outcomes.

Several studies have previously considered the relative importance of office blood pressure and ambulatory blood pressure on the risk of cardiovascular events [15, 16], and the relative importance of daytime blood pressure and nighttime blood pressure [9, 43]. These studies all claim that ambulatory blood pressure measurements provide prognostic value beyond that of office blood pressure, and similar nighttime blood pressure provides prognostic value beyond that of daytime blood pressure. Thus, the general recommendations from these studies are that it is import to measure the ambulatory blood pressure, in particular ambulatory nighttime blood pressure, in order to provide an accurate prognosis for a patient. These recommendations have also been included in the current guidelines regarding cardiovascular disease

prevention [68]. The motivation for our study was that these previous studies all used Cox regression analyses and conclusions are drawn solely based on hazard ratios and corresponding p-values. No attempt has been made to assess the accuracy of person-specific risk estimates, and thus the studies do not formally consider prediction. This demonstrate a point that is not widely appreciated in applied medical settings, namely that a strong statistical association does not necessary translate into strong predictive power, and thus *explanatory power* is often confused with *predictive power*.

In our study we used cause-specific Cox regression models to obtain 10-year risk estimates of cardiovascular events, and the predictive accuracy of these risk estimates were assessed by the time-dependent AUC for competing risks. The conclusions from the study was that ambulatory blood pressure measurements did not provide predictive value beyond that of office blood pressure measurements, and nighttime blood pressure did not provide predictive value beyond that of daytime blood pressure.

2 Paper II

Rikke Nørmark Mortensen, Tianxi Cai, and Thomas Alexander Gerds. Uncertainty of bootstrap cross-validation estimates of prediction performance in censored survival data. *(Manuscript)*

In paper II we considered extensions of a leave-one-out bootstrap estimator suggested by Efron and Tibshirani [29]. This estimator was originally introduced to estimate the error rate in a binary outcome setting. In our study we suggested a leave-one-out bootstrap IPCW estimator of the time-dependent Brier score and a leave-pair-out bootstrap IPCW estimator of the time-dependent AUC for right-censored survival data. To deal with censored event times we used the inverse probability of censoring weighting technique.

We used von Mises calculus to obtain large sample properties for each of the two estimators and derived the influence functions. From estimates of these influences functions one can get estimates of the standard error of the two bootstrap estimates and confidence intervals can be constructed.

The leave-one-out bootstrap IPCW estimator and the leave-pair-out bootstrap IPCW estimator are implemented in the `Score` function from the R-package `riskRegression` [37, 83] that is public available at the Comprehensive R Archive Network (CRAN) site [69].

3 Paper III

Rikke Nørmark Mortensen, Michael Mørk Petersen, Michala Skovlund Sørensen and Thomas Alexander Gerds. On risk predictions from the logistic re-

calibration method. *(Manuscript)*

In paper III we review a logistic re-calibration approach that provide updated risk estimates when an established prediction model is applied in a data set different from the one in which the model was developed [19, 62, 79, 90]. The case considered in this paper were limited to a logistic prediction model.

The idea is that if the established model provides invalid risk estimates in the new data then the calibration of the model can be improved by updating the prediction model as follows. First risk predictions obtained from the established prediction model is used to fit a logistic calibration model in the new data. This step yields estimates of the regression coefficients a and b of the calibration model. In a second step, the risk predictions are updated according to the estimates $\hat{a}$ and $\hat{b}$. The resulting re-calibrated risk estimates will have better calibration properties in the new data compared to the established model.

Because of this two-step updating approach, the re-calibrated risk estimates will inherent variability from both the prediction model and the calibration model, and for this reason it is not straight forward to construct confidence intervals for the re-calibrated risk estimates. In the paper we suggested a method to construct confidence intervals that incorporate variability from both the prediction model and the calibration model; for this task we used von Mises calculus. In a small simulation study we demonstrated that the coverage of the confidence intervals were close to the nominal level of 95%.

