Università degli Studi di Padova
Dipartimento di Scienze Statistiche

Corso di Laurea Magistrale in

Scienze Statistiche

Accounting for uncertainty in the predictive calibration of prognostic models constructed using multiple imputations with cross-validatory assessment

Relatore: Prof. Livio Finos
Dipartimento di Psicologia dello Sviluppo e della Socializzazione

Correlatori: Prof. Bart Mertens e Prof.ssa Liesbeth de Wreede
Medical Statistics, Department of Biomedical Data Sciences, Leiden
University Medical Center, Leiden, Netherlands

Laureanda Erika Banzato
Matricola n. 1131385

Anno Accademico 2017/2018

Contents

Introduction

In a predictive context, where the aim of a study is to calibrate a predictive model, a very important step is to assess its performance. A common technique is to use cross-validation, which consists of splitting the dataset into subsets and using in turn one of them as an independent validation set and all the others to calibrate the model. Once the predictive rules have been defined, we use them to predict the outcome on the validation set and model performance can be assessed by comparing predictions with the observed outcome. At the same time though, we may need to use multiple imputation to account for missing data in the dataset. This technique, developed by Rubin (1978), imputes missing values by generating a set of several possible values from the predictive distribution of the missing values given the observed values. This is done in order to add some uncertainty to the imputation process. Combining validation and imputation may be problematic though. Indeed, on one hand we have cross-validation, that requires outcome to be removed from the calibration set to build the predictive model. While on the other hand, we have multiple imputation that requires the outcome as integral part of the estimation of the imputation model, in order to preserve the association between predictors and outcome in the imputation (White, Royston, and Wood 2011).

A second issue concerns how to obtain final predictions. Multiple imputation procedure, indeed, replaces the missing data with multiple possible values, that means creating several complete datasets. The model calibration has to be computed in each of them and the results have to be combined

somehow. In the presence of missing values in the predictors, much of classical biostatistics data analysis practice in predictive calibration focuses on the application of the so-called Rubin's rules (Rubin 2004). Basically, all the sets of parameters derived from the separate analysis on the imputed datasets are pooled together, in order to get the overall effect estimates, and plugged into the assumed substantive model, which will be used for the prediction of new outcome. In the predictive scenario however, and from a formal probabilistic point of view, the effect estimates are nuisance parameters and predictions should be obtained from the calibrated posterior predictive density with the missing observations and effect measures integrated out (Lesaffre and Lawson 2012). Hence, if the aim of the study is to get predictions, these should be obtained by pooling together the single predictions from the calibrations on the complete datasets, instead of applying Rubin's rules to the sets of parameters.

The aim of this work is to propose methodologies to combine multiple imputation with cross-validation for the assessment of prediction rules, which can also be implemented using existing imputation software. Our approaches allow outcomes of the left-out fold to be set-aside from the calibration of the final prediction model, in order to use it for validation of the estimated prediction rules. In addiction, we develop methodology to directly calibrate the required marginal density of future predictive outcomes in the presence of missing values and compare this method with direct applications of Rubin's rules. Finally, we also compare these approaches with their corresponding naïve implementations, which imply to compute multiple imputation prior to the cross-validation procedure. Since this work primary idea came from the analysis of clinical survival data, proposed methods are then described to account for lifetime outcome.

Proposed approaches performance is then evaluated by applying these methods to real and simulated data. First of all, we introduce application in prognosis and describe two real datasets with lifetime outcomes subject to censoring. The CRT (*Cardiac Resynchronization Therapy*) dataset concerns

a study from the department of cardiology of *Leiden University Medical Center* (LUMC), while the CLL (*Chronic Lymphocytic Leukemia*) dataset has been extracted from the registry of the *European Society for Blood and Marrow Transplantation* (EBMT). Finally, we also perform a simulation study to better investigate statistical performance of the proposed approaches in different scenarios.

Chapter 1 contains a brief introduction on the statistical methods used in this work. Chapter 2 concerns a general description of the theoretical background of predictive calibration, multiple imputation and validation, with particular attention on predictions. It also describes two basic approaches to the problem of predictive calibrations and assessment when multiple imputation is used to account for the presence of missing values in predictors. In addiction, it also provides a specific description on the application of these ideas in survival analysis. Chapter 3 briefly describes the real datasets we use for the analysis and presents the simulation study and the statistical measures used to assess the performance of the proposed methodologies. Finally, chapter 4 presents the results from application of the proposed approaches on the two real datasets and simulations.

Chapter 1

Statistical methods

This chapter contains a brief introduction to the statistical methods used in this work. First of all, section 1.1 presents an introduction to survival analysis and the Cox regression model. Section 1.2 is about missing data and multiple imputation, a technique to deal with them, and finally, section 1.3 deals with cross-validation.

1.1 Survival analysis

Survival analysis is the study of time-to-event data, where the dependent variable is the waiting time until the occurrence of a well-defined event. Time can be measured in years, months, weeks or days from the beginning of the follow-up of an individual, until the event occurs. For example, individuals might be followed from birth to the diagnosis of a certain disease, or from the day of surgery to death. The waiting time until the event occurs is usually called *survival time*. The most important thing is that the starting point and the event of interest must be well defined. This means that the time origin must be the same for each individual in the study, even if it can occur in different calendar years (e.g. birth, recovery, day of surgery,...) and furthermore, the endpoint has to be appropriately specified, in order to be able to calculate the time until event occurs. One could also be interested in

the occurrence of more than one event, which could be a recurrent event or a competing risk problem. In this work we only consider the possibility of one event occurring and that this occurs with certainty.

1.1.1 Censoring

An important issue concerning survival analysis is the presence of censored data. Censoring occurs when we do not know exactly the survival time of some individuals, but we still have some information about their survival times. There are three types of censoring: *right censoring*, *left censoring* and *interval censoring*. The most common one in survival analysis is right censoring, that occurs when, at the end of the followed up time, an individual has not yet experienced the event and thus we only know the time interval in which it did not occur.
This might be due to three main reasons (Kleinbaum and Klein 2012):

- a subject does not experience the event before the end of the study;
- a subject is lost to follow-up during the study period;
- a subject withdraws from the study because of death (if death is not the event of interest) or some other reason.

On the other hand, left censoring occurs when it is known that an individual experienced the event of interest before a specific time point, but that could be any time before the censoring time. Finally, interval censoring defines a situation where it is only known that the event occurred between two different time points, without knowing the exact time.

For example, we may be interested in studying the onset of HIV in a subset of the population. People in the study are then followed until they become HIV positive and the event occurs when the test for the virus is positive. In this situation it could happen that some people die during the study without the event occurring, or that at some point the study ends and some people have not yet experienced the event. For those individuals we

therefore have a *right censoring*. On the other hand, we might also not know the exact time of exposure to the virus, because when an individual gets a positive test, the follow-up period ends, but the exposure time could be any time between the starting point (say the day of born) and that moment. In this case, we have a *left censoring*, since the true survival time, which ends at the time of exposure, is shorter than the follow-up one, which ends when the test is positive. Finally, if the test turns out to be negative the first time an individual does it and positive the second time it is done, we would have *interval censorship*, because we do not know exactly when the exposure happened, but we know that it occurred between two well defined time points. (Kleinbaum and Klein 2012)

For survival data, the most important assumption about censoring is that it should be *non-informative*. This means that the distribution of survival times provides no information about the distribution of censorship times, and vice versa (Kleinbaum and Klein 2012). Under this assumption, inference is not biased. On the other hand, *informative* censoring occurs when, for example, in a survival study after a disease diagnosis, patients are lost to follow-up because their health conditions no longer allow them to attend appointments (Kartsonaki 2016).

1.1.2 Notation

Let T be a continuous non-negative random variable representing the survival time with probability density function $f(t)$. The probability of observing an event before time t is given by the cumulative distribution function:

$$F(t) = Pr(T < t) = \int_0^t f(x)\,dx \tag{1.1}$$

While the *survival function* is given by:

$$S(t) = Pr(T \geq t) = 1 - F(t) = \int_t^\infty f(x)\,dx \tag{1.2}$$

which is the probability that the event has not occurred before time t.
The *hazard function* $h(t)$ represents the instantaneous rate of occurrence and it is the probability to observe a failure in the infinitesimal interval $[t, t+\Delta t)$. It is defined as

$$h(t) = \lim_{\Delta t \to 0} \frac{Pr(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \tag{1.3}$$

and it can also be seen as the density of events at time t, divided by the probability of surviving to that duration without experiencing the event:

$$h(t) = \frac{f(t)}{S(t)} \tag{1.4}$$

Since $-f(t)$ is the derivative of $S(t)$, the hazard function can also be defined as

$$h(t) = -\frac{d}{dt} \ln S(t) \tag{1.5}$$

Then, the survival function can be written as

$$S(t) = exp(-H(t)) \tag{1.6}$$

where $H(t)$ is the cumulative hazard function:

$$H(t) = \int_0^t h(x)\, dx \tag{1.7}$$

which can be interpreted as the total amount of risk that has been accumulated up to time t.

1.1.3 Cox proportional hazard models

The main purpose of a survival study is, usually, to measure the association between the time to event with a set of covariates. This can be done using several different models, which can be *parametric*, if the distribution of T is considered known, *non-parametric*, if no assumption about the distribu-

tion of T is made, or *semi-parametric*, if the model combines both parametric and non-parametric assumptions.

Cox model assumes that the hazard at time t for an individual with covariates $\boldsymbol{x}_i$ has the form:

$$h(t|\boldsymbol{x}_i) = h_0(t)\ exp\{\boldsymbol{x}_i'\boldsymbol{\beta}\}. \tag{1.8}$$

In this equation, $h_0(t)$ is the baseline hazard function, that describes the risk for an individual with all the covariates equal to 0, $exp\{\boldsymbol{x}_i'\boldsymbol{\beta}\}$ is the parametric component and $\boldsymbol{\beta}$ is the coefficients vector. The parametric component defines the relative risk associated with the covariates and represents a proportional increase or reduction in risk, that is the same for each value of t.

The corresponding estimates of these parameters are derived by maximizing a likelihood function. The formula for the Cox model likelihood function is actually called a "partial" likelihood function. For the Cox PH model, in fact, a full likelihood based on the outcome distribution cannot be formulated, since there is not an assumed distribution for the outcome variable. Hence, the construction of the Cox likelihood is based on the observed order of events rather than the joint distribution of events and the formula considers probabilities only for those subjects who fail and does not consider probabilities for those who are censored. In particular, the partial likelihood can be written as the product of several likelihoods, one for each failure time. Each of them represents the likelihood of failing at that specific time point, given survival up to that time (Kleinbaum and Klein 2012).

Furthermore, an important feature of this formula concerns the main assumption on which this class of models is based: the *proportional hazard assumption*. This assumption implies that the hazard ratio comparing any two specifications of predictors is constant over time, or equivalently, that the hazard for one individual is proportional to the hazard for any other individual, where the proportionality constant is independent of time. We can easily check that, once we have written the hazard ratio that compares

two different specifications for the explanatory variables (using equation 1.8), the baseline hazard function $\widehat{h}_0(t)$ cancels out of the formula and the final expression no longer involves time t (Kleinbaum and Klein 2012). The proportional hazard assumption has to be checked every time the Cox model is used.

Finally, since no assumptions are made about the nature or the shape of the baseline hazard function, the Cox model can be considered as a *semi-parametric* model. This implies that this class of models does not rely on distributional assumptions for the outcome and it is the main reason why the Cox model is widely popular (Kleinbaum and Klein 2012).

1.2 Multiple imputation

The problem of missing data occurs frequently in almost all fields of research and it has to be taken into account when data are analysed. If missing data are inadequately handled, this could lead to biased or inefficient estimates of parameters and it affects the whole analysis. Since the presence of missing values may lead to technical difficulties, an approach used to handle this problem is to delete them, if they are not too many, i.e. ignoring rows of individuals with missing data (*listwise deletion*), or, otherwise, not consider in the analysis covariates with too many incomplete records. Usually, the problem is also downplayed by authors and presence of missing data and the use of listwise deletion are not even explicitly mentioned in the text, or sometimes it also happens that different tables are based on different sample sizes (Van Buuren 2012).

This section provides a brief introduction to the kinds of missing data mechanisms, along with an explanation of the *multiple imputation* procedure to deal with this problem, both in a univariate and multivariate context.

1.2.1 Missing data mechanisms

Defining the missing data mechanism is a key point in the analysis, since the properties of methods used to deal with them depend very strongly on the nature of the dependence on these mechanisms. This means that it has to be defined whether the fact that some observations have missing values is related somehow to the values of the other variables in the dataset, or not (Little and Rubin 2014). Rubin (1976) formalized this concept, by treating the missing data indicator as a random variable and assigning it a distribution. He defined three different scenarios to describe the mechanism underlying the presence of missing data: *missing completely at random* (MCAR), *missing at random* (MAR) and *missing not at random* (MNAR).

The explanation of missing data mechanisms described in this section is based on the books of Little and Rubin (2014) and of Van Buuren (2012). Let $\boldsymbol{Z}$ be a $n \times p$ matrix with n observations of p variables. The $\boldsymbol{Z}$ matrix is partially observed, so that $\boldsymbol{Z} = (\boldsymbol{Z}_{obs}, \boldsymbol{Z}_{mis})$, where $\boldsymbol{Z}_{obs}$ and $\boldsymbol{Z}_{mis}$ denote respectively the subset of fully observed data and the one with missing values. Define the missing data indicator $\boldsymbol{R}$, which is a $n \times p$ matrix, that takes values in $\{0, 1\}$ to define observed and missing values in $\boldsymbol{Z}$ respectively. The relation that might exist between $\boldsymbol{R}$ and $\boldsymbol{Z}$ is described by a missing data model, which is characterized by the conditional distribution of $\boldsymbol{R}$ given $\boldsymbol{Z}$, $f(\boldsymbol{R}|\boldsymbol{Z}, \varphi)$, where φ denotes a vector containing the unknown parameters of the model. Hence, there are three possible scenarios.

MCAR Data are said to be missing completely at random if missingness does not depend on the values of $\boldsymbol{Z}$, that means, if

$$f(\boldsymbol{R}|\boldsymbol{Z}, \varphi) = f(\boldsymbol{R}|\varphi) \tag{1.9}$$

and hence, the probability of being missing only depends on φ, the overall probability of being missing. This assumption does not mean that the

mechanism itself is random, but rather that causes of the missing data are unrelated to the data. In this case, ten, we might also ignore the process that leads to missing data and many of the complexities that arise because of that, apart from the loss of information, and do a *complete case* analysis taking into account only the fully observed records. MCAR data may be generated because, for example, a weighing scale might run out of batteries, a questionnaire may be lost in the post or a blood sample might be damaged in the lab. This assumption can be tested by separating the missing and the complete cases and examining the characteristics of these two groups. If characteristics are equal for both groups, we can assume that data are MCAR, otherwise this assumption does not hold.

MAR Assuming that data are missing at random makes a less restrictive assumption on the underlying mechanism and it defines a scenario where the missingness depends only on the observed components $\boldsymbol{Z}_{obs}$, and not on the missing values. That is, if

$$f(\boldsymbol{R}|\boldsymbol{Z}, \varphi) = f(\boldsymbol{R}|\boldsymbol{Z}_{obs}, \varphi). \tag{1.10}$$

MAR assumption is more general and more realistic than MCAR. For example, if a weighing scale is placed on a soft surface, it may lead to more missing values than when it is placed on a hard surface. In this case, data cannot be MCAR, however, if we know the surface type and we assume MCAR within the type of surface, then the data are MAR. Another example, people who come from poorer families might be less inclined to complete the questionnaire, thus the missingness would be related to family income. Also in this case, if we know the family income and, stratifying for that, missingness can be assumed random, then we can say data are MAR. The key aspect about MAR is that the values of the missing data can somehow be predicted from some of the other variables being studied. The assumption that the mechanism is MAR cannot be confirmed, because it cannot be tested whether

the probability of missing data on a variable is solely a function of other measured variables.