Chapter 6

Discussion

The aim of most medical research is to identify patients with an increase risk of some fatal or non-fatal event. In this thesis it was demonstrated how statistical risk prediction models can be used to pursue this aim. Specially, the results in paper I demonstrated that risk factors that are highly significant associated to an event of interest may not provide predictive value that is meaningful in a clinical application. It was demonstrated how statistical risk prediction models provide a better tool to identify biomarkers that actual improves the clinicians ability to predict the event for a patient. In paper II and paper III we suggested statistical tools that can be used to assess the predictive value of such prediction models. We now discuss some further details of the materials presented in the three papers of the thesis.

1 Added value of a new marker

In paper I we used statistical risk prediction models to study the predictive value of various blood pressure measurements on the risk of fatal and nonfatal cardiovascular events. In clinical applications, it is well established that a raised blood pressure is associated with an increased risk of cardiovascular events. Thus, a patient with a raised blood pressure may benefit from treatment with antihypertensive drugs. As pointed out in chapter 5 it is generally accepted that ambulatory blood pressure, measured over a period of 24-hours, is the strongest marker for prediction of cardiovascular events [15, 16, 68]. The results of our study as presented in Paper I seemingly contradict this contention. The approach taken in our study goes beyond the often met, but too simple, approach which simply considers the p-value associated with a predictor variable in Cox regression. However, if the biomarker is intended to aid clinical decision-making, this etiological approach does not address the question of direct interest, namely whether the biomarker im-

proves our ability to predict the patients future outcome [4, 17, 54, 55, 67]. We assess the accuracy of the statistical prediction models obtained by including versus excluding the predictor variable in question. The general idea behind our approach is not new but it is rarely meet in medical applications [21, 54, 63]. In our Paper I, we deal with competing risks and right censored data and we investigate the magnitude of the changes in patient individual risk predictions that occur when a single new biomarker is added to a standard prediction model.

2 Internal validation in paper I

As discussed in chapter 3 it is usually desirable to use cross-validation to detect overfitting of a prediction model. In paper I we did not pursue this aspect of model validation, and some comments regarding this will now be made. A prediction model is said to be overfitted if the model specification is too complex to capture the true underlying data structure. This will for example occur if the model contains more variables than the data can justify. In this case the trained model will contain "noise", i.e. trends inherent in the training data that does not generalize to other data sets. The intend of internal validation by cross-validation is to detect this "noise". In the case of the study in paper I we compared the predictive value of two nested models; a model containing office blood pressure and a model containing both office blood pressure and ambulatory blood pressure (both of these models were adjusted for the same set of risk factors). The results revealed that the AUC of the two models were nearly identical, and we concluded that ambulatory blood pressure did not add predictive value beyond that of office blood pressure. As we did not attempt to detect overfitting by internal validation we do not know whether the two models are overfitted; they might indeed be. However, the bigger model did not contain predictive information beyond that of the smaller model. This means that the possible "noise" in the bigger model is not any worse than the "noise" in the smaller model, and thus it is unlikely that cross-validation would change the conclusion of the study.

3 Internal validation of prediction performance

In paper II we discussed a leave-one-out bootstrap IPCW estimator of the time-dependent Brier score, and a leave-pair-out bootstrap IPCW estimator of the time-dependent AUC for right-censored survival data. In the paper we used von Mises calculus to study large sample properties, and we derived the influence function. An estimate of the standard error can be obtained from the estimate of the influence function. An attractive feature of the estimate

of the influence function is that it only depends on the same set of bootstrap samples that were used to calculate the point estimate, this means that the standard error can be estimated in a computationally efficient manner. We used the results to construct confidence intervals for the target parameters, i.e., for the internally validated prediction performance of a model and for differences between models. As explained in more detail in the paper, the target parameters estimated by the leave-one-out bootstrap estimator and the leave-pair-out bootstrap estimator are the expected Brier score and the expected AUC, respectively. These expected performance measures can be interpreted when the aim is to assess the performance of the prediction modelling strategy rather than the performance of a prediction model which was trained in a given dataset. However, the latter will often be the parameter of actual interest for the applied researcher. For example, the expected Brier score may be of little interest if the leave-one-out bootstrap estimator is used in an internal validation analysis to detect overfitting of a prediction model, and it may seem irrelevant to construct a confidence interval for this parameter. However, if the aim of an analysis is to choose between two rival prediction models it may be highly relevant to assess the standard error of the leave-one-out estimator in order to construct statistical tests to compare the performance of the two prediction models.

In the paper we use the IPCW estimation approach to deal with right-censored event times. This approach has in the past decade become a popular method to estimate performance measures for risk prediction models in censored survival data [6, 7, 34, 36, 51, 74, 85, 94]. The IPCW method is in general recommended over other model-based estimation methods that depend on a correctly specified model of the event time distribution, especially if the aim is to compare two competing prediction models [7, 33, 36]. The main reason for choosing an IPCW estimator over a model-based estimate [56] is that a model-based estimate of the performance measures will be biased as soon as the model of the event time distribution is wrong. Also the bias will increase with increasing misspecification of the model. Comparing two statistical prediction models is, by its very nature, a comparison of two rival models for the event time distribution, and it follows that a comparison based on a model-based estimator cannot be interpreted solely in terms of the accuracy of the two models. The IPCW estimation approach does however depend on a model of the survival function of the censoring distribution, and the IPCW estimate will be biased if the model for the censoring distribution is wrong. From a modeling perspective it may or may not be more difficult to model the event time distribution than to model the censoring distribution. However, for the aim of comparing two prediction models the IPCW estimate appears much more tractable compared to a model-based estimate. The reason for this is that even if the model for the censoring distribution is misspecified, then at least the bias in the estimated performance measures

will be the same for the two rival prediction models.

4 Re-calibration of risk predictions

In paper III we considered the logistic re-calibration method to update a logistic risk prediction model which has been proven not to provide valid risk estimates in a data setting that is different from the one in which the model was developed. The logistic re-calibration method is one among many methods that can be used to update prediction models for use in a new dataset [18, 53, 79, 82, 89]. It is relevant to discuss why we need to update a model with poor calibration performance in the situation where we have a new dataset that could in principle be used to refit the model formula. In applied settings it is most common to refit a new model rather than to use an updating method. In certain applied areas this practice has led to a large number of different prediction models that all predict the same outcome, and it may be hard to navigate between all these models when the aim is to choose the model that is best suited for a given setting [22, 61]. For this reason several studies have advocated the use of an updating method to re-calibrate an existing model rather than refitting new models [52, 61, 82, 90]. The main argument is that by refitting a model one is potentially loosing a lot of information regarding the choice of risk factors and their effect on outcome. Another argument is that the data in which the established model was developed is usually much larger than the data that is available from the new data setting, and thus the refitted model is often less generalizable compared to an updated model that contains information from both data sets.

The logistic re-calibration method is a quite parsimonious updating method as it only updates by adding and multiplying the intercept and slope from the calibration model. Two important limitations are related to this updating method. First, the method assumes that the difference in effect of the risk factors on the outcome between the two data sets can be explained by the same multiplying factor for all risk factors. This assumption simplifies the problem considerably and one may wonder if it holds, for example, when the age-gender distributions differ a lot between the two datasets. Secondly, the logistic re-calibration method does not allow for incorporation of new biomarkers. This may for example be problematic if the model updated in calendar time as new important biomarkers may have emerged. Other more flexible updating methods accommodate these limitations by allowing for e.g. more flexible updating of the effect of the risk factors on the outcome, or by allowing incorporation of new biomarkers. However, as these more flexible updating methods are much more complex it may happen that the resulting updated prediction models will require another validation analysis.

Risk prediction from an updated model will inherent variability from both the prediction model and the updated prediction model, and it is likely that the variability of the risk estimates will increase with the flexibility of the updating method. Thus, if it is necessary to use a highly complicated updating method to adapt the prediction model to the new data setting it may be more sensible to simply refit a model to the new data.

Chapter 7

Perspectives

1 Extension of the leave-out-out bootstrap estimator

In paper II we defined leave-one-out bootstrap estimates of the expected Brier score and the expected AUC in right-censored data. In future studies it may be interesting to pursue this further by considering extension to other performance measures. It is straight forward to extend the estimates to time-dependent Brier score and time-dependent AUC for competing risks. However, other extension may be considered such as other scoring rules and the time-dependent c-index.