MNAR If neither MCAR nor MAR hypothesis holds, then data are missing not at random, which means that the probability of being missing varies for unknown reasons. This means that the general expression of the missing data model does not simplify and the distribution of $\boldsymbol{R}$ depends on both observed and unobserved information and on the parameters:

$$f(\boldsymbol{R}|\boldsymbol{Z}_{obs}, \boldsymbol{Z}_{mis}, \varphi) \tag{1.11}$$

In public opinion research, an example of MNAR data may occur if those with weaker opinions respond less often than the others, or this is also the case where people with the lowest education are missing on education or the sickest people are most likely to drop out of the study. MNAR is the most complex case. Strategies to handle MNAR are to find more data about the causes of missingness, or to perform sensitivity analyses to see how sensitive the results are under various scenarios.

1.2.2 Multiple imputation

Multiple imputation is a statistical technique to handle missing data, that was developed by Rubin (1978). He thought that imputing only one value (single imputation), in order to estimate the "best", one could not be correct in general, because we cannot know which value to impute with certainty, otherwise it would not be missing. Hence, since the observed and the unobserved data are connected to each other by a statistical model, the method used to impute missing values should reflect this uncertainty. His idea was to create multiple versions of the data, drawing imputations from a distribution. This approach is a Bayesian perspective, where the missing values have a distribution given the observed values. Thus, what we really

want to impute is the predictive distribution of the missing values given the observed values and not a single value (Rubin 1978).

The key point of the multiple imputation procedure is to use the distribution of the observed data to estimate a set of plausible values for the missing data. In this way, multiple datasets are created and subsequently analyzed individually and identically in order to obtain a set of parameter estimates, that are combined together to obtain the final results. When correctly implemented, multiple imputation is asymptotically efficient and produces asymptotically unbiased estimates and standard errors. Two key requirements to gain precision and avoid bias are using all the available covariates for the imputation model. To avoid bias in the analysis model, all the variables that are then used for calibrating the model have to be included in the imputation model, as well as the outcome itself. This point is important to ensure that the imputation model has the ability to reconstruct all the relationships between the variables in the dataset. Moreover, including also predictors of the incomplete variable in the imputation model can improve the analysis. In fact, this makes the MAR assumption more plausible, since it assumes that the probability of data being missing does not depend on the unobserved, conditional on the observed data that are included in the imputation model. Doing that can reduce the bias and improve the imputations (White, Royston, and Wood 2011).

Procedure

The multiple imputation technique consists of three main stages: generating multiply imputed datasets, analyzing multiply imputed datasets and combining estimates from multiply imputed datasets. Figure 1.1 illustrates the three main steps as depicted in the book of Van Buuren (2012, p.17).

As it is shown in Figure 1.1, multiple imputation replaces every missing value with M plausible values drawn from a distribution specifically modelled using the observed data. This results in M completed datasets, which differ

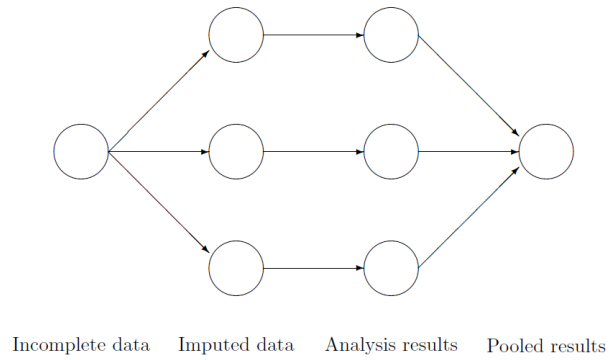

Figure 1.1: Schematic illustration of the main steps in multiple imputation (Van Buuren 2012, p.17)

from each other only for the entries that were missing. After that, each dataset is analyzed and the results are pooled together. These three steps are explained more detailed below.

Step 1: Generating multiply imputed datasets. To generate the imputed values, three tasks exist: the *modelling task*, the *estimation task* and the *imputation task*. The modelling task chooses a model for the data, the estimation task computes a posterior distribution for the parameters of this model and finally, the imputation task takes one random draw from the associated predictive distribution of the missing data given the observed data (Rubin 1978).

Missing values are therefore replaced by M independent set of values, simulated from the posterior predictive distribution of the missing data conditional on the observed data. For a single incomplete variable Z, this means defining an imputation model which regresses Z on a set of completed variables, say $\boldsymbol{X} = (X_1, X_2, \ldots, X_q)$, among all the individuals with the observed Z.

We then have to choose a specific model for the imputation of the missing data, $f(Z_{mis}|\boldsymbol{X};\varphi)$, parametrized by φ. This might be a linear regression model, if we want to impute normally distributed continuous variables, or a logistic regression model to impute binary variables. Several models are possible and the choice has to be made according to the type of variable

whose missing values we wish to impute. For more details about choices of imputation model, see White and Royston (2009). Formally, multiple imputation involves drawing values of the missing data Z_{mis} from the predictive distribution

$$f(Z_{mis}|Z_{obs}, \boldsymbol{X}) = \int f(Z_{mis}|Z_{obs}, \boldsymbol{X}; \varphi) f(\varphi|Z_{obs}, \boldsymbol{X})\, d\varphi \tag{1.12}$$

where $f(\varphi|Z_{obs}, \boldsymbol{X})$ is the Bayesian distribution of φ. Once the imputation model has been chosen, the regression parameters, φ, and the relative covariance matrix have to be estimated. In practice, this may be achieved, with implicit vague priors, by fitting the model $f(Z_{mis}|Z_{obs}; \varphi)$ to the case with Z observed, estimating $\widehat{\varphi}$ with covariance matrix V_φ and drawing a value of φ, say $\varphi*$, from its posterior (which may be approximated by $N(\varphi*, V_\varphi)$). Finally, imputations for Z_{mis} are drawn from $f(Z_{mis}|\boldsymbol{X}; \varphi*)$ (Rubin and Schenker 1986; White and Royston 2009). The estimation and imputation procedure have to be repeated M times, so at the end M datasets are generated.

Step 2: Analyzing multiply imputed datasets. After multiple imputation, the M different imputed datasets are separately analyzed in order to obtain, from each dataset, the quantities of interest (usually regression coefficients). The results of these M analysis differ because the procedure generated different datasets (White, Royston, and Wood 2011).

Step 3: Combining estimates from multiply imputed datasets. The M estimates are finally combined together into an overall estimate and variance-covariance matrix using *Rubin's rules*, which are based on asymptotic theory in a Bayesian framework (Rubin 2004; White, Royston, and Wood 2011). In this case, the goal of multiple imputation is to find an estimate of the quantity of interest that is unbiased and with correct confident

coverage (Rubin 1996). This means that the estimate should be equal, on average, to the value of the population parameter and the associated confidence intervals and hypothesis tests should achieve at least the stated nominal value (Van Buuren 2012).

Suppose $\widehat{\theta}_m$ is an estimate of a quantity of interest obtained from the analysis of the m^{th} imputed dataset and $\boldsymbol{W}_m$ is the corresponding estimate variance. Then, the combined overall estimate $\widehat{\theta}$ is equal to the average of the individual estimates:

$$\widehat{\theta} = \frac{1}{M} \sum_{m=1}^{M} \widehat{\theta}_m \tag{1.13}$$

While the variance of $\widehat{\theta}$ is the sum of the within-imputation variance:

$$\boldsymbol{W} = \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{W}_m \tag{1.14}$$

and the between-imputation variance

$$\boldsymbol{B} = \frac{1}{M-1} \sum_{m=1}^{M} (\widehat{\theta}_m - \widehat{\theta})^2 \tag{1.15}$$

Combining these two measures together leads to the total variance:

$$var(\widehat{\theta}) = \boldsymbol{W} + \left(1 + \frac{1}{M}\right) \boldsymbol{B}. \tag{1.16}$$

1.2.3 Imputing multivariate missing data

When we have a large dataset, with many predictors, it is common that missing values occur in several variables. In this case, the main problem arises when we want to use a regression-based imputation as described in the previous section to impute missing values in X_j. To do that, we need to use all the other predictors X_{-j}, but those variables themselves contain

missing values. Several other practical problems may also occur: for example, the "circular" dependence, that arises when the missing values of two incomplete variables depend on each other because of their correlation, variables may also be of different types (e.g., binary, unordered, ordered, continuous) or collinearity or empty cells might occur as well (Van Buuren et al. 2006). These and many others complexities may arise when we have to deal with multivariate missing data. A strategy to impute missing values in a multivariate context is *fully conditional specification* (Van Buuren 2007).

Fully conditional specification

Fully conditional specification (FSC), also known as chained equations (Van Buuren and Groothuis-Oudshoorn 2011) and sequential regression multivariate imputation (Raghunathan et al. 2001), is a method to impute data in a variable-by-variable basis, by specifying an imputation model per variable (Van Buuren 2007).

Suppose that $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_k)$ is a set of variables which contains missing values and $\boldsymbol{X}$ is the set of completely observed variables, while $\boldsymbol{R}$ is the already defined indicator of missing values. This approach defines $P(\boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{R}|\varphi)$ by specifying a conditional density $P(Z_j|\boldsymbol{Z}_{-j}, \boldsymbol{X}, \boldsymbol{R}, \varphi_j)$ for each Z_j. Hence, Z_j^{mis} values are imputed given $\boldsymbol{Z}_{-j}$, $\boldsymbol{X}$ and $\boldsymbol{R}$, while the multivariate distribution of φ is obtained (either explicitly or implicitly) by sampling iteratively from conditional distributions. This procedure starts from simple guessed values and then, imputation under FCS is done by iterating over all conditionally specified imputation models (Van Buuren 2007). FCS is the natural generalization of univariate imputation discussed in the previously section, the main difference is that FSC does not need to specify a multivariate model for the data, because it directly defines the conditional distributions from which draws should be made (Van Buuren 2012).

Multiple imputation by chained equations

Multiple imputation by chained equations (MICE) is an algorithm proposed by Van Buuren and Oudshoorn (2000) as a practical approach to generate imputations, under conditionally specified models, one for each variable with missing values.

The algorithm starts filling in all missing values by simple random drawing from the observed values. Then, the first variable with missing values, Z_1 is regressed on all the other variables, $Z_2, \ldots, Z_k$, restricted to individuals with the observed Z_1 and missing values in Z_1 are replaced with simulated draws from the corresponding posterior predictive distribution of Z_1. After that, the second variable with missing values, Z_2, is regressed on all the other variables, restricted to those observations with observed Z_2, but this time, the new imputed values of Z_1 are used. After missing values of Z_2 have been imputed, the process is repeated for all the other variables with missing values: this is called a cycle. To stabilize the results, the procedure is repeated for several cycles, usually 10 or 20, to produce a single imputed dataset. Finally, in order to have M imputed datasets, the entire procedure is repeated M times (White, Royston, and Wood 2011).

The MICE algorithm can handle different types of variables, because each variable is imputed using its own imputation model. Moreover, the choice of the conditional distributions is made by the user and so, the joint distribution is only implicitly known (Van Buuren 2012).

1.3 Cross-validation

Assessing performance of a model relates to its predictive capability on independent data and it is very important, especially in practice. Ideally, we would like to assess the performance of our model using a set of observations, that is independent from the one used to calibrate the model. If we had enough data, we could split the original dataset in two parts and use one of

them for the calibration and set aside the other one as validation set. This is not always possible, since the available datasets might be too small to allow this kind of procedure. To deal with that, several techniques have been developed and *cross-validation* is one of the most famous and widely used methods to estimate prediction error of a model using part of the available dataset to fit the model and a different one to test it.

First of all we split the dataset in K equal-sized parts and we set aside the k^{th} fold (*validation set*) and fit the model using the other $K-1$ parts of the data (*calibration set*). Finally, we calculate the prediction error of the fitted model when predicting the k^{th} part of the data. The previous steps have to be done for $k = 1, 2, \ldots, K$ and then the K estimates have to be combined together in order to obtain the prediction error of the model (Friedman, Hastie, and Tibshirani 2009).

The most common choices for K are 5 or 10, but it is intuitive that the method becomes more accurate with increasing K. The maximum possible value for K is N and this procedure is called *leave-one-out* cross-validation. With $K = N$, the cross-validation estimator is approximately unbiased for the true (expected) prediction error, but can have high variance because the N *calibration set* are obviously very similar to one another. The opposite problem occurs with low values of folds, for example with $K = 5$, because even if in this case the variance is lower, the bias could be a problem and the procedure may overestimate the true prediction error (Friedman, Hastie, and Tibshirani 2009). Overall, five- or tenfold cross-validation are a good compromise (Kohavi 1995). The crucial point of this procedure is that it must be done at the very beginning of the analysis and, in case of a multistep modeling procedure, it must be applied to the entire sequence of modeling steps. This is a crucial point of the cross-validation procedure, because it basically does the analysis K times and each time is independent of each other. Thus, to ensure the analysis is not biased, when the model is calibrated in one set of the data, the procedure must not "see" the outcome of the test set.

Chapter 2

Combining multiple imputation and cross-validation

This chapter presents the theoretical issues that arise when multiple imputation and cross-validation are used at the same time and finally, our proposal to deal with that. A general theoretical explanation of the problem is described in section 2.1, together with the description of the most proper way to get predictions in this situation. In section 2.2, instead, we define 2 approaches specially designed to handle with these issues, along with an implementation for survival data.

2.1 Theoretical issues

2.1.1 Imputing missing values

When we want to asses the performance of a prediction model using cross-validatory assessment and, at the same time, multiple imputation is used to account for missing values, a problematic conflict between these two procedures arises. In fact, as it is described in section 1.3, validation, and above all cross-validation, requires outcome to be removed from the calibration data and then predicted applying the calibrated prediction rules,

in order to compare predicted values with the actually observed outcome. On the other hand, multiple imputation requires the outcome data as an integral part of the estimation of the imputation model, in order to preserve the association between predictors and outcome in the imputed values. Basically, during the cross-validation procedure, calibration models should not "see" the outcome in the validation set, but this actually implicitly happens if multiple imputation is done once at the beginning of the analysis, because the outcome is used to generate the imputed values.

This problematic can also be extended to a general situation, where the aim of the analysis is to predict a future outcome. For example, when we have a set of individuals, of which we only know the predictors values, and we would like to predict their future outcome using the model calibrated in advance on a previous set of individuals. In this case, the new set of observations may also contain missing values, which should be imputed using both the predictor variables as well as the outcome information to preserve all the relationships between covariates in the dataset. The main problem in this case is that we do not have the outcome yet. To solve this problem, we should then estimate the imputation model borrowing information from both sets of observations, that means by building the imputation model on the two sets together, treating them as a unique dataset. Once all the missing values have been imputed, we can then use the "old" set to calibrate the model and finally get the predictions for the "new" one.

This argumentation can also be generalized to a cross-validation scenario where, for example, the aim of the study is to calibrate a predictive model and using a cross-validatory assessment on data that contains missing values. The conflict generated by these two approaches can actually be solved, as we have seen in the previous paragraph. In fact, once we have split the data into K folds and have defined the k^{th} fold as validation set and all the others as calibration set, we are exactly in the same situation described above, where the validation set represents the "new" observations set, while the calibration set the "old" one. Now, to correctly use the cross-validation

technique and to then take advantage of its theoretical properties, the outcome of the validation set cannot be seen by the calibration procedure and we could actually consider it as missing.

2.1.2 Getting predictions

Another issue, regarding above all multiple imputation, concerns Rubin's rules, as they are described in section 1.2. The Rubin's rules are generally used after the imputation procedure to summarize the estimates obtained from the calibration of the M models in order to get the assumed substantial model, which is used later for the prediction of the new outcome. In the predictive scenario, however, and from a formal probabilistic point of view, the effect estimate are nuisance parameters and predictions are obtained from the calibrated posterior predictive density with the missing observations and effect measures integrated out (Lesaffre and Lawson 2012). Hence, final predictions could actually be obtained in two different ways. The most common one is to use straightforwardly Rubin's rules, that means pooling together the M sets of coefficients obtained from the calibration of the model on the M imputed datasets and using them to get final predictions. While the other one implies to use separately the M sets of coefficients obtained from the M calibrations in order to get M predictions, that will be finally pooled together to get the final predictions.