2 Re-calibration for survival data and competing risks

The re-calibration method considered in paper III is limited to a logistic prediction model and a logistic calibration model. However, as censored survival data and competing risks are common in medical applications it is relevant to discuss extensions to such settings. In survival analysis without competing risks the most popular model is the Cox proportional hazards model. Using the same notation as in chapter 3, the Cox model can be written as

$$S(t|Z) = S_0(t)\exp(\beta Z),$$

where $S_0(t)$ is a baseline survival function and β is a vector of regression coefficients. Suppose the Cox model has been fitted in the data $\mathcal{L}_n$ which yield a partial maximum likelihood estimate $\hat{\beta}$ of the regression coefficients and the usual Breslow estimate of the baseline hazard function. Suppose the aim

is to re-calibrate the resulting prediction model to adjust to the data $\mathcal{V}_m$. Fix the prediction horizon $t > 0$ and let $\hat{\beta}Z_j$ be the linear predictor for person j in $\mathcal{V}_m$. For this setting, one can obtain a re-calibration method that is analogue to the logistic re-calibration method by fitting a Cox regression model in $\mathcal{V}_m$ with the linear predictor $\hat{\beta}Z_j$ as the only covariate this will yield a calibration slope $\hat{b}$ [78, p. 381] [90]. Then, for person j in $\mathcal{V}_m$ a re-calibrated risk estimate can be calculated by multiplying the linear $\hat{\beta}Z_j$ with $\hat{b}$ and updating the Breslow estimator at time t according to this re-calibrated linear predictor. Other related re-calibration methods have likewise been suggested to re-calibrate a Cox prediction model in right-censored survival data [20, 90].

However, none of these methods seem to generalize to the case of competing risks. A heuristic method to re-calibrate risk estimates for competing risk is now discussed. Prediction models for competing risks settings aim at predicting the cumulative incidence function $F_1(t|Z) = \mathrm{P}_n(\tilde{T} \leq t, \eta = 1|Z)$; here it is assumed that the event of interest is event $\eta = 1$. The cumulative incidence function may be estimates by fitting a cause-specific Cox regression model for event $\eta = 1$ and a cause-specific Cox regression model for event $\eta = 2$. Then one can get an estimate of $F_1(t|Z)$ by using the formula

$$F_1(t|Z) = \int_0^t h_1(s|Z) \exp\Big(- \int_0^s h_1(u|Z) + h_2(u|Z) \mathrm{d}u \Big) \mathrm{d}s,$$

here $h_1(s|Z)$ is the cause-specific hazard function for event $\eta = 1$ and $h_2(s|Z)$ is the cause-specific hazard function for event $\eta = 2$. As this prediction model relies on Cox regression models it may be possible to use the re-calibration method for Cox models to re-calibrate estimates of $F_1(t|Z)$. However, this approach requires re-calibration of two Cox models and it is not clear whether this will result in valid re-calibrated risk estimates. A more sensible approach is to use a calibration model that regress the risk estimate $R^1_{t,n}(Z_j)$ directly to $F_1(t|R^1_{t,n}(Z_j))$. One solution is to use an absolute risk regression model [35]

$$\log\big(F_1(t|R^1_{t,n}(Z_j))\big) = a(t) + b \cdot \log(R^1_{t,n}(Z_j)),$$

with $a(t) = \log(F_0(t))$. Parameters of this model can be estimated using either estimation based on IPCW [73] or pseudo-values for survival data [3]. A re-calibrated risk estimate is then given by

$$\hat{F}_1(t|R^1_{t,n}(Z_j)) = \hat{F}_0(t) \exp\big(\hat{b} \cdot \log(R^1_{t,n}(Z_j))\big),$$

Future work could examine if this re-calibration approach provides risk-estimates that have better calibration properties compared to the estimate $R^1_{t,n}(Z_j)$.

References

[1] Per Kragh Andersen, Steen Z Abildstrom, and Susanne Rosthøj. Competing risks as a multi-state model. *Statistical Methods in Medical Research*, 11(2):203–215, 2002.

[2] Per Kragh Andersen, Ornulf Borgan, Richard D. Gill, and Niels Keiding. *Statistical models based on counting processes*. Springer, 1993.

[3] Per Kragh Andersen and Maja Pohar Perme. Pseudo-observations in survival analysis. *Statistical Methods in Medical Research*, 19(1):71–99, 2010.

[4] Aasthaa Bansal and Margaret Sullivan Pepe. When does combining markers improve classification performance and what are implications for practice? *Statistics in Medicine*, 32(11):1877–1892, 2013.

[5] Peter J. Bickel, Chris A. Klaassen, Ya'acov Ritov, and Jon A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins series in the mathematical sciences. Johns Hopkins University Press, 1993.

[6] Paul Blanche, Jean-François Dartigues, and Hélène Jacqmin-Gadda. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Statistics in Medicine*, 32(30):5381–5397, 12 2013.

[7] Paul Blanche, Jean-François Dartigues, and Hélène Jacqmin-Gadda. Review and comparison of ROC curve estimators for a time-dependent outcome with marker-dependent censoring. *Biometrical Journal*, 55(5):687–704, 2013.

[8] Paul Blanche, Michael W Kattan, and Thomas A Gerds. The c-index is not proper for the evaluation of t-year predicted risks. *Biostatistics*, page kxy006, 2018.

[9] José Boggia, Yan Li, Lutgarde Thijs, Tine W Hansen, Masahiro Kikuya, Kristina Björklund-Bodegård, Tom Richart, Takayoshi Ohkubo, Tatiana Kuznetsova, Christian Torp-Pedersen, Lars Lind, Hans Ibsen, Yutaka Imai, Jiguang Wang, Edgardo Sandoya, Eoin O'Brien, and Jan A Staessen. Prognostic accuracy of day versus night ambulatory blood pressure: a cohort study. *The Lancet*, 370(9594):1219–1229, 2007.

[10] Ulisses M. Braga-Neto and Edward R. Dougherty. Is cross-validation valid for small-sample microarray classification? *Bioinformatics*, 20(3):374–380, 2004.

[11] Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, 2001.

[12] Glenn W. Brier. Verication of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

[13] Hege M. Bøvelstad and Ørnulf Borgan. Assessment of evaluation criteria for survival prediction from genomic data. *Biometrical Journal*, 53(2):202–216, 2011.

[14] Lloyd E. Chambless and Guoqing Diao. Estimation of time-dependent area under the roc curve for long-term risk prediction. *Statistics in Medicine*, 25(20):3474–3486, 2006.

[15] Denis L. Clement, Marc L. De Buyzere, Dirk A. De Bacquer, Peter W. de Leeuw, Daniel A. Duprez, Robert H. Fagard, Peter J. Gheeraert, Luc H. Missault, Jacob J. Braun, Roland O. Six, Patricia Van Der Niepen, and Eoin O'Brien. Prognostic value of ambulatory blood-pressure recordings in patients with treated hypertension. *New England Journal of Medicine*, 348(24):2407–2415, 2003.

[16] David Conen and Fabian Bamberg. Noninvasive 24-h ambulatory blood pressure and cardiovascular disease: a systematic review and meta-analysis. *Journal of Hypertension*, 26(7):1290–1299, 2008.

[17] Nancy R. Cook. Quantifying the added value of new biomarkers: how and how not. *Diagnostic and Prognostic Research*, 2(1):14, 2018.

[18] J. B. Copas. Regression, prediction and shrinkage. *Journal of the Royal Statistical Society. Series B (Methodological)*, 45(3):311–354, 1983.

[19] D. R. Cox. Two further applications of a model for binary regression. *Biometrika*, 45(3/4):562–565, 1958.

[20] Cynthia S. Crowson, Elizabeth J. Atkinson, and Terry M. Therneau. Assessing calibration of prognostic risk scores. *Statistical Methods in Medical Research*, 25(4):1692–1706, 2016.

[21] Marc L. De Buyzere. Multi-biomarker risk stratification in heart failure: a story of diminished marginal returns after herculean efforts? *European Journal of Heart Failure*, 20(2):278–280, 2018.

[22] Thomas P.A. Debray, Hendrik Koffijberg, Daan Nieboer, Yvonne Vergouwe, Ewout W. Steyerberg, and Karel G.M. Moons. Meta-analysis and aggregation of multiple published prediction models. *Statistics in Medicine*, 33(14):2341–2362, 2014.

[23] Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 01 1979.

[24] Bradley Efron. Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association*, 78(382):316–331, 06 1983.

[25] Bradley Efron. How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association*, 81(394):461–470, 06 1986.