Pooling coefficients

Let Y be the outcome of interest and $\boldsymbol{X}$ a set of predictors. We assume a substantive prediction model $f(Y|\boldsymbol{X}, \boldsymbol{\beta})$, which describes the variation in an univariate outcome Y conditional on the predictor matrix $\boldsymbol{X}$ and depending on $\boldsymbol{\beta}$, which is an unknown vector of regression parameters. The latter has to be estimated in order to subsequent use of the model. In this work we only consider scenarios with fully observed Y and missing values in the predictors,

such that $\boldsymbol{X} = (\boldsymbol{X}_{mis}, \boldsymbol{X}_{obs})$, where $\boldsymbol{X}_{mis}$ is the set of predictors with missing values, while $\boldsymbol{X}_{obs}$ is the set with fully observed components. If we were interested in estimating the parameters $\boldsymbol{\beta}$ in the presence of missing values, then we can calibrate the conditional density:

$$\begin{aligned} p(\boldsymbol{\beta}|\boldsymbol{X}_{obs}, Y) &= \int p(\boldsymbol{\beta}, \boldsymbol{X}_{mis}|\boldsymbol{X}_{obs}, Y)\ d\boldsymbol{X}_{mis} \\ &= \int p(\boldsymbol{\beta}|\boldsymbol{X}_{mis}, \boldsymbol{X}_{obs}, Y)\ p(\boldsymbol{X}_{mis}|\boldsymbol{X}_{obs}, Y)\ d\boldsymbol{X}_{mis} \end{aligned} \tag{2.1}$$

which is obtained as the marginalized joint density on the two unknown components $\boldsymbol{\beta}$ and $\boldsymbol{X}_{mis}$, marginalized across the unobserved values in $\boldsymbol{X}_{mis}$. The first equality may also be written as the probability density of the parameters vector $\boldsymbol{\beta}$, conditional on the unknown quantities $\boldsymbol{X}_{mis}$ and averaged across the uncertainty in $\boldsymbol{X}_{mis}$, both conditional on the observed data.

The multiple imputation procedure, based on Rubin's rules, represents a practical approximation to this latter integration, by first generating imputed data, $\widehat{\boldsymbol{X}}_m^{mis}$, sampling from the conditional density $p(\boldsymbol{X}_{mis}|\boldsymbol{X}_{obs}, Y)$, with $m = 1, \ldots, M$ for a total number of M imputations. After that, we estimate the modes $\widehat{\boldsymbol{\beta}}_m$ of the conditional densities $p(\boldsymbol{\beta} \mid \widehat{\boldsymbol{X}}_m^{mis}, \boldsymbol{X}_{obs}, Y)$ evaluated at the "completed" datasets $(\widehat{\boldsymbol{X}}_m^{mis}, \boldsymbol{X}_{obs}, Y)$ for all m. Finally, the conditional density $p(\boldsymbol{\beta} \mid \boldsymbol{X}_{obs}, Y)$ is approximated using classical frequentist theory and this gives rise to the so-called Rubin's rules estimate of the expectation as

$$\widehat{\boldsymbol{\beta}}_{MI} = \frac{1}{M} \sum_{m=1}^{M} \widehat{\boldsymbol{\beta}}_m. \tag{2.2}$$

In a predictive scenario, where the study aim is to get predictions, for a new set of data, these can be finally obtained using the pooled model.

Pooling predictions

In the predictive scenario, the averaging described in equation 2.1 should be expanded to average across the regression coefficients in order to account

for both the missing values $\boldsymbol{X}_{mis}$ and the uncertainty in $\boldsymbol{\beta}$.

Let $\widetilde{Y}$ be a future outcome, which we want to predict using covariates $\widetilde{\boldsymbol{X}}$. To simplify the discussion, in the first instance we assume no missing data in future data. The calibration data we use to calibrate the model includes the outcome Y and the regression variables $\boldsymbol{X}$. To predict a future $\widetilde{Y}$, we calibrate the target density as $p(\widetilde{Y}|\boldsymbol{X}_{obs}, Y)$, which denotes the conditional dependence of $\widetilde{Y}$ on the past observed calibration data Y and $\boldsymbol{X}_{obs}$.

In the presence of missing values, the predictive density for future outcome outcomes $\widetilde{Y}$ can be calibrated as

$$\begin{aligned} p(\widetilde{Y}|\boldsymbol{X}_{obs}, Y) &= \int f(\widetilde{Y}, \boldsymbol{\beta}, \boldsymbol{X}_{mis}|\boldsymbol{X}_{obs}, Y)\ d\boldsymbol{\beta}\, d\boldsymbol{X}_{mis} \\ &= \int f(\widetilde{Y}|\boldsymbol{\beta}, \boldsymbol{X}_{mis}, \boldsymbol{X}_{obs}, Y)\ p(\boldsymbol{\beta}, \boldsymbol{X}_{mis}|\boldsymbol{X}_{obs}, Y)\ d\boldsymbol{\beta}\, d\boldsymbol{X}_{mis}. \end{aligned} \tag{2.3}$$

The integration is then achieved by averaging across both imputations $\widehat{\boldsymbol{X}}_m^{mis}$ and simulations $\widehat{\boldsymbol{\beta}}_m$ from the density $p(\boldsymbol{\beta}, \boldsymbol{X}_{mis}|\boldsymbol{X}_{obs}, Y)$, always conditioning on the observed calibration data. In analogy with parameter estimation, we then could calculate expectations:

$$\widehat{Y}_m = E[f(\widetilde{Y} \mid \widehat{\boldsymbol{\beta}}_m, \widehat{\boldsymbol{X}}_m^{mis}, \boldsymbol{X}_{obs}, Y)] \tag{2.4}$$

for each pair of imputed values $\widehat{\boldsymbol{\beta}}_m$, $\widehat{\boldsymbol{X}}_m^{mis}$, from the conditional density $p(\boldsymbol{\beta}, \boldsymbol{X}_{mis}|\boldsymbol{X}_{obs}, Y)$. The set of predictions $\widehat{Y}_m$ for $m = 1, \ldots, M$, might be summarized using a suitable summary measure, like the mean or the median, in order to get the final prediction estimate $\widehat{Y}$. For example, using Rubin's rules to summarize the set of predictions $\widehat{Y}_m$, $m = 1, \ldots, M$, the quantity $E[\widehat{Y}\ |\boldsymbol{X}_{obs}, Y]$ would be estimated using

$$\widehat{Y}_{MI} = \frac{1}{M} \sum_{m=1}^{M} \widehat{Y}_m. \tag{2.5}$$

In the previous paragraph we only describe how to get predictions when

we have no missing values in the new observations. It may also happen though that the future outcomes have themselves missing values in the predictors, such that $\widetilde{\boldsymbol{X}} = (\widetilde{\boldsymbol{X}}_{obs}, \widetilde{\boldsymbol{X}}_{mis})$, and they also might not occur in the same covariates containing missing values in the calibration data. In case of missing values, the equation 2.4 should then be expanded in order to include averaging across $\widetilde{\boldsymbol{X}}_{miss}$ and to obtain predictions we will then have

$$\widehat{Y} = E[\widetilde{Y} \mid \widetilde{\boldsymbol{X}}_{obs}, \boldsymbol{X}_{obs}, Y] \tag{2.6}$$

2.2 Methods

This section presents a general approach to validation, which enables to deal with the problem discussed above. First of all, this approach allows the outcome of the validation set $\widetilde{Y}$ to be set-aside during the calibration of the imputation model to impute $\widetilde{\boldsymbol{X}}_{mis}$ and $\boldsymbol{X}_{mis}$ and thus, this subsequently allows to use it for the validation of the prediction rules. In section 2.2.1, we then propose two different algorithms to get final predictions, that means by directly estimating the outcome by pooling predictions or, in contrast, by applying Rubin's rules for the parameter estimation and afterwards getting predictions. In section 2.2.2 we also define the naïve implementation of the previous approaches in order to use them as comparison during the analysis. Finally, in section 2.2.3, we describe the implementation for survival outcomes.

This discussion focuses on cross-validation, but it could also be adapted for a single set-aside validation set.

2.2.1 Combining cross-validation and multiple imputation

A general approach to generate imputations without considering the outcome of the validation test is to remove it ($\widetilde{Y}$) from the left-out fold defined within the cross-validation procedure. This can be achieved by setting the outcome of the left-out fold as "missing". After that, multiple imputation can

be used to impute missing values in $\widetilde{\boldsymbol{X}}_{mis}$ and $\boldsymbol{X}_{mis}$ by calibrating the imputation model on the remainder of the observed data $(\widetilde{\boldsymbol{X}}_{obs}, \boldsymbol{X}_{obs}, Y)$. After missing values have been imputed using the multiple imputation procedure, a prediction model can be fitted on the calibration data $(\widehat{\boldsymbol{X}}_{mis}, \boldsymbol{X}_{obs}, Y)$ and subsequently applied to predict the outcome of the validation set, using $\widehat{\widetilde{\boldsymbol{X}}}_{mis}$ and $\widetilde{\boldsymbol{X}}_{obs}$. Imputed values of $\widetilde{Y}$ in the left-out fold are then discarded and the real outcome values are returned in order to repeat the entire procedure for the next fold within the whole sequence defined by the cross-validation at the beginning.

As mentioned before, the multiple imputation procedure generates M imputed datasets and final predictions can be obtained in different ways, by pooling predictions obtained from the analysis of the M imputed datasets, as described in *Approach 1*, or by applying Rubin's rules to get a pooled parameters vector and then getting predictions from that one, as described in *Approach 2*. Note that the approaches coincide for $M = 1$.

Once we have the final predictions, these can then be compared with the original outcome to get assessment measures.

Approach 1

The first approach starts defining K folds on the entire dataset. After that, for each left-out fold, one realization of the multiple imputation procedure is run to get a complete dataset on which a suitable model is then fitted and corresponding predictions for the outcome of the left-out fold are generated, as described above. This procedure is then repeated M times in order to get M predictions for each individual. The K folds may also be re-defined each time in order to add extra variation. In this way then, M predictions for each individual are generated and the final predictions vector can be derived by taking the mean, the median or other suitable summary measure within each individual. Note that individual final predictions are derived by using M different models and extra variation is add by fold definitions.

A schematic diagram of this approach is shown at the end of this chapter in figure 2.3.

Approach 2

The second approach starts defining K folds on the whole dataset, but this time they will be kept fixed for the entire procedure. For each left-out fold, multiple imputation is run M times, so that at the end, M completed datasets are generated. For each of them, a suitable model is then fitted on the calibration set to get the corresponding parameters. These M parameters are then pooled together using Rubin's rules in order to obtain the "final" model. The latter is then applied to the validation set of each imputed dataset to get predictions. Since we have M validation sets, at the end of this procedure, each subject has M predictions which will be pooled together within each individual using a suitable summary measure to get the final predictions. Note that the latter will of course all coincide for complete records.

A schematic diagram of this approach is shown at the end of this chapter in figure 2.4.

2.2.2 Naïve approaches

Naïve approaches are defined in analogy to the above approaches. Essential difference is that, in this case, first of all multiple imputation is used to get a set of complete datasets and, only after that, cross-validation is separately done using the complete datasets. Thus, both of the naïve methods start computing multiple imputation M times on the whole dataset, in order to obtain M complete datasets for the analysis. At this point, the *Naïve 1* approach defines K folds in each imputed dataset for the cross-validation procedure and in turn the k^{th} fold is selected to be the validation set, while on the others, a prediction model is fitted in order to obtain predictions for

the individuals in the left-out fold, in analogy with *Approach 1*. After that this procedure has been completed for each fold, each subject has M predictions, one from each complete dataset, which will be pooled together using a suitable summary measure. Alternatively, the *Naïve 2* approach defines K fixed folds, which will be the same on each imputed dataset, and in turn, the k^{th} fold is defined as validation set, while on all the others a predictive model is calibrated. Once this operation has been applied in each imputed dataset, the M resulted parameters are pooled together using Rubin's rules to get the "final" model, in accordance with *Approach 2*. The pooled parameters vector is then used to obtain predictions for the individuals in the validation sets. After this procedure has been computed for each fold, also in this case, each subject has M predictions, which will be pooled together to obtain the final predictions vector.

2.2.3 Implementation for survival data

The previous sections presented a general approach to solve the problem of the combination of multiple imputation and cross-validation, but in principle, this question arose during the analysis of survival data. These data will be presented in the next chapter and the approaches will be also tested using real and simulated survival data (chapter 3). However, some additional aspects must be considered to apply these approaches to survival outcomes. Moreover, to emphasize the specific application for survival outcome, from now on, the notation will be switched from Y to T, when discussing lifetime outcome, in addition to a status indicator δ to denote censoring.

First of all, it is really important to find the right way to include the survival outcome in the imputation model because, otherwise, the association between covariates and survival is likely to be biased. For this reason, White and Royston (2009) showed how imputation models should be constructed considering the censoring indicator and an estimate of the cumulative hazard for the observed individual follow-up time, in addition to the regression

covariates. The estimate of the cumulative hazard can be obtained by using the Nelson-Aalen estimator and this should also be used instead of the survival time. Denoting by $t_1 < t_2 < \ldots$ the times when events are observed and defining d_j as the number of individuals who experienced the event at time t_j, the Nelson-Aalen estimator for the cumulative hazard rate function has the form:

$$\widehat{A}(t) = \sum_{t_j \leq t} \frac{d_j}{r_j} \tag{2.7}$$

where r_j is the number of individuals at risk just prior to time t_j. To respect the cross-validatory logic, the Nelson-Aalen estimate is computed from the data only for the calibration set corresponding to any left-out fold. The Cox models are instead estimated using the original outcome data within the calibration set.

A second issue concerns the *Approach 2* and in particular, how to construct the combined model for subsequent application in prediction. In fact, the Cox regression models involve both the regression parameters (hazard ratios) as well as the baseline hazards to vary across imputations and thus, both sources of variation must be considered. There are two methods to do that, which we will denote as *Approach 2A* and *2B* respectively. The first one (*2A*) consists of averaging both the regression parameters as well as the baseline hazards separately, and use them to define the final model for predictions. The second one (*2B*) applies Rubin's rules only for the combination of the regression parameters, which will then be averaged together, in order to get the estimate of the cumulative baseline hazard with the Breslow estimator. Denoting by T_i the observed survival time and by δ_i the status indicator, the Breslow estimator takes the form:

$$\widehat{\Lambda}_0(t) = \sum_{i=1}^{N} \frac{I(T_i \leq t)\ \delta_i}{\sum_{j \in R_i} exp\{\boldsymbol{x}_j' \widehat{\boldsymbol{\beta}}\}} \tag{2.8}$$

where $R_i = \{j : T_j \geq T_i\}$.

We have then defined 3 implementations for Cox proportional hazards

modeling, which we will refer to in tables and graphs as *Approach 1, 2A* and *2B*. A summary in pseudo-code of *Approach 1* is shown in figure 2.1, while *Approaches 2A* and *2B* are shown in figure 2.2.

The naïve approaches for Cox proportional hazards modeling are defined in analogy to those presented in section 2.2.2 making the same changes discussed above. We have then also defined 3 naïve approaches which we will refer to as *Naïve 1, 2A* and *2B*.

All analyses were performed using R Statistical Software (version 3.4.3, R Core Team (2017)) and multiple imputations were generated using the chained equations methodology, already discussed in section 1.2.3 and implemented by Van Buuren and Groothuis-Oudshoorn (2011) in the package MICE. All covariates have also been included in the same functional form as in analysis model and no variable selection has been done.

Approach 1

Define M and repeat the following steps M times:

1. Define K folds for the CV procedure.
2. Select each fold in turn as the validation set and use the others as calibration set. Run the following steps for each selected fold:
   (a) Remove the outcome from the validation data.
   (b) Compute the Nelson-Aalen estimate for the cumulative hazard in the calibration data and replace the original calibration values of the "Time" variable with it.
   (c) Run a single imputation on this dataset.
   (d) Remove the Nelson-Aalen estimate and restore the original "Time" data to the calibration data only.
   (e) Fit a Cox PH model on the calibration set.
   (f) Derive predictions from this model for subjects in the imputed validation set, using the equation:

$$\widehat{S}_{i,m}(t) = \widehat{S}_{0,m}(t)^{exp\{\boldsymbol{x}'_{i,m}\ \widehat{\boldsymbol{\beta}}_m\}}$$

Compute the final prediction $\widehat{S}_i(t)$ for each individual as the average of all the M individual predictions.

Figure 2.1: Algorithmic description of Approach 1 for combination of multiple imputation and cross-validation using the Cox model.