[26] Bradley Efron. *The Jackknife, the bootstrap, and other resampling plans*. SIAM, Philadelphia, 1990.

[27] Bradley Efron and Robert Tibshirani. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1(1):54–75, 02 1986.

[28] Bradley Efron and Robert Tibshirani. Cross-validation and the bootstrap: estimating the error rate of a prediction rule. Technical Report 176, Division of biostatistics, Stanford University, 05 1995.

[29] Bradley Efron and Robert Tibshirani. Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association*, 92(438):548–560, 06 1997.

[30] Jason P. Fine and Robert J. Gray. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94(446):496–509, 1999.

[31] Seymour Geisser. The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70(350):320–328, 1975.

[32] Thomas A. Gerds, Per K. Andersen, and Michael W. Kattan. Calibration plots for risk prediction models in the presence of competing risks. *Statistics in Medicine*, 33(18):3191–3203, 2013.

[33] Thomas A. Gerds, Tianxi Cai, and Martin Schumacher. The performance of risk prediction models. *Biometrical Journal*, 50(4):457–479, 2008.

[34] Thomas A. Gerds, Michael W. Kattan, Martin Schumacher, and Changhong Yu. Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Statistics in Medicine*, 32(13):2173–2184, 2012.

[35] Thomas A. Gerds, Thomas H. Scheike, and Per K. Andersen. Absolute risk regression for competing risks: interpretation, link functions, and prediction. *Statistics in Medicine*, 31(29):3921–3930, 2012.

[36] Thomas A. Gerds and Martin Schumacher. Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal*, 48(6):1029–1040, 12 2006.

[37] Thomas Alexander Gerds and Brice Ozenne. *riskRegression: Risk Regression Models and Prediction Scores for Survival Analysis with Competing Risks*. R package version 2018.07.06.

[38] Richard D. Gill. Non- and semi-parametric maximum likelihood estimators and the von mises method (part 1). *Scandinavian Journal of Statistics*, 16:97–128, 1989.

[39] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[40] Erika Graf, Claudia Schmoor, Willi Sauerbrei, and Martin Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(18):2529–2545, 1999.

[41] Erika Graf and Martin Schumacher. An investigation on measures of explained variation in survival analysis. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 44(4):497–507, 1995.

[42] James A. Hanley and Barbara J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

[43] Tine W. Hansen, Yan Li, José Boggia, Lutgarde Thijs, Tom Richart, and Jan A. Staessen. Predictive role of the nighttime blood pressure. *Hypertension*, 57(1):3–10, 2011.

[44] Frank E. Harrell, Jr. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer series in statistics. Springer, 2. ed. edition, 2015.

[45] Frank E. Harrell, Jr., Robert M. Califf, David B. Pryor, Kerry L. Lee, and Robert A. Rosati. Evaluating the yield of medical tests. *Journal of the American Medical Association*, 247(18):2543–2546, 1982.

[46] Patrick J. Heagerty, Thomas Lumley, and Margaret S. Pepe. Time-dependent roc curves for censored survival data and a diagnostic marker. *Biometrics*, 56(2):337–344, 2000.

[47] Patrick J. Heagerty and Yingye Zheng. Survival model predictive accuracy and roc curves. *Biometrics*, 61(1):92–105, 2005.

[48] Jørgen Hilden and Thomas Gerds. A note on the evaluation of novel biomarkers: do not rely on integrated discrimination improvement and net reclassification index. *Statistics in medicine*, 33(19):3405–3414, 2014.

[49] Jørgen Hilden, J. Habbema, and Beth 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.

[50] David W. Hosmer and Stanley Lemeshow. *Applied logistic regression*. Wiley series in probability and statistics. Wiley, 2. ed. edition, 2000.

[51] Hung Hung and Chin-Tsang Chiang. Estimation methods for time-dependent auc models with survival data. *Canadian Journal of Statistics*, 38(1):8–26, 2010.

[52] Joan Ivanov, Jack Ven Tu, and C. David Naylor. Ready-made, recalibrated, or remodeled? issues in the use of risk indexes for assessing mortality after coronary artery bypass graft surgery. *Circulation*, 99 16:2098–2104, 1999.

[53] K.J.M. Janssen, K.G.M. Moons, C.J. Kalkman, D.E. Grobbee, and Y. Vergouwe. Updating methods improved the performance of a clinical prediction model in new patients. *Journal of Clinical Epidemiology*, 61(1):76–86, 2008.

[54] Michael Kattan. Statistical prediction models, artificial neural networks, and the sophism "I am a patient, not a statistic". *Journal of Clinical Oncology*, 20(4):885–887, 2002.

[55] Michael W. Kattan. Judging new markers by their ability to improve predictive accuracy. *Journal of the National Cancer Institute*, 95(9):634–635, 2003.

[56] Edward L. Korn and Richard Simon. Measures of explained variation for survival data. *Statistics in Medicine*, 9(5):487–503, 1990.

[57] Jerald F. Lawless and Yan Yuan. Estimation of prediction error for survival models. *Statistics in Medicine*, 29(2):262–274, 2009.

[58] S. le Cessie and J. C. van Houwelingen. A goodness-of-fit test for binary regression models, based on smoothing methods. *Biometrics*, 47(4):1267–1282, 1991.

[59] Kristian Linnet. Assessing diagnostic tests by a strictly proper scoring rule. *Statistics in Medicine*, 8(5):609–618, 1989.

[60] Gregory Y. H. Lip, Robby Nieuwlaat, Ron Pisters, Deirdre A. Lane, and Harry J. G. M. Crijns. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: The euro heart survey on atrial fibrillation. *Chest*, 137(2):137, 2010.

[61] Glen P. Martin, Mamas A. Mamas, Niels Peek, Iain Buchan, and Matthew Sperrin1. Clinical prediction in defined populations: a simulation study investigating when and how to aggregate existing models. *BMC Medical Research Methodology*, 17(1), 2017.

[62] Michael E Miller, Siu L Hui, and William M Tierney. Validation techniques for logistic regression models. *Statistics in Medicine*, 10(8):1213–1226, 1991.

[63] Tuan V Nguyen and John A Eisman. Assessment of fracture risk: Population association versus individual prediction. *Journal of Bone and Mineral Research*, 33(3):386–388, 2018.

[64] Michael J. Pencina, Ralph B. D'Agostino Sr, Ralph B. D'Agostino Jr, and Ramachandran S. Vasan. Evaluating the added predictive ability of a new marker: From area under the roc curve to reclassification and beyond. *Statistics in Medicine*, 27(2):157–172, 2007.

[65] Margaret Pepe and Holly Janes. Methods for evaluating prediction performance of biomarkers and tests. In *Risk assessment and evaluation of predictions*, pages 107–142. Springer, 2013.

[66] Margaret Sullivan Pepe. *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press, Oxford, 2003.

[67] Margaret Sullivan Pepe, Holly Janes, Gary Longton, Wendy Leisenring, and Polly Newcomb. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *American Journal of Epidemiology*, 159(9):882–890, 2004.

[68] Massimo F Piepoli, Arno W Hoes, Stefan Agewall, Christian Albus, Carlos Brotons, Alberico L Catapano, Marie-Therese Cooney, Ugo Corrà, Bernard Cosyns, Christi Deaton, Ian Graham, Michael Stephen Hall, F D Richard Hobbs, Maja-Lisa Løchen, Herbert Löllgen, Pedro Marques-Vidal, Joep Perk, Eva Prescott, Josep Redon, Dimitrios J Richter, Naveed Sattar, Yvo Smulders, Monica Tiberi, H Bart van der Worp, Ineke van Dis, W M Monique Verschuren, Simone Binno, and ESC Scientific Document Group. 2016 european guidelines on cardiovascular disease prevention in clinical practicethe sixth joint task force of the european society of cardiology and other societies on cardiovascular disease prevention in

clinical practice (constituted by representatives of 10 societies and by invited experts)developed with the special contribution of the european association for cardiovascular prevention; rehabilitation (eacpr). *European Heart Journal*, 37(29):2315–2381, 2016.

[69] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.

[70] M. Shafiqur Rahman, Gareth Ambler, Babak Choodari-Oskooei, and Rumana Z. Omar. Review and evaluation of performance measures for survival prediction models in external validation settings. *BMC medical research methodology*, 17(60), 2017.