Approaches 2A and 2B

Define K folds for the CV procedure and select each fold in turn as the validation set and use the others as calibration set. Run the following steps for each selected fold:

1. Remove the outcome from the validation data.
2. Compute the Nelson-Aalen estimate for the cumulative hazard in the calibration data and replace the original calibration values of the "Time" variable with it.
3. Run M imputations on this dataset.
4. Remove the Nelson-Aalen estimate from each imputed dataset and restore the original "Time" data to the calibration sets only.
5. Fit separate Cox PH model on the calibration set of each of the M imputed datasets.
6. Compute the average $\overline{\boldsymbol{\beta}}$ of the M coefficients vectors from these models.
7. Compute the baseline survival:
   - **For Approach 2A**, calculate the combined baseline hazard as the average of the M baseline hazards
   - **For Approach 2B**, calculate the Breslow estimate of the baseline hazard from $\overline{\boldsymbol{\beta}}$
8. Derive predictions from the combined model for subjects in the imputed validation sets, using the equation:

$$\widehat{S}_{i,m}(t) = \widehat{S}_0(t)^{exp\{\boldsymbol{x}'_{i,m}\ \overline{\boldsymbol{\beta}}\}}$$

Compute the final prediction $\widehat{S}_i(t)$ for each individual as the average of all the M individual predictions.

Figure 2.2: Algorithmic description of Approaches 2A and 2B for combination of multiple imputation and cross-validation using the Cox model.

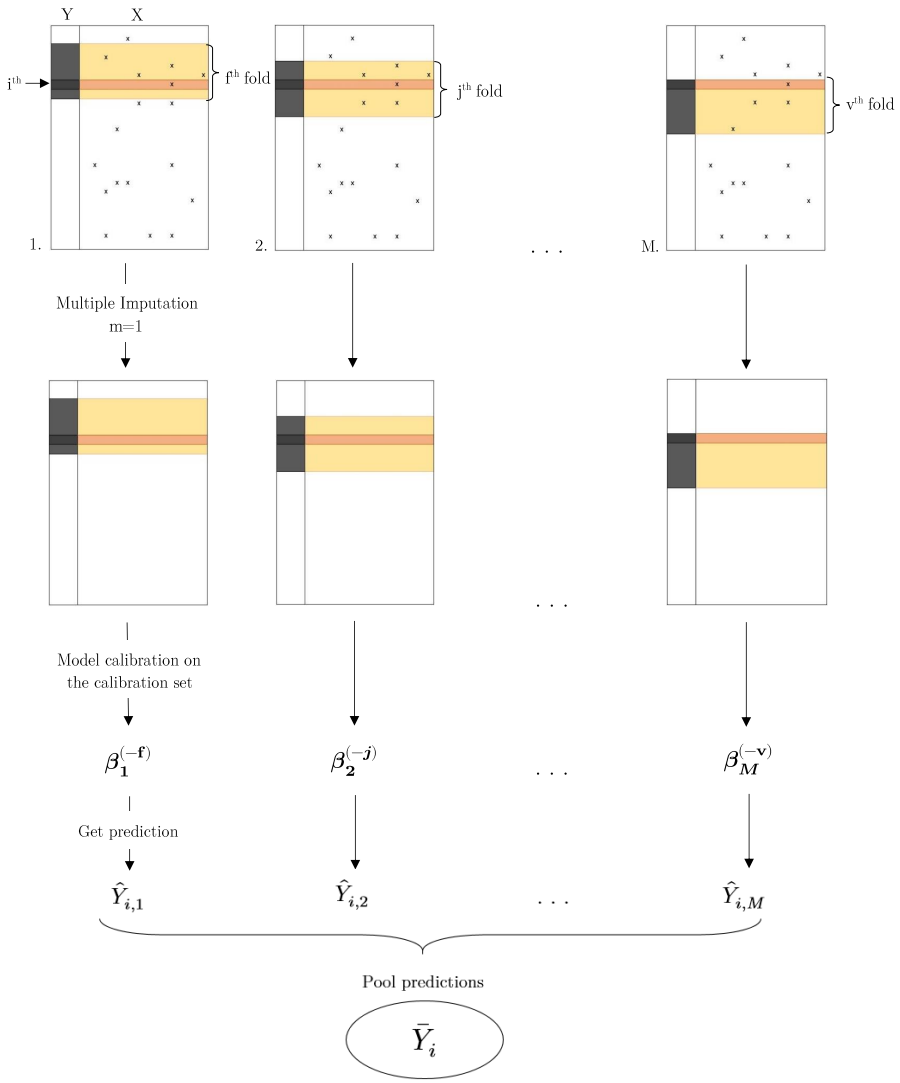

Figure 2.3: Schematic representation of *Approach 1* for combination of multiple imputation and cross-validation. Missing values are represented with "x".

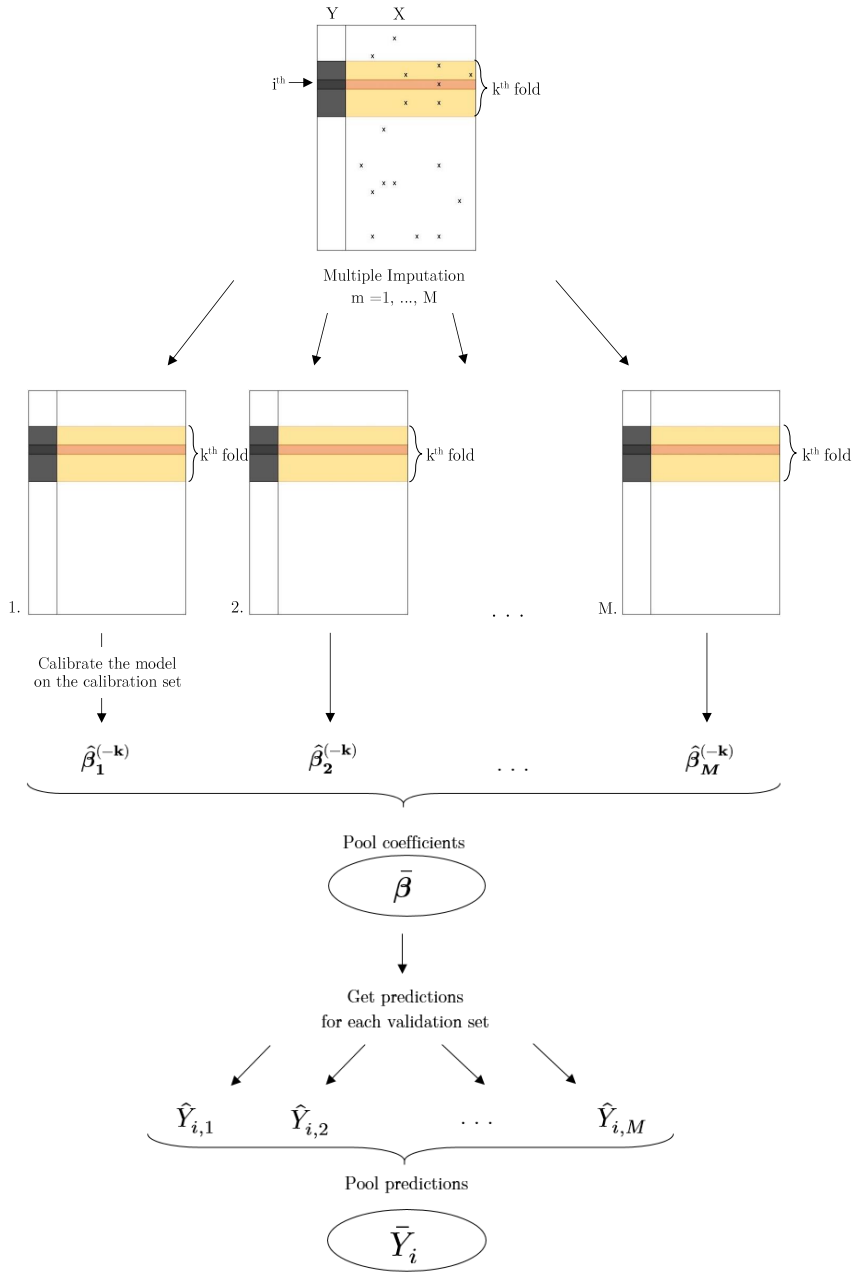

Figure 2.4: Schematic representation of *Approach 2* for combination of multiple imputation and cross-validation. Missing values are represented with "x".

Chapter 3

Application in real and simulated data

This chapter investigates performance of the proposed methodologies. First of all, we will use two real datasets, which generated the interest in this research field, to study performance in real data applications in clinical survival analysis. The Cox proportional hazards model is used to calibrate prognostic models to take censoring into account, regressing on all variables without variable selection. Results for predicted survival probabilities at 1 and 5 years of follow-up are presented. Furthermore, performance of the proposed methods is also investigated using simulations, which are designed to test methodologies under various combinations of missing values patterns and strength of association of predictors with the outcome.

Since this work aims to investigate performances of the proposed methodologies and it is not focused on the data themselves, section 3.1 provides just a brief introduction to the datasets used for the analyses, as they have already been described in previous works. In particular, section 3.1.1 describes the CRT dataset analysed by Hoke et al. (2017), while section 3.1.2 presents the CLL dataset studied by Schetelig et al. (2017a,b). Section 3.2 provides a description of the simulation study and finally, section 3.3 presents the summary measures used to evaluate approaches on real and simulated data.

3.1 Data

3.1.1 CRT data

The CRT (Cardiac Resynchronization Therapy) data has been collected by the Department of Cardiology of *Leiden University Medical Center* (LUMC) and consists of an observational cardiology cohort of 1053 patients. Cardiac resynchronization therapy is a treatment option for individuals with heart failure, especially for those resistant to drugs, and consists of the implantation of a specific device in the heart, which sends small electrical impulses to help both chambers of the heart to beat together in a more synchronized pattern. However, the benefits of this treatment are not guaranteed, because they depend on the characteristics of the patient. To avoid patients have to undergo unsuccessful implantations, it would be helpful to be able to predict their short- and long-term survival probabilities. For this reason, the study of Hoke et al. (2017) aimed to derive a multi-parametric prognostic risk score (CRT-SCORE) using pre-implantation variables, for use in the shared decision-making between patients with heart failure and their physicians.

Data consists of 1053 patients, who underwent CRT implantation between 1999 and 2003. Survival outcome was defined as all-cause mortality (494 deaths (47%), of which 438 are cardiovascular related). The median follow-up is 60 months, while the median survival time is 85 months. Furthermore, data were artificially censored after 7 years (84 months). A total of 430 deaths occurred during this period of follow-up. There are 14 predictor variables: age at implantation (*Age*, continuous), gender (*Gender*, two categories), New York Heart Association functional class (*Nyha*, three categories), etiology of heart failure (*Et*, two categories), diabetes mellitus (*Dm*, two categories), mitral regurgitation (*Mr*, two categories), left ventricular diastolic dysfunction (*Lvdias*, two categories), left bundle branch block (*Lbbb*, two categories), atrial fibrillation (*Af*, two categories), estimated glomerular filtration rate (*Egfr*, continuous), hemoglobin levels (*Hb*, continuous), left ventricular ejection fraction (*Lvef*, continuous) and QRS duration (*Qrs*, two

categories). Information is missing in 529 records, of which the majority is concentrated in the *Lvdias* variable, which is missing in 524 cases (50%). In addition, missing values occur also in *Egfr* (2 cases), *Hb* (7 cases), *Lvef* (20 cases) and in *Mr* (30 cases). Missing observations in *Lvdias* were due to failure of the measuring device, which give some credence to the missing completely at random assumption.

3.1.2 CLL data

The CLL (Chronic Lymphocytic Leukemia) data has been extracted from the registry of the *European Society for Blood and Marrow Transplantation* (EBMT) and describes the risk factors and outcomes of a cohort of patients with chronic lymphocytic leukemia who received an allogeneic hematopoietic stem cell transplantation. Data have already been analysed in two papers by Schetelig et al. (2017a,b) in order to study the impact on several outcomes of a large series of risk factors, including patient-, disease-, procedure- and center-related information. For this work, a simplified version of the data analysed in Schetelig et al. (2017b) is used.

Data consists of 694 retrospective observations of patients who were transplanted between 2000 and 2011. The outcome of interest is overall survival up to 5 years after first allogeneic stem cell transplantation and it was 64% at 2 years and 47% at 5 years. In addition, data were artificially censored after 5 years and a total of 314 deaths were observed during this period of interest. There are 8 predictor variables: age at transplantation (*age10*, continuous), performance status indicated by the Karnofsky Index (*perfstat*, four categories), remission status at transplantation (*remstat*, three categories), cytogenetic abnormalities (*cyto*, four categories), previous autologous transplantation (*asct*, two categories), donor type (*donor*, three categories), patient-donor sex match (*sex_match*, four categories) and conditioning regimen (*cond*, three categories). Information is missing in 241 records, in particular for *cyto* (171 cases, 25%), *perfstat* (63 cases, 9%), *remstat* (42

cases, 6%), *cond* (9 cases, 1%) and finally, for *sex_match* (8 cases, 1%).

3.2 Simulation study

3.2.1 Simulating lifetimes

Dataset are randomly generated resembling the CRT data to some extent, especially the fact that missing values are almost uniquely confined to a single predictor. In particular, simulated data consists of survival time T, a censoring status indicator δ and a predictor matrix $\boldsymbol{X}$ with 4 continuous covariates, which are drawn from a multivariate normal distribution $N_4(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\mu_j = 0$, $j = 1, \ldots, 4$. The covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1. & -0.5486 & -0.1442 & 0.0617 \\ -0.5486 & 1. & 0.2970 & 0.1189 \\ -0.1442 & 0.2970 & 1. & -0.0210 \\ 0.0617 & 0.1189 & -0.0210 & 1. \end{bmatrix}$$

is chosen to equal the sample covariance matrix between the standardized continuous variables in the CRT data (Age, Egfr, Hb, Lvef).

The survival times T_i, where i denotes the i^{th} individual, with $i = 1, \ldots, N$ and $N = 1000$, are drawn from an Exponential distribution with hazard

$$h(t|\boldsymbol{x}_i) = \lambda \; exp\{\boldsymbol{x}_i^{'}\boldsymbol{\beta}\} \tag{3.1}$$

where $\boldsymbol{x}_i$ is the corresponding vector of predictors and the baseline hazard is fixed $\lambda = 0.0073$. The hazard ratios are chosen as

$$\boldsymbol{\beta}^{'} = (\beta_1, \; log(1.2), \; log(0.85), \; log(0.75)) \tag{3.2}$$

such that $\beta_2, \; \beta_3, \; \beta_4$ are fixed, while β_1 varies across simulation scenarios. Censoring times are drawn from a Uniform distribution between 13.5 and 167.5, resembling the CRT dataset. The observed follow-up time T is defined

for each individual by taking the minimum between the generated survival and censoring times. The status indicator δ is set to 0 when T corresponds to a censoring time, and to 1 when T is an event time point. Finally, administrative censorship is applied at $t = 84$ months, as for the CRT data. The choices for λ and $\boldsymbol{\beta}$ were made so that the simulated data has similar survival proportions and levels of censoring as in the CRT data at 1 and 5 years.

3.2.2 Missing values scenarios

Once the simulated outcomes and predictors have been generated, missing values are introduced by removing a percentage of observations from the X_1 variable. As in the CRT data, missing values are concentrated in one variable. The percentage of missing values in X_1 and the value of the regression coefficient β_1 are chosen in order to generate 4 different scenarios, which are defined as all combinations of low or high association of X_1 with the outcome ($\beta_1 = log(1.1)$ or $log(2)$) and low or high percentage of missing values in X_1 (10% or 50%), as shown in table 3.1.

Association between X_1 and the outcome	% of missing values in X_1	
	Low [10%]	High [50%]
Low [$\beta_1 = log(1.1)$]	Scenario 1	Scenario 3
High [$\beta_1 = log(2)$]	Scenario 2	Scenario 4

Table 3.1: Definition of the simulation scenarios.

For each scenario, missing values are introduced *completely at random* (MCAR) or *at random* (MAR), such that we have 8 scenarios in total. MAR observations are generated by calculating for each individual i the probability of being missing, given X_2, defined by the equation below:

$$p_i^{MAR} = min\left[\frac{x_{2,i}^* \, L}{\overline{X_2^*}}, 1\right] \tag{3.3}$$

where $x^*_{2,i} = (x_{2,i} - min(X_2))/(max(X_2) - min(X_2))$, L is a fraction between 0 and 1 chosen to define the percentage of missing values as described in table 3.1 and $\overline{X}^*_2$ is the mean of the X_2 variable. We take the minimum value between 1 and the generated value to avoid exceeding 1. For each individual a draw is then generated from the Bernoulli density with probability p_i^{MAR}.