[71] James M. Robins and Andrea Rotnitzky. *Recovery of Information and Adjustment for Dependent Censoring Using Surrogate Markers*, pages 297–331. Birkhäuser Boston, Boston, MA, 1992.

[72] James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

[73] Thomas H. Scheike, Mei-Jie Zhang, and Thomas A. Gerds. Predicting cumulative incidence probability by direct binomial regression. *Biometrika*, 95(1):205–220, 2008.

[74] Rotraut Schoop, Jan Beyersmann, Martin Schumacher, and Harald Binder. Quantifying the predictive accuracy of time-to-event models in the presence of competing risks. *Biometrical Journal*, 53(1):88–112, 2011.

[75] Galit Shmueli. To explain or to predict? *Statistical Science*, 25(3):289–310, 2010.

[76] Richard Simon. Roadmap for developing and validating therapeutically relevant genomic classifiers. *Journal of Clinical Oncology*, 23(29):7332–7341, 10 2005.

[77] Xiao Song and Xiao-Hua Zhou. A semiparametric approach for the covariate-specific roc curve with survival outcome. *Statistica Sinica*, 18(3):947–965, 2008.

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

[79] Ewout W. Steyerberg, Gerard J. J. M. Borsboom, Hans C. van Houwelingen, Marinus J. C. Eijkemans, and J. Dik F. Habbema. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Statistics in Medicine*, 23(16):2567–2586, 2004.

[80] Ewout W. Steyerberg, Andrew J. Vickers, Nancy R. Cook, Thomas Gerds, Mithat Gonen, Nancy Obuchowski, Michael J. Pencina, and Michael W. Kattan. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, 21(1):128–138, 2010.

[81] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):111–147, 1974.

[82] Ting-Li Su, Thomas Jaki, Graeme L Hickey, Iain Buchan, and Matthew Sperrin. A review of statistical updating methods for clinical prediction models. *Statistical Methods in Medical Research*, 27(1):185–197, 2016.

[83] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.

[84] Anastasios Tsiatis. *Semiparametric theory and missing data*. Springer, 2007.

[85] Hajime Uno, Tianxi Cai, Lu Tian, and L. J. Wei. Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association*, 102(478):527–537, 2007.

[86] Mark J. van der Laan and James M. Robins. *Unified Methods for Censored Longitudinal Data and Causality*. Springer, New York, 2003.

[87] A. W. van der Vaart. *Asymptotic statistics*. Cambridge : Cambridge University Press, 2007.

[88] Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics, New York, 1996.

[89] Kirsten Van Hoorde, Yvonne Vergouwe, Dirk Timmerman, Sabine Van Huffel, Ewout W. Steyerberg, and Ben Van Calster. Simple dichotomous updating methods improved the validity of polytomous prediction models. *Journal of Clinical Epidemiology*, 66(10):1158–1165, 2013.

[90] Hans C. van Houwelingen and Jane Thorogood. Construction, validation and updating of a prognostic model for kidney graft survival. *Statistics in Medicine*, 14(18):1999–2008, 1995.

[91] Richard von Mises. On the asymptotic distribution of differentiable statistical functions. *The Annals of Mathematical Statistics*, 18(3):309–348, 1947.

[92] Sonja Wehberg and Martin Schumacher. A comparison of nonparametric error rate estimation methods in classification problems. *Biometrical Journal*, 46(1):35–47, 2004.

[93] Peter W. F. Wilson, Ralph B. D'Agostino, Daniel Levy, Albert M. Belanger, Halit Silbershatz, and William B. Kannel. Prediction of coronary heart disease using risk factor categories. *Circulation*, 97(18):1837–1847, 1998.

[94] Marcel Wolbers, Paul Blanche, Michael T. Koller, Jacqueline C. M. Witteman, and Thomas A. Gerds. Concordance for prognostic models with competing risks. *Biostatistics*, 15(3):526–539, 2014.

[95] Yingye Zheng, Tianxi Cai, Yuying Jin, and Ziding Feng. Evaluating prognostic accuracy of biomarkers under competing risk. *Biometrics*, 68(2):388–396, 2012.

ISSN (online): 2246-1302
ISBN (online): 978-87-7210-335-8

AALBORG UNIVERSITY PRESS