For each of the above described 4 simulation scenarios and for both MAR and MCAR, we generate $S = 100$ simulated datasets, with $N = 1000$.

3.3 Comparison statistics to assess performance of a predictive model

This section introduces the comparison statistics used to assess the performance of the proposed methodologies in a survival analysis context. First of all, we present a calibration and discrimination measure, as they are used to assess survival predictions. Finally, we present some *ad hoc* statistics, expressly created to evaluate the effect of multiple imputation on predictions.

All statistics and summary measures are calculated based on the output from the K-folds cross-validatory approaches described in section 2.2, using $K = 10$. Furthermore, for each simulated dataset, $R = 10$ replications of each approach are run to account for imputation variation.

3.3.1 Calibration and discrimination measures

When the study aims to build a prediction model, it is very important to assess its predictive performance in a new set of data. This model evaluation process is usually called *model validation*. The general idea of validating a prediction model is to establish that it performs well also for new observations, and this is very important, especially in a health research context. When validating a prediction model, the predictive performance of the

model is commonly addressed by quantifying the agreement between the observed and predicted outcomes, *calibration*, and the ability of the model to distinguish between low and high risk patients, *discrimination*. Several performance measures based on these concepts are well established for risk models for binary outcomes, but for survival prediction models the presence of censoring in the validation data has to be considered to avoid biased results (Rahman et al. 2017).

In this work we use the *Brier score* as calibration measure and the *C-index* to evaluate the discriminative ability of the models, both of them adjusted for survival outcomes. Furthermore, for both CRT and CLL datasets, as well as for the simulated datasets, we evaluate all methods described in section 2.2 using the Brier score and the C-index measures at both 1 and 5 years follow-up. Calculations were carried out in R using the packages *pec* (Mogensen, Ishwaran, and Gerds 2012) for the Brier score and *timeROC* (Blanche, Dartigues, and Jacqmin-Gadda 2013) for the C-index. We evaluated each method for both $M = 10$ and $M = 100$. For the CRT and CLL data, the above measures are also based on 10 applications of each methodology on the single data and the resulting final measures are then averaged across replications. While for the simulated data, Brier score and C-index statistics are calculated for each simulation scenario, averaging across replications and the 100 simulations from that scenario.

Brier score

The Brier score measures the predictive performance by measuring the "predictor error". There are several versions of this statistic, but the most popular one has been introduced by Graf et al. (1999). In a survival context, prediction is not to be understood as an absolute prediction whether an individual will survive beyond t_0 or not, but as a *probabilistic prediction* quantifying the probability of survival beyond t_0 (Van Houwelingen and Putter 2011).

Let $\widehat{S}(t_0|x)$ be the predicted survival probability for an individual beyond t_0 given the predictor x and let $y = I(T > t_0)$ be the actual observation (ignoring censoring). Brier score is then defined as follows:

$$BS(y, \widehat{S}(t_0|x)) = (y - \widehat{S}(t_0|x))^2. \tag{3.4}$$

With respect to a new observation y_{new} under the true model $S(t_0|x)$, the expected value of this measure can be seen as the sum of two components: the "true variation" and the "model error" due to misspecification of the model. It can in fact be written as:

$$E[BS(y_{new}, \widehat{S}(t_0|x))] = S(t_0|x)(1 - S(t_0|x)) + (S(t_0|x) - \widehat{S}(t_0|x))^2. \tag{3.5}$$

In survival context, censoring has to be considered. Graf et al. (1999) suggested a weighted derivation of the Brier score, based on the assumption that the censoring mechanism is independent of the covariates, called *Inverse Probability of Censoring Weighting* (IPCW). To compensate for the loss of information due to censoring, the individual contributions have to be weighted. For each patient we observe $T_i = min(\tilde{T}_i, C_i)$ and $\delta_i = I(\tilde{T}_i \leq C_i)$ where $\tilde{T}_i$ is the time to the event of interest and C_i the censoring time. For a fixed time point t_0, the weighted Brier score equation for the entire model is then:

$$BS(t_0) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \widehat{S}_i(t_0|x_i))^2 \; w_i \tag{3.6}$$

where the weight function w_i is defined as:

$$w_i = \begin{cases} 0, & \text{if } T_i < t_0 \text{ and } d_i = 0 \\ \dfrac{1}{\widehat{G}(t_0)}, & \text{if } T_i > t_0 \\ \dfrac{1}{\widehat{G}(T_i)}, & \text{if } T_i < t_0 \text{ and } d_i = 1 \end{cases} \tag{3.7}$$

where $\widehat{G}(t)$ is the estimate of the censoring distribution $G(t) = P(C > t)$.

The Brier score can take values between 0 and 1, where 0 denotes a

model with no predictor error. When comparing two different models, the best model is the one with the smallest model error.

C-index

Harrell's (1996) C-index is the most commonly used performance measure to indicate the discriminative ability of generalized linear regression models. For a binary outcome, it corresponds to the area under the receiver operating characteristic (ROC) curve, also called AUC, which plots the true positive rate against the false positive rate for consecutive cut-offs for the probability of an outcome. This measure can be seen as a rank-order statistic for predictions against true outcomes and can also be extended to censored data ignoring the pairs that cannot be ordered (Steyerberg et al. 2010). C-index is then the fraction of pairs of observations for which the order of survival times and model predictions are correctly ordered among all pairs that can be ordered. In a survival context, a pair (i, j) is considered usable if both individuals are non-censored or if at least the individual with the shortest time has an event. The pair is then considered concordant if the one who dies first has the largest x-value (Van Houwelingen and Putter 2011).

A version of the C-index corrected for censoring can be obtained by IPCW, as it has been shown in Uno et al. (2007). Assuming that random censoring time C is independent of predictors, the C-index can be estimated as follows:

$$\widehat{C}(t) = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} I(T_i \leq t)\ I(T_j > t)\ I(X_i > X_j)\ \frac{\delta_i}{\hat{S}_C(Z_i)\ \hat{S}_C(t)}}{n^2\ \hat{S}_{KM}(t)[1 - \hat{S}_{KM}(t)]} \tag{3.8}$$

where $\hat{S}_C(.)$ is the Kaplan-Meier estimator of the survival function of the censoring time C, while $\hat{S}_{KM}(t)$ is the Kaplan-Meier estimator of $P(T > t)$. Furthermore, T_i and T_j are the observed survival times for individuals i and j respectively, δ_i is the status indicator for the individual i and finally, X_i and X_j are the marker values for individuals i and j respectively. X can be

a single marker or several markers combined into a predictive model, we also assume that larger values of X are associated with greater risks.

The C-index can take values between 0 and 1, with 1 denoting a perfect discrimination and 0.5 corresponding to a model with no predictive ability, as a random guess model. When comparing two different models, the best model is the one with the higher discrimination ability.

3.3.2 Summary measures for the CRT and CLL data

In addition to the above calibration (Brier score) and discrimination (C-index) measures, we also investigated the variation of predictions for individual patients in the data across several repeated calibrations of the methods. The objective of the latter is to investigate the sensitivity of prediction at the patient-level due to imputation variation.

Let $\widehat{S}_{i,r}(t)$ be the predicted survival probability at time t for the patient i, while r denotes the replicate (calibration) of the model. We then first calculate the mean of the final predictions across replications, $\overline{S_i(t)}$, for each patient and then the deviations $D_{i,r}(t) = \widehat{S}(t)_{i,r} - \overline{S_i(t)}$. While the latter are heteroscedastic, their variation will be approximately constant for patients with $0.2 \leq \overline{S_i(t)} \leq 0.8$, we therefore discard all the deviations corresponding to patients with $\overline{S_i(t)} < 0.2$ or $\overline{S_i(t)} > 0.8$ and compute the 90^{th} and 10^{th} percentiles, $Q_{0.90}$ and $Q_{0.10}$, across all the remaining deviations $D_{i,r}(t)$.
We then report

$$R(t) = Q_{0.90} - Q_{0.10} \tag{3.9}$$

as a measure of spread of predictive probabilities induced by imputation variation at the probability scale. We calculate this measure for $M = 10, 100$ and 1000, we set the number of replicates to $r = 10$ and investigate measures for $t = 1$ and $t = 5$ years.

3.3.3 Summary measures for simulated data

For the simulated data, in addition to the Brier score and C-index measure, we could also investigate variance as well as bias, as we have the true survival fractions $S_{i,TRUE}(t)$ available for each simulated individual i at any time t, based on the assumed simulation model. We therefore define a measure to asses variation of predictions based on percentiles and a measure of bias.

Let $\widehat{S}(t)_{i,r,s}$ be the fitted survival probability from any approach at time t for the individual i within the simulated dataset s and for the r^{th} replicate analysis. We then first calculate the mean of the final predictions across replications, $\overline{S_{i,s}(t)}$, for each individual i within the s^{th} simulation. We then compute the deviations $H_{i,r,s}(t) = \widehat{S}(t)_{i,r,s} - \overline{S_{i,s}(t)}$. In analogy to the above description of the measure D, we now calculate the 90^{th} and 10^{th} percentiles, $Q_{0.90}$ and $Q_{0.10}$, across all the deviations corresponding to individuals with $0.2 \leq S_{i,TRUE}(t) \leq 0.8$ within each s^{th} simulated dataset. We then define $V_s(t) = Q_{0.90} - Q_{0.10}$ as a measure of variation for the s^{th} simulation and report as final summary measure of variance

$$V(t) = \overline{V_s(t)} \tag{3.10}$$

that is the mean across simulations of these measures.

To define a measure of bias, we proceed by first calculating the average values of predictions $\overline{S_{i,s}(t)}$ across replicates and subsequently computing the deviations within the s^{th} simulation from the true survival fraction:

$$B_{i,s}(t) = \overline{S_{i,s}(t)} - S_{i,TRUE}(t). \tag{3.11}$$

We report the summary measure $B(t)$ defined as the mean across i and across all the simulations s of all the $B_{i,s}(t)$ measures of those individuals with $0.2 \leq S_{i,TRUE}(t) \leq 0.8$:

$$B(t) = \overline{B_{i,s}(t)}. \tag{3.12}$$

Chapter 4

Results

This chapter discusses the results from application of the proposed methodologies on the two real datasets (section 4.1) and simulations (section 4.2). All the measures described in section 3.3 are reported, but only the most interesting tables and graphs are shown in the chapter, other materials can be found in the appendix.

4.1 CRT and CLL data

4.1.1 Calibration and discrimination

Tables 4.1 and 4.2 report results on the evaluation of calibration and discrimination performance of the proposed methods on the CRT and CLL data. These tables show the Brier score and the C-index statistics based on 10 multiple imputations and on 10-folds cross-validation. Approaches have been applied 10 times to the data and, as final measures, we took the average values of these statistics across repetitions. Results are tabulated for both 1 and 5 years follow-up. The calculation has been done on the full set of observations in column "*All obs.*" and repeated for only those observations containing missing values, in column "*Missing*", and for the completely observed records, in column "*Fully obs.*" In addition we also show the results

based on 10-fold cross-validation in the complete case, that means considering only completely observed records, which does not require imputation. In this case, predictions for each left-out fold have been obtained after a single estimation of the Cox model in the corresponding calibration sets. The complete case analysis can be seen as reference performance.

CRT		**Brier Score**			**C-index**		
M=10		Missing	Fully obs.	All obs.	Missing	Fully obs.	All obs.
1 yr	A 1	0.0711	0.0628	0.0670	0.8230	0.7111	0.7703
	A 2A	0.0712	0.0629	0.0671	0.8146	0.7113	0.7654
	A 2B	0.0713	0.0629	0.0671	0.8146	0.7113	0.7654
	N 1	0.0702	0.0627	0.0665	0.8395	0.7125	0.7794
	N 2A	0.0704	0.0627	0.0666	0.8209	0.7021	0.7647
	N 2B	0.0704	0.0627	0.0666	0.8209	0.7021	0.7647
Compl. case			0.0629			0.7353	
5 yrs	A 1	0.2104	0.1761	0.1904	0.6863	0.7693	0.7371
	A 2A	0.2121	0.1767	0.1914	0.6747	0.7573	0.7243
	A 2B	0.2120	0.1767	0.1914	0.6747	0.7573	0.7243
	N 1	0.2043	0.1757	0.1876	0.7032	0.7714	0.7450
	N 2A	0.2048	0.1765	0.1882	0.6988	0.7548	0.7327
	N 2B	0.2048	0.1764	0.1882	0.6988	0.7548	0.7327
Compl. case			0.1780			0.7346	

Table 4.1: Brier score and C-index statistics for the CRT data based on 10 multiple imputations and on 10-fold cross-validation. We report the average values of these statistics across 10 replicates of the approaches. Results are shown for both 1 and 5 years of follow-up, for those observations with missing values ("Missing"), for the completely observed records ("Fully obs.") and for the full set of observations ("All obs."). Results for the complete case analysis are shown as well.

Tables show that there are no relevant differences in the performance of the proposed methods, neither at 1 nor at 5 years of follow-up. We might see a small difference in the summary measures computed on those observations with missing values, where naïve approaches seem to have slightly higher C-index values. Results for Brier score and C-index computed on the whole set of data are a mixture between the corresponding results on the missing and completely observed cases. Brier score is slightly higher for those obser-

vations with missing values, as a consequence of the increased uncertainty in prediction. C-index score calculated on observations with missing values is also slightly larger, which seems counter-intuitive and it might be due to a misspecification of the imputation methods, which underestimates the variation in imputed values. Finally, we can also note that the Brier score for the complete case analysis for the CRT data closely matches the results obtained for the fully observed data across approaches. As has been discussed before, the missingness pattern in the CRT dataset could be seen as a MCAR scenario and thus, complete case analysis could be actually done.

For $M = 100$ we obtained the same results as for $M = 10$. Since numbers are almost indistinguishable, results for $M = 100$ are shown in the appendix, in table 9 for the CRT dataset and in table 10 for the CLL.

CLL		**Brier Score**			**C-index**		
M=10		Missing	Fully obs.	All obs.	Missing	Fully obs.	All obs.
1 yr	A 1	0.2019	0.1815	0.1886	0.6698	0.5982	0.6260
	A 2A	0.2026	0.1816	0.1889	0.6667	0.6006	0.6256
	A 2B	0.2026	0.1816	0.1889	0.6667	0.6006	0.6256
	N 1	0.1972	0.1808	0.1864	0.6952	0.6011	0.6374
	N 2A	0.1979	0.1813	0.1870	0.6883	0.6006	0.6341
	N 2B	0.1980	0.1813	0.1870	0.6883	0.6006	0.6341
Compl. case			0.1837			0.5950	
5 yrs	A 1	0.2362	0.2416	0.2404	0.6462	0.6109	0.6235
	A 2A	0.2362	0.2411	0.2401	0.6451	0.6137	0.6248
	A 2B	0.2363	0.2411	0.2401	0.6451	0.6137	0.6248
	N 1	0.2293	0.2396	0.2362	0.6750	0.6166	0.6393
	N 2A	0.2307	0.2407	0.2375	0.6656	0.6168	0.6353
	N 2B	0.2308	0.2407	0.2375	0.6656	0.6168	0.6353
Compl. case			0.2560			0.6046	

Table 4.2: Brier score and C-index statistics for the CLL data based on 10 multiple imputations and on 10-fold cross-validation. We report the average values of these statistics across 10 replicates of the approaches. Results are shown for both 1 and 5 years of follow-up, for those observations with missing values ("Missing"), for the completely observed records ("Fully obs.") and for the full set of observations ("All obs."). Results for the complete case analysis are shown as well.

4.1.2 Variation of the individual predictions

In addition to the above measures to asses classification performance of the proposed methods, we also study the variability of predictions at individual-level to account for variation due to imputations. We therefore investigate the variation in individual predictions within methods, after a single calibration of any approach, and after several replications of the analysis. That means that we study the variation in predictions at two different levels: first of all, looking at how much individual predictions $\widehat{S}_{i,m}(t)$ vary before averaging them to get the final predictions $\widehat{S}_i(t)$ and second, we also investigate how much the latter vary, this time across different replicates of the same method.

Within a single multiple-imputation based calibration

We start investigating variation of $\widehat{S}_{i,m}(t)$. Figures 4.1 and 4.2 plot the individual survival predictions $\widehat{S}_{i,m}(t)$ versus the final ones $\widehat{S}_i(t)$ at 5 years for all the approaches and using $M = 1000$ imputations. Note that the final predictions correspond to the mean of the individual survival predictions within individuals. We also distinguish between predictions corresponding to fully observed records and to those with missing values, respectively marked with black and red dots. For fully observed records, variation of predictions at individual level is zero for approaches $2A$ and $2B$, by design. These figures show that the variation in individual predictions is very large, especially for those predictions with averaged value around 0.5 and for those observations with missing values, as we expected.

These figures are further summarized in table 4.3. In this case, we calculate the same statistic defined in section 3.3.2 to express variation, but in this case within a single application of the approaches. Which means that we calculate the distance between the 90^{th} and 10^{th} percentiles of the deviations $\widehat{S}_{i,m}(t) - \widehat{S}_i(t)$ for those observations with average survival rate $\widehat{S}_i(t)$ between

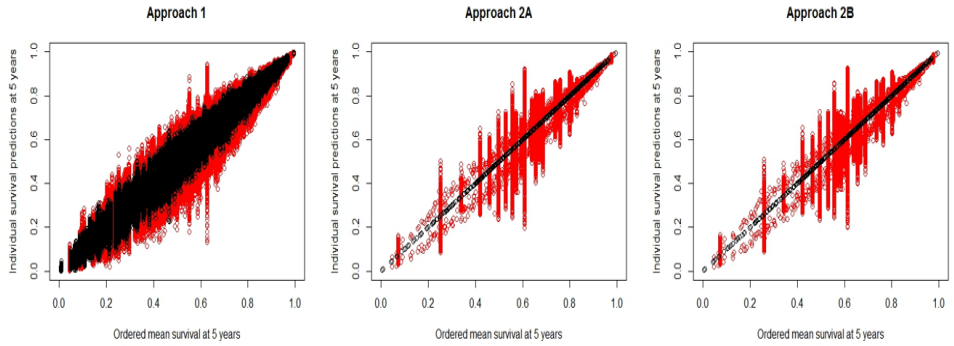

Figure 4.1: Survival prediction $\widehat{S}_{i,m}(t)$ at 5 years for the CRT data within approaches 1, $2A$ and $2B$ versus the mean (final) predictions $\widehat{S}_i(t)$. Results are shown for 1000 multiple imputations. Red dots show predictions for individuals with missing values in the covariates, while black dots denote predictions based on fully observed records.

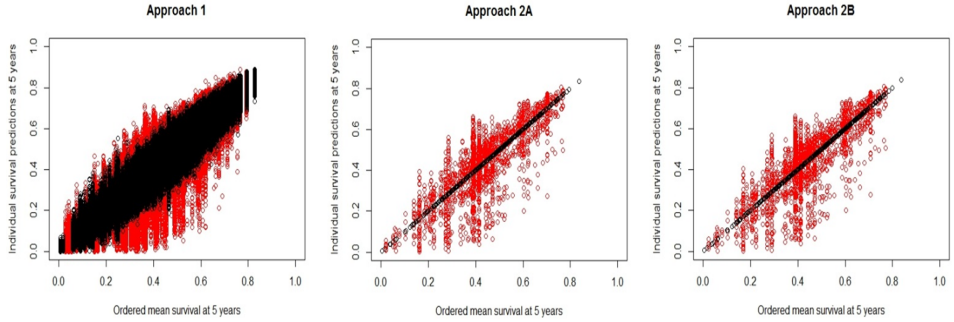

Figure 4.2: Survival prediction $\widehat{S}_{i,m}(t)$ at 5 years for the CLL data within approaches 1, $2A$ and $2B$ versus the mean (final) predictions $\widehat{S}_i(t)$. Results are shown for 1000 multiple imputations. Red dots show predictions for individuals with missing values in the covariates, while black dots denote predictions based on fully observed records.

0.2 and 0.8. Results are shown for both 1 year and 5 years follow-up and are also separately calculated for those individuals with missing values, as well as for those with fully observed records.

M=1000		CRT		CLL	
		Missing	Fully obs.	Missing	Fully obs.
1 year	A 1	0.117	0.061	0.138	0.075
	A 2A	0.103	0	0.115	0
	A 2B	0.102	0	0.115	0
	N 1	0.115	0.060	0.135	0.071
	N 2A	0.103	0	0.113	0
	N 2B	0.103	0	0.113	0
5 years	A 1	0.153	0.066	0.173	0.099
	A 2A	0.136	0	0.144	0
	A 2B	0.135	0	0.144	0
	N 1	0.151	0.065	0.169	0.094
	N 2A	0.134	0	0.141	0
	N 2B	0.135	0	0.141	0

Table 4.3: Variation between predictions within a single calibration of any approach, using 1000 imputations. Results are shown for both CRT and CLL data at 1 and 5 years follow-up and distinguishing between fully observed records and those with missing values. Refer to section 3.3.2 for the precise definition of the measure.

Both for the CRT and CLL data, the difference between the deviation of the individual predicted survival probabilities at the 90^{th} and 10^{th} percentiles can be larger then 10% in case of records with missing values, both at 1 and 5 years of follow-up. For the fully observed records, numbers are smaller and, for example, for approach 1 at 1 year, variation is around 6% for the CRT data and around 7% for the CLL data.

Within a single method calibration, approach 1 has higher variation of the individual predictions than the two approaches 2, based on Rubin's rules. This depends on the fact that approach $2A$ and $2B$ are based on the direct averaging of the model coefficients and some of the between-models variation is removed before the application to the individual observations. On the other

hand, approach 1 approximates the posterior predictive density and this leads to a higher variation between the individual predictions.
Fully observed records have in general a smaller variation and if we now look at the difference between 1 and 5 years of follow-up, for the CLL data, predictions after 5 years have always a larger variation, as would be expected. The same behaviour can be observed for the approach 1 for the CRT data, even if the difference is smaller.

Finally, note that variation observed for approach 1 may also be interpreted as the variation between predictions which would be observed if we applied single-imputation and then repeated the analysis. Table 4.3 shows that variation is very high and single-imputation should then be avoided in the predictive calibration of prognostic rules. It is also clear that a larger number of imputations is preferable.

Between replicates of the imputed-based approaches

In this section we investigate the predictive variation at the individual level, due to the variation in multiple imputation. We then recalibrate each approach 10 times and study the variation in the individual final predictions ($\widehat{S}_{i,r}(t)$), which are the average of the imputation-based individual survival predictions($\widehat{S}_{i,r,m}(t)$). Results for the $R(t)$ statistic, as discussed in section 3.3.2, are shown in table 4.4 for the CRT data and in table 4.5 for the CLL data. Results are presented for $M = 10, 100$ and 1000 imputations, at 1 and 5 years of follow-up and distinguishing between records with missing values and those completely observed.

First of all, we can notice that approach 1 has lower between-replicates variation in prediction if compared with approaches 2A and 2B. This result is true independently of the number of multiple imputations considered, for the prediction of both fully observed and with partially missing records set of data and also at both years of follow-up. Furthermore, the difference in the values of $R(t)$, comparing between approach 1 versus 2A and 2B, increases

CRT		M=10		M=100		M=1000	
		Missing	Fully obs.	Missing	Fully obs.	Missing	Fully obs.
1 yr	A 1	0.048	0.023	0.019	0.007	0.004	0.002
	A 2A	0.080	0.050	0.050	0.050	0.043	0.049
	A 2B	0.076	0.046	0.051	0.045	0.042	0.048
	N 1	0.054	0.023	0.015	0.007	0.005	0.003
	N 2A	0.085	0.048	0.053	0.051	0.051	0.049
	N 2B	0.078	0.044	0.047	0.047	0.049	0.050
5 yrs	A 1	0.059	0.024	0.017	0.008	0.005	0.003
	A 2A	0.097	0.053	0.057	0.049	0.049	0.050
	A 2B	0.095	0.051	0.056	0.055	0.050	0.048
	N 1	0.063	0.024	0.018	0.008	0.006	0.004
	N 2A	0.092	0.050	0.053	0.047	0.066	0.066
	N 2B	0.094	0.049	0.055	0.048	0.064	0.067

Table 4.4: Variation measure in prediction between replicate analysis ($R(t)$) using the same approach for either $M = 10, 100$ or 1000, for CRT data at 1 and 5 years follow-up and distinguishing between fully observed records and those with missing values. Refer to section 3.3.2 for the precise definition of the measure.

CLL		M=10		M=100		M=1000	
		Missing	Fully obs.	Missing	Fully obs.	Missing	Fully obs.
1 yr	A 1	0.051	0.026	0.016	0.009	0.005	0.003
	A 2A	0.073	0.057	0.059	0.054	0.052	0.052
	A 2B	0.076	0.057	0.056	0.055	0.052	0.054
	N 1	0.046	0.027	0.016	0.009	0.005	0.003
	N 2A	0.069	0.050	0.056	0.053	0.051	0.050
	N 2B	0.067	0.051	0.054	0.051	0.049	0.050
5 yrs	A 1	0.061	0.036	0.002	0.012	0.006	0.004
	A 2A	0.095	0.076	0.077	0.073	0.068	0.068
	A 2B	0.093	0.076	0.071	0.072	0.068	0.072
	N 1	0.057	0.035	0.002	0.011	0.006	0.004
	N 2A	0.084	0.067	0.071	0.070	0.066	0.066
	N 2B	0.085	0.069	0.072	0.070	0.064	0.067

Table 4.5: Variation measure in prediction between replicate analysis ($R(t)$) using the same approach for either $M = 10, 100$ or 1000, for CLL data at 1 and 5 years follow-up and distinguishing between fully observed records and those with missing values. Refer to section 3.3.2 for the precise definition of the measure.

with the rise in the number of imputations and the biggest gap is observed for the complete records at $M = 1000$.
Second, the number of imputations seems to have an important role in reducing variation. Results show that 10 imputations are not enough for predictive calibration in the presence of missing data and that a substantial improvement can be made by increasing this number at least to 100. In general, in fact, variation reduces when increasing the number of imputations and this is especially true for approach 1. For the latter, indeed, reduction in predictive variation is still achieved when increasing the number of imputation from 100 to 1000 and this also leads to a predictive variation below 1%, which is highly desirable for practical use in any medical application. While, in contrast, for approaches 2A and 2B, reduction in predictive variation does not considerably improve when increasing number of imputations beyond 100 and moreover, the predictive variation measure $R(t)$, for both CRT and CLL data, is stuck above 4.5% for both $M = 100$ and $M = 1000$. These conclusions can be drawn for predictions based on the fully observed records as well for those based on records with missing values. Furthermore, we can see that the difference in variation between completely observed records and those with some missing values is higher for approach 1, while approach 2A and 2B do not present a big gap between the two groups, especially for M bigger than 10.

The above argumentations can be easily better realized by looking at figures 4.3 and 4.4, respectively for CRT and CLL data, which show the same results presented in tables 4.4 and 4.5 for the three approaches. The red lines correspond to results for partially observed records, while the black ones are for fully observed records. Finally, the solid lines are for results at 1 year of follow-up, while the dashed lines correspond to results at 5 years.

From these graphics, the difference between approach 1 and the other two is even more understandable. As we can easily see, for $M = 10$, approach 1 achieves a precision that is not matched for approaches 2A and 2B, not even with $M = 1000$. We can also notice an effective reduction in variation when

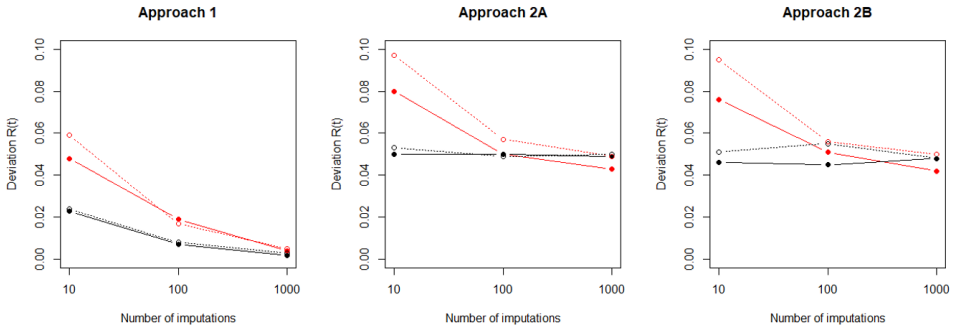

Figure 4.3: Deviation of predictions $R(t)$ across replicate calibration for approaches 1, $2A$ and $2B$ versus the number of imputations for the CRT data. Results are shown at 1 and 5 years follow-up, respectively denoted by solid and dashed lines. Red lines correspond to results for predictions with missing values, while black lines are for fully observed records.

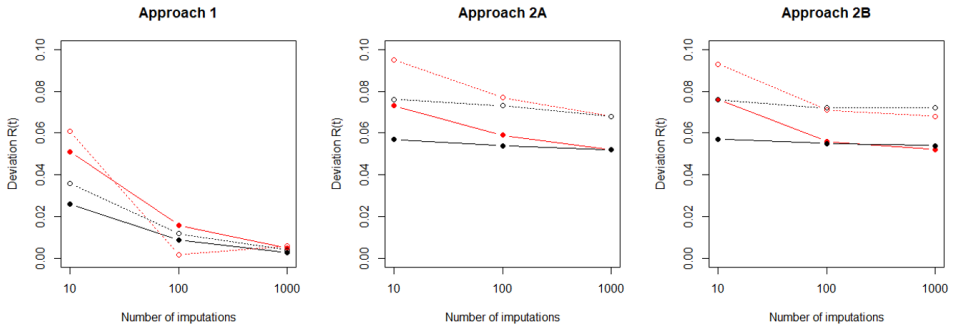

Figure 4.4: Deviation of predictions $R(t)$ across replicate calibration for approaches 1, $2A$ and $2B$ versus the number of imputations for the CLL data. Results are shown at 1 and 5 years follow-up, respectively denoted by solid and dashed lines. Red lines correspond to results for predictions with missing values, while black lines are for fully observed records.

increasing the number of imputations for approach 1, behaviour that we do not see for the other approaches, which seem cannot profit from the increase of number of imputations. The fact that the above results are common features of both the CRT and CLL data finally suggests that they might actually represent general properties of the methods and are not data-specific.

Furthermore, we can also compare these results with the ones in table 4.3, which can be seen as the single-imputation scenario, with $M = 1$. We can then see that, especially for approach 1, the reduction in predictive variation is achieved by using multiple imputation instead of one.

Finally, regarding the naïve implementation of the proposed approaches, we note that their predictive variation is not much different from the one of the proposed methods.

4.2 Simulated data

This section presents the results of the simulation study, as described in section 3.2. Analyses have been carried out for both MCAR and MAR scenarios, with $M = 10$ and 100. Summary measures, as described in section 3.3, have been computed at both 1 and 5 years of follow-up and separately for fully observed records and for those with missing values. For scenarios 1 and 3, measures of variation and bias at 1 year of follow-up are not shown, since these measures are defined for all those observations with "true" survival outcome between 0.2 and 0.8, but because of the way simulations have been set up, there were not enough observations within this range of interest. For comparison, we have also performed analysis for a single calibration of the Cox model with cross-validatory assessment, considering first, the complete cases only, and second, the *original* simulated datasets, before the introduction of the missing values. Results of the analyses on the original datasets can be seen as benchmarks, since they represent the optimal values, those we would expect if we had no lack of information.

In the next sections we present tables with the most interesting results,

complete tables with all the results of the simulation study can be found in the appendix in section .2.

4.2.1 Calibration and discrimination

Table 4.6 shows the simulation results for the Brier score and the C-index at 1 year of follow-up and using $M = 10$ multiple imputations on simulated datasets with MCAR values. Results of the analyses on the complete cases and on the original datasets are shown as well.

Simulations show that the Brier score, as we would expect, presents in general higher values for those records with missing values. If we compare the performance of our proposed approaches with the corresponding naïve implementations, we can also note that naïve approaches present lower values of this index and this fact may be seen as in favour of the naïve implementation. However, if we compare these numbers with the benchmark values, we can see that naïve approaches give actually too optimistic results, especially for scenarios 2 and 4. The analysis on the original data, indeed, represents what we would expect if we had no missing data and it then can be seen as the best achievement we can have for those datasets.

On the other hand, C-index generally shows lower values for those records with missing values, as we would expect, since they are characterized by more uncertainty. As for the Brier score, if we compare the proposed approaches with the corresponding naïve implementations for those observations with missing values, we would say that the latter present better performance than the proposed ones. For example, in scenario 2, C-index for approach 1 is around 59% and reaches 76% for the corresponding naïve implementation. Also in this case though, this hypothesis is contradicted by the comparison with the result obtained from the analysis in the original datasets, which gives a C-index around 69%. Discrimination performance of the naïve implementations seems indeed to be better than the one we would have in the best scenario we could have, that is the one with no missing values. Once

MCAR		**Brier Score**		**C-index**	
M=10, 1 year		Missing	Fully obs.	Missing	Fully obs.
	A 1	0.0780	0.0803	0.5911	0.5867
	A 2A	0.0780	0.0803	0.5911	0.5862
Scen 1	A 2B	0.0780	0.0803	0.5911	0.5862
	N 1	0.0779	0.0803	0.5978	0.5870
	N 2A	0.0779	0.0803	0.5969	0.5864
	N 2B	0.0779	0.0803	0.5969	0.5864
Complete Case		0.0803		0.5859	
Original Dataset		0.0801		0.5884	
	A 1	0.0928	0.0871	0.5876	0.6880
	A 2A	0.0929	0.0872	0.5862	0.6878
Scen 2	A 2B	0.0929	0.0872	0.5862	0.6878
	N 1	0.0879	0.0871	0.7566	0.6882
	N 2A	0.0879	0.0871	0.7562	0.6880
	N 2B	0.0879	0.0871	0.7562	0.6880
Complete Case		0.0872		0.6878	
Original Dataset		0.0873		0.6876	
	A 1	0.0793	0.0801	0.5872	0.5829
	A 2A	0.0793	0.0801	0.5870	0.5827
Scen 3	A 2B	0.0793	0.0801	0.5870	0.5827
	N 1	0.0791	0.0801	0.5984	0.5849
	N 2A	0.0791	0.0801	0.5966	0.5845
	N 2B	0.0791	0.0801	0.5966	0.5845
Complete Case		0.0803		0.5773	
Original Dataset		0.0797		0.5873	
	A 1	0.0906	0.0865	0.5934	0.6845
	A 2A	0.0907	0.0865	0.5927	0.6843
Scen 4	A 2B	0.0907	0.0865	0.5927	0.6843
	N 1	0.0859	0.0864	0.7613	0.6855
	N 2A	0.0859	0.0864	0.7600	0.6853
	N 2B	0.0859	0.0864	0.7600	0.6853
Complete Case		0.0867		0.6834	
Original Dataset		0.0866		0.6881	

Table 4.6: Results of the simulation study for the Brier score and the C-index at 1 year of follow-up and using $M = 10$. Results refer to simulated datasets with MCAR values and are reported separately for fully observed records and for those with missing values. Results of the analyses on the complete cases and on the original datasets are shown as well.

again, this affirmation is particularly true for scenarios 2 and 4. As well as for the Brier score before, this might be due to the strong association between the variable with missing values and the outcome and it clearly shows how naïve implementations in predictive calibration should be avoided.

Increasing M to 100 or applying the proposed approaches to MAR data leads to the same conclusions and values are almost indistinguishable from those in table 4.6. For this reason, results of the analyses carried out with $M = 100$ and on MAR data are reported in the appendix, along with the results at 5 years of follow-up.

For both statistics, not much difference has been observed between analysis on the original datasets and on the complete cases only, even if we can see that complete case analysis has slightly worse performance than the other one. We would have expected to gain something more from doing multiple imputation, at least for the MAR scenarios. These results may then be due to the way simulations have been set up and especially to the fact that predictors are not strongly correlated to the outcome. Multiple imputation is shown to be preferable than complete cases analysis to avoid bias in the parameters estimation, but since this work is focused on predictions, further research to better understand this phenomenon is suggested.

4.2.2 Variation and bias

Tables 4.7 and 4.8 show the simulation results for variation and bias measures as described in section 3.3.3. Results refer to MCAR scenarios at 5 years of follow-up (more informative than those at 1 year, reported in the appendix) and using respectively $M = 10$ and 100 multiple imputations. Values are reported separately for fully observed records and for those with missing values. Tables with the summary measures for datasets with MAR values can be found in the appendix, since numbers are very similar to those of MCAR scenarios and lead to the same conclusions.

MCAR		Variation		Bias	
M=10, 5 years		Missing	Fully obs.	Missing	Fully obs.
	A 1	0.0191	0.0088	-0.0000	-0.0006
	A 2A	0.0293	0.0257	-0.0001	-0.0006
Scen 1	A 2B	0.0293	0.0258	-0.0001	-0.0006
	N 1	0.0190	0.0086	-0.0001	-0.0006
	N 2A	0.0292	0.0254	-0.0001	-0.0006
	N 2B	0.0292	0.0254	-0.0001	-0.0006
	A 1	0.1186	0.0090	0.0182	-0.0002
	A 2A	0.1211	0.0256	0.0182	-0.0002
Scen 2	A 2B	0.1211	0.0261	0.0182	-0.0002
	N 1	0.1117	0.0089	0.0172	-0.0002
	N 2A	0.1136	0.0248	0.0171	-0.0002
	N 2B	0.1137	0.0254	0.0172	-0.0001
	A 1	0.0249	0.0127	0.0033	0.0033
	A 2A	0.0319	0.0287	0.0035	0.0035
Scen 3	A 2B	0.0320	0.0289	0.0033	0.0033
	N 1	0.0242	0.0122	0.0033	0.0033
	N 2A	0.0314	0.0260	0.0035	0.0035
	N 2B	0.0315	0.0262	0.0033	0.0033
	A 1	0.1166	0.0143	0.0204	0.0036
	A 2A	0.1183	0.0297	0.0205	0.0036
Scen 4	A 2B	0.1190	0.0329	0.0206	0.0036
	N 1	0.1085	0.0133	0.0184	0.0038
	N 2A	0.1103	0.0258	0.0185	0.0039
	N 2B	0.1110	0.0293	0.0183	0.0037

Table 4.7: Results of the simulation study for the variation and bias statistics described in section 3.3.3 at 5 years of follow-up and using $M = 10$. Results refer to simulated datasets with MCAR values and are reported separately for fully observed records and for those with missing values.

MCAR M=100, 5 years		Variation Missing	Variation Fully obs.	Bias Missing	Bias Fully obs.
Scen 1	A 1	0.0060	0.0028	-0.0000	-0.0006
	A 2A	0.0234	0.0255	-0.0000	-0.0006
	A 2B	0.0234	0.0255	-0.0001	-0.0006
	N 1	0.0060	0.0027	-0.0000	-0.0006
	N 2A	0.0233	0.0251	-0.0000	-0.0006
	N 2B	0.0233	0.0251	-0.0000	-0.0006
Scen 2	A 1	0.0381	0.0029	0.0209	0.0010
	A 2A	0.0453	0.0253	0.0209	0.0010
	A 2B	0.0455	0.0259	0.0209	0.0010
	N 1	0.0358	0.0028	0.0186	0.0010
	N 2A	0.0420	0.0247	0.0187	0.0010
	N 2B	0.0423	0.0254	0.0187	0.0011
Scen 3	A 1	0.0080	0.0040	-0.0019	-0.0023
	A 2A	0.0239	0.0273	-0.0017	-0.0021
	A 2B	0.0241	0.0275	-0.0019	-0.0022
	N 1	0.0077	0.0038	-0.0019	-0.0023
	N 2A	0.0237	0.0241	-0.0017	-0.0021
	N 2B	0.0239	0.0243	-0.0020	-0.0024
Scen 4	A 1	0.0362	0.0045	0.0214	0.0042
	A 2A	0.0440	0.0279	0.0214	0.0043
	A 2B	0.0459	0.0315	0.0214	0.0043
	N 1	0.0337	0.0042	0.0203	0.0042
	N 2A	0.0400	0.0235	0.0204	0.0042
	N 2B	0.0422	0.0276	0.0207	0.0046

Table 4.8: Results of the simulation study for the variation and bias statistics described in section 3.3.3 at 5 years of follow-up and using $M = 100$. Results refer to simulated datasets with MCAR values and are reported separately for fully observed records and for those with missing values.

Variation in predictive probabilities is systematically smaller for approach 1 than for approaches 2A and 2B and this can especially be seen for scenarios 1 and 3. As we expected, observations with missing values have generally larger variation of predictions than those with fully observed records. Furthermore, for the latter group, the difference in variation between approach

1 and the other two is bigger than for the ones with missing values. For fully observed observations, indeed, variation of approaches 2A and 2B is always at least two times the one of approach 1. Predictions at 5 years of follow-up are generally more variable than those at 1 year and the highest values are registered in scenarios 2 and 4, with $M = 10$, where variation presents values around 12%. These two scenarios are characterized by a strong association between the predictor with missing values and the outcome, the effect of multiple imputation is then amplified in predictions by leading to more variable results. By increasing the number of imputations from 10 to 100, this variability is reduced in all scenarios and the biggest gain is observable for approach 1, as we have already seen for the application in the real data. On the other hand, reduction in variation observed for approaches 2A and 2B is not very considerable. Furthermore, no relevant difference in variation can be observed between the proposed methods and their naïve implementations.

Bias in predictive probabilities is small in all scenarios, but especially for scenarios 1 and 3, and can be both positive and negative. Bias for fully observed records presents almost the same values across all scenarios. The highest values can be observed in scenarios 2 and 4, for incomplete observations, but in any case, bias is still always lower than 0.03. This seems to confirm that including the survival information in the form of the Nelson-Aalen estimate of the cumulative hazard, together with the status information, gives satisfactory results within this context of simulation scenarios where the goal is getting predictions and $\boldsymbol{\beta}$ is not investigated. No relevant improvement in bias reduction has been observed when increasing the number of imputations from 10 to 100. Regarding naïve approaches, it might seem that for scenarios 2 and 4 they lead to a slightly lower bias than our proposed methods do, but this can actually be seen as a too "optimistic" results, connected with what has already been said for the Brier score. The latter, in fact, can be seen as the sum of a measure of variance and model error and the fact that bias presents smaller values may then be related to the smaller values also observed for the Brier score.

Discussion

This work aimed to compare two approaches to the calibration of prediction models when multiple imputation is used to deal with missing data and cross-validatory assessment is required to asses models performance. We have first defined two general approaches to the combination of cross-validation and multiple imputation, further specified for the application on survival data. The first approach aims to calibrate the predictive density by averaging predictions of multiple models, which have been separately estimated on distinct imputed datasets. The second approach is based on the application of the so-called Rubin's rules to combine the model parameters across multiple imputations. Once we have the pooled set of parameters, we can use it to get final predictions. In addition, two versions of this second approach have been implemented to take into account the particular features of the Cox regression model. When the final aim is getting predictions in a Cox regression context, in fact, the substantive model is a combination of the usual regression parameters, say $\boldsymbol{\beta}$, and the cumulative baseline hazard, which has to be summarized as well. Hence, we proposed two different ways to get the final estimate of this latter measure. The first one is a straightforward application of the Rubin's rules, which means that the pooled cumulative baseline hazard is obtained by averaging the single estimates across multiple imputations. On the other hand, the second one gets the estimate of the quantity of interest by plugging the pooled set of parameters $\boldsymbol{\beta}$ into the Breslow estimator. These approaches are respectively called 2A and 2B. All the above mentioned methods have been implemented to combine cross-validatory as-

sessment with multiple imputation, which means avoiding the re-use of the set-aside data in the cross-validation when computing imputations. We also compared these approaches with the corresponding naïve implementations, which first derive imputations on the full dataset and subsequently compute cross-validation on the already-imputed data.

We investigated application of the proposed methods in prognosis, using two real datasets, and we also presented a simulation study where we generated lifetime outcome data subject to censoring and with missing data in predictors. We used simulations to assess the proposed methodologies in different scenarios. Finally, to generate multiple imputations for the survival data, we replaced the observed follow-up times with the Nelson-Aalen estimator of the cumulative hazard.

Results demonstrate, both for the real data and across simulations, that the first approach is vastly superior in terms of variation of the achieved predictions due to multiple imputations. This seems to be true irrespective of the number of imputations used and when comparing the first approach with both the approaches 2A and 2B. Indeed, even when increasing the number of imputations to 1000, the variability of the two approaches based on Rubin's rules is outperformed by approach 1 when using only 10 imputations.
Regarding the naïve implementations of the proposed methods, simulations have shown that they should be avoided, as they may exhibit optimistic bias. As we expected, indeed, computing multiple imputation prior to cross-validation leads to overrate the performance of the calibrated predictive model, since the predictive rules have been calibrated already "knowing" exactly what they have to predict.
Finally, we have shown that the number of multiple imputations should be much higher than current practice would suggest for predictive purpose and it should then be likely closer to 1000 imputations (or even more) in order to achieve reliable predictions which can be used in practical clinical applications. Moreover, the most important thing is that any implementation based on single imputation should be treated with greatest caution, since it may

lead to misleading results.

Preferring the first approach, of course, goes against the desire to report interpretable models, which makes the use of models based on Rubin's rules more attractive. Approach 1 is, in fact, specially built for predictive purpose and any attention has been paid to the parameters. However, Rubin's rules pooled estimates and standard errors may also be reported for interpretation of effects side to side with performance measures for approach 1. This issue is then left to further research and considerations.

Variable selection is another aspect that would require more investigation, since in this work we simply used all the available predictors, without any kind of selection.

Another topic that should be explored is the extension to handling other outcomes, such as continuous or binary, even though in theory there are no restriction to the general applicability of the discussed methodologies.

Appendix

.1 Real data

CRT		Brier Score			C-index		
M=100		Missing	Fully obs.	All obs.	Missing	Fully obs.	All obs.
1 yr	A 1	0.0711	0.0627	0.0670	0.8261	0.7134	0.7723
	A 2A	0.0710	0.0628	0.0669	0.8140	0.7039	0.7611
	A 2B	0.0711	0.0628	0.0669	0.8140	0.7039	0.7611
	N 1	0.0702	0.0627	0.0665	0.8416	0.7131	0.7803
	N 2A	0.0701	0.0626	0.0663	0.8281	0.7052	0.7696
	N 2B	0.0701	0.0626	0.0664	0.8281	0.7052	0.7696
Compl. case			0.0629			0.7353	
5 yrs	A 1	0.2104	0.1760	0.1903	0.6869	0.7706	0.7377
	A 2A	0.2110	0.1763	0.1907	0.6808	0.7590	0.7287
	A 2B	0.2109	0.1763	0.1907	0.6808	0.7590	0.7287
	N 1	0.2040	0.1757	0.1874	0.7058	0.7710	0.7457
	N 2A	0.2039	0.1755	0.1872	0.6946	0.7634	0.7367
	N 2B	0.2038	0.1754	0.1872	0.6946	0.7634	0.7367
Compl. case			0.1780			0.7346	

Table 9: Brier score and C-index statistics for the CRT data based on 100 multiple imputations and on 10-folds cross-validation. We report the average values of these statistics across 10 replicates of the approaches.Results are shown for both 1 and 5 years of follow-up, for the full set of observations (All obs.), for those observations with missing values (Missing) and for the completely observed records (Fully obs.). Results for the complete case analysis are shown as well.

CLL		Brier Score			C-index		
M=100		Missing	Fully obs.	All obs.	Missing	Fully obs.	All obs.
1 yr	A 1	0.2020	0.1814	0.1885	0.6713	0.5980	0.6266
	A 2A	0.2026	0.1818	0.1890	0.6683	0.5996	0.6260
	A 2B	0.2026	0.1818	0.1890	0.6683	0.5996	0.6260
	N 1	0.1965	0.1807	0.1861	0.7000	0.6032	0.6401
	N 2A	0.1980	0.1812	0.1870	0.6879	0.5999	0.6331
	N 2B	0.1980	0.1812	0.1870	0.6879	0.5999	0.6331
Compl. case			0.1837			0.5950	
5 yrs	A 1	0.2362	0.2412	0.2402	0.6463	0.6119	0.6243
	A 2A	0.2362	0.2410	0.2400	0.6496	0.6147	0.6268
	A 2B	0.2361	0.2411	0.2401	0.6496	0.6147	0.6268
	N 1	0.2288	0.2396	0.2360	0.6752	0.6179	0.6402
	N 2A	0.2316	0.2414	0.2383	0.6640	0.6118	0.6310
	N 2B	0.2317	0.2414	0.2383	0.6640	0.6118	0.6310
Compl. case			0.2560			0.6046	

Table 10: Brier score and C-index statistics for the CLL data based on 100 multiple imputations and on 10-folds cross-validation. We report the average values of these statistics across 10 replicates of the approaches.Results are shown for both 1 and 5 years of follow-up, for the full set of observations (All obs.), for those observations with missing values (Missing) and for the completely observed records (Fully obs.). Results for the complete case analysis are shown as well.

.2 Simulation Study

MCAR		1 year			5 years		
Compl. Case		Bias	Brier Score	C-index	Bias	Brier Score	C-index
M=10	Scen 1	-	0.0803	0.5859	0.0007	0.2243	0.6032
	Scen 2	-0.0014	0.0872	0.6878	-0.0012	0.2041	0.7162
	Scen 3	-	0.0803	0.5773	0.0020	0.2256	0.5932
	Scen 4	0.0041	0.0867	0.6834	0.0033	0.2053	0.7113
M=100	Scen 1	-	0.0803	0.5859	0.0007	0.2243	0.6032
	Scen 2	0.0005	0.0865	0.6900	0.0007	0.2043	0.7143
	Scen 3	-	0.0803	0.5930	-0.0017	0.2253	0.6037
	Scen 4	-0.0035	0.0884	0.6853	0.0008	0.2044	0.7123

Table 11: Results for bias, Brier score and C-index statistics for predictions obtained with a single calibration of the Cox model with cross-validatory assessment for the 4 scenarios with MCAR values and considering for the analysis only the fully observed records (complete cases) of the datasets which have been generated to test the approaches with $M = 10$ or 100. Results are shown at both 1 and 5 years of follow-up.

MAR		1 year			5 years		
Compl. Case		Bias	Brier Score	C-index	Bias	Brier Score	C-index
M=10	Scen 1	-	0.0801	0.5881	-0.0007	0.2242	0.6049
	Scen 2	-0.0039	0.0880	0.6871	0.0005	0.2042	0.7149
	Scen 3	-	0.0783	0.5852	0.0014	0.2239	0.6005
	Scen 4	-0.0042	0.0950	0.6776	-0.0011	0.2086	0.7136
M=100	Scen 1	-	0.0801	0.5898	-0.0007	0.2241	0.6056
	Scen 2	0.0013	0.0861	0.6949	0.0015	0.2031	0.7170
	Scen 3	-	0.0779	0.5910	0.0023	0.2232	0.6006
	Scen 4	-0.0104	0.0951	0.6866	0.0025	0.2076	0.7157

Table 12: Results for bias, Brier score and C-index statistics for predictions obtained with a single calibration of the Cox model with cross-validatory assessment for the 4 scenarios with MAR values and considering for the analysis only the fully observed records (complete cases) of the datasets which have been generated to test the approaches with $M = 10$ or 100. Results are shown at both 1 and 5 years of follow-up.

MCAR		1 year			5 years		
Original Data		Bias	Brier Score	C-index	Bias	Brier Score	C-index
M=10	Scen 1	-	0.0801	0.5884	-0.0007	0.2243	0.6041
	Scen 2	-0.0019	0.0873	0.6876	-0.0010	0.2039	0.7160
	Scen 3	-	0.0797	0.5873	0.0031	0.2238	0.6015
	Scen 4	0.0019	0.0866	0.6881	0.0001	0.2040	0.7154
M=100	Scen 1	-	0.0801	0.5884	-0.0007	0.2243	0.6041
	Scen 2	0.0004	0.0865	0.6903	0.0004	0.2041	0.7146
	Scen 3	-	0.0791	0.5947	-0.0024	0.2245	0.6075
	Scen 4	0.0004	0.0866	0.6897	0.0018	0.2035	0.7150

Table 13: Results for bias, Brier score and C-index statistics for predictions obtained with a single calibration of the Cox model with cross-validatory assessment for the 4 scenarios with MCAR values and considering for the analysis the original datasets (without missing values) which have been generated to test the approaches with $M = 10$ or 100. Results are shown at both 1 and 5 years of follow-up.

MAR		1 year			5 years		
Original Data		Bias	Brier Score	C-index	Bias	Brier Score	C-index
M=10	Scen 1	-	0.0801	0.5884	-0.0007	0.2243	0.6041
	Scen 2	-0.0043	0.0873	0.6877	0.0002	0.2037	0.7147
	Scen 3	-	0.0792	0.5865	0.0016	0.2239	0.6037
	Scen 4	-0.0052	0.0881	0.6862	-0.0004	0.2038	0.7150
M=100	Scen 1	-	0.0801	0.5894	-0.0005	0.2242	0.6048
	Scen 2	-0.0018	0.0856	0.6954	0.0013	0.2028	0.7165
	Scen 3	-	0.0792	0.5968	0.0027	0.2234	0.6047
	Scen 4	-0.0080	0.0881	0.6896	0.0020	0.2026	0.7174

Table 14: Results for bias, Brier score and C-index statistics for predictions obtained with a single calibration of the Cox model with cross-validatory assessment for the 4 scenarios with MAR values and considering for the analysis the original datasets (without missing values) which have been generated to test the approaches with $M = 10$ or 100. Results are shown at both 1 and 5 years of follow-up.

MCAR		Variation		Bias		Brier Score		C-index	
M=10, 1 year		Missing	Fully obs.	Missing	Fully obs.	Missing	Fully obs.	Missing	Fully obs.
	A 1	-	-	-	-	0.0780	0.0803	0.5911	0.5867
	A 2A	-	-	-	-	0.0780	0.0803	0.5911	0.5862
Scen 1	A 2B	-	-	-	-	0.0780	0.0803	0.5911	0.5862
	N 1	-	-	-	-	0.0779	0.0803	0.5978	0.5870
	N 2A	-	-	-	-	0.0779	0.0803	0.5969	0.5864
	N 2B	-	-	-	-	0.0779	0.0803	0.5969	0.5864
	A 1	0.0544	0.0098	0.1251	0.0004	0.0928	0.0871	0.5876	0.6880
	A 2A	0.0552	0.0278	0.1249	0.0004	0.0929	0.0872	0.5862	0.6878
Scen 2	A 2B	0.0551	0.0279	0.1249	0.0005	0.0929	0.0872	0.5862	0.6878
	N 1	0.0533	0.0096	0.1123	0.0004	0.0879	0.0871	0.7566	0.6882
	N 2A	0.0551	0.0266	0.1123	0.0004	0.0879	0.0871	0.7562	0.6880
	N 2B	0.0551	0.0269	0.1123	0.0004	0.0879	0.0871	0.7562	0.6880
	A 1	-	-	-	-	0.0793	0.0801	0.5872	0.5829
	A 2A	-	-	-	-	0.0793	0.0801	0.5870	0.5827
Scen 3	A 2B	-	-	-	-	0.0793	0.0801	0.5870	0.5827
	N 1	-	-	-	-	0.0791	0.0801	0.5984	0.5849
	N 2A	-	-	-	-	0.0791	0.0801	0.5966	0.5845
	N 2B	-	-	-	-	0.0791	0.0801	0.5966	0.5845
	A 1	0.0528	0.0153	0.1318	0.0115	0.0906	0.0865	0.5934	0.6845
	A 2A	0.0543	0.0315	0.1318	0.0117	0.0907	0.0865	0.5927	0.6843
Scen 4	A 2B	0.0547	0.0331	0.1318	0.0116	0.0907	0.0865	0.5927	0.6843
	N 1	0.0529	0.0140	0.1191	0.0115	0.0859	0.0864	0.7613	0.6855
	N 2A	0.0542	0.0268	0.1191	0.0119	0.0859	0.0864	0.7600	0.6853
	N 2B	0.0546	0.0283	0.1190	0.0117	0.0859	0.0864	0.7600	0.6853

Table 15: Results of the simulation study for the statistics described in section 3.3 at 1 year of follow-up and using $M = 10$ multiple imputations. Results are reported separately for fully observed records and for those with missing values and referred to simulated datasets with MCAR values.

MCAR		Variation		Bias		Brier Score		C-index	
M=10, 5 years		Missing	Fully obs.	Missing	Fully obs.	Missing	Fully obs.	Missing	Fully obs.
	A 1	0.0191	0.0088	-0.0000	-0.0006	0.2265	0.2240	0.6001	0.6044
	A 2A	0.0293	0.0257	-0.0001	-0.0006	0.2266	0.2241	0.6001	0.6040
Scen 1	A 2B	0.0293	0.0258	-0.0001	-0.0006	0.2266	0.2241	0.6001	0.6040
	N 1	0.0190	0.0086	-0.0001	-0.0006	0.2254	0.2240	0.6082	0.6048
	N 2A	0.0292	0.0254	-0.0001	-0.0006	0.2255	0.2241	0.6074	0.6043
	N 2B	0.0292	0.0254	-0.0001	-0.0006	0.2255	0.2241	0.6074	0.6043
	A 1	0.1186	0.0090	0.0182	-0.0002	0.2303	0.2039	0.5955	0.7165
	A 2A	0.1211	0.0256	0.0182	-0.0002	0.2306	0.2040	0.5948	0.7162
Scen 2	A 2B	0.1211	0.0261	0.0182	-0.0002	0.2306	0.2040	0.5948	0.7162
	N 1	0.1117	0.0089	0.0172	-0.0002	0.1893	0.2038	0.8075	0.7166
	N 2A	0.1136	0.0248	0.0171	-0.0002	0.1894	0.2039	0.8070	0.7164
	N 2B	0.1137	0.0254	0.0172	-0.0001	0.1894	0.2039	0.8070	0.7164
	A 1	0.0249	0.0127	0.0033	0.0033	0.2237	0.2243	0.6001	0.5993
	A 2A	0.0319	0.0287	0.0035	0.0035	0.2237	0.2245	0.6000	0.5987
Scen 3	A 2B	0.0320	0.0289	0.0033	0.0033	0.2237	0.2245	0.6000	0.5987
	N 1	0.0242	0.0122	0.0033	0.0033	0.2218	0.2240	0.6144	0.6017
	N 2A	0.0314	0.0260	0.0035	0.0035	0.2222	0.2241	0.6118	0.6014
	N 2B	0.0315	0.0262	0.0033	0.0033	0.2222	0.2241	0.6118	0.6014
	A 1	0.1166	0.0143	0.0204	0.0036	0.2309	0.2044	0.6009	0.7130
	A 2A	0.1183	0.0297	0.0205	0.0036	0.2309	0.2045	0.6010	0.7127
Scen 4	A 2B	0.1190	0.0329	0.0206	0.0036	0.2309	0.2045	0.6010	0.7127
	N 1	0.1085	0.0133	0.0184	0.0038	0.1902	0.2039	0.8133	0.7141
	N 2A	0.1103	0.0258	0.0185	0.0039	0.1905	0.2040	0.8118	0.7139
	N 2B	0.1110	0.0293	0.0183	0.0037	0.1905	0.2040	0.8118	0.7139

Table 16: Results of the simulation study for the statistics described in section 3.3 at 5 years of follow-up and using $M = 10$ multiple imputations. Results are reported separately for fully observed records and for those with missing values and referred to simulated datasets with MCAR values.

MCAR		Variation		Bias		Brier Score		C-index	
M=100, 1 year		Missing	Fully obs.	Missing	Fully obs.	Missing	Fully obs.	Missing	Fully obs.
	A 1	-	-	-	-	0.0780	0.0803	0.5919	0.5868
	A 2A	-	-	-	-	0.0780	0.0803	0.5917	0.5866
Scen 1	A 2B	-	-	-	-	0.0780	0.0803	0.5917	0.5866
	N 1	-	-	-	-	0.0779	0.0803	0.5982	0.5870
	N 2A	-	-	-	-	0.0779	0.0803	0.5972	0.5866
	N 2B	-	-	-	-	0.0779	0.0803	0.5972	0.5866
	A 1	0.0168	0.0031	0.1335	0.0025	0.0901	0.0864	0.6206	0.6906
	A 2A	0.0221	0.0269	0.1334	0.0026	0.0901	0.0864	0.6196	0.6903
Scen 2	A 2B	0.0223	0.0271	0.1334	0.0026	0.0901	0.0864	0.6196	0.6903
	N 1	0.0171	0.0030	0.1202	0.0025	0.0854	0.0864	0.7965	0.6907
	N 2A	0.0222	0.0259	0.1203	0.0025	0.0854	0.0864	0.7956	0.6905
	N 2B	0.0223	0.0263	0.1203	0.0026	0.0854	0.0864	0.7956	0.6905
	A 1	-	-	-	-	0.0782	0.0801	0.5880	0.5962
	A 2A	-	-	-	-	0.0782	0.0801	0.5874	0.5958
Scen 3	A 2B	-	-	-	-	0.0782	0.0801	0.5874	0.5958
	N 1	-	-	-	-	0.0781	0.0801	0.5985	0.5980
	N 2A	-	-	-	-	0.0781	0.0801	0.5961	0.5976
	N 2B	-	-	-	-	0.0781	0.0801	0.5961	0.5976
	A 1	0.0166	0.0049	0.1315	0.0115	0.0889	0.0881	0.5957	0.6867
	A 2A	0.0221	0.0299	0.1316	0.0118	0.0889	0.0881	0.5956	0.6865
Scen 4	A 2B	0.0229	0.0317	0.1315	0.0117	0.0889	0.0882	0.5956	0.6865
	N 1	0.0165	0.0044	0.1200	0.0113	0.0843	0.0880	0.7748	0.6876
	N 2A	0.0215	0.0239	0.1201	0.0116	0.0843	0.0881	0.7732	0.6874
	N 2B	0.0221	0.0258	0.1201	0.0117	0.0843	0.0881	0.7732	0.6874

Table 17: Results of the simulation study for the statistics described in section 3.3 at 1 year of follow-up and using $M = 100$ multiple imputations. Results are reported separately for fully observed records and for those with missing values and referred to simulated datasets with MCAR values.

MCAR		Variation		Bias		Brier Score		C-index	
M=100, 5 years		Missing	Fully obs.	Missing	Fully obs.	Missing	Fully obs.	Missing	Fully obs.
	A 1	0.0060	0.0028	-0.0000	-0.0006	0.2265	0.2240	0.6006	0.6045
	A 2A	0.0234	0.0255	-0.0000	-0.0006	0.2265	0.2241	0.6006	0.6042
Scen 1	A 2B	0.0234	0.0255	-0.0001	-0.0006	0.2265	0.2241	0.6006	0.6042
	N 1	0.0060	0.0027	-0.0000	-0.0006	0.2253	0.2239	0.6089	0.6048
	N 2A	0.0233	0.0251	-0.0000	-0.0006	0.2255	0.2241	0.6080	0.6043
	N 2B	0.0233	0.0251	-0.0000	-0.0006	0.2255	0.2241	0.6080	0.6043
	A 1	0.0381	0.0029	0.0209	0.0010	0.2293	0.2041	0.6070	0.7145
	A 2A	0.0453	0.0253	0.0209	0.0010	0.2295	0.2042	0.6062	0.7142
Scen 2	A 2B	0.0455	0.0259	0.0209	0.0010	0.2295	0.2042	0.6062	0.7142
	N 1	0.0358	0.0028	0.0186	0.0010	0.1888	0.2041	0.8378	0.7146
	N 2A	0.0420	0.0247	0.0187	0.0010	0.1889	0.2042	0.8370	0.7144
	N 2B	0.0423	0.0254	0.0187	0.0011	0.1889	0.2042	0.8370	0.7144
	A 1	0.0080	0.0040	-0.0019	-0.0023	0.2254	0.2240	0.6026	0.6089
	A 2A	0.0239	0.0273	-0.0017	-0.0021	0.2255	0.2242	0.6021	0.6082
Scen 3	A 2B	0.0241	0.0275	-0.0019	-0.0022	0.2255	0.2242	0.6021	0.6082
	N 1	0.0077	0.0038	-0.0019	-0.0023	0.2234	0.2237	0.6162	0.6112
	N 2A	0.0237	0.0241	-0.0017	-0.0021	0.2238	0.2238	0.6135	0.6108
	N 2B	0.0239	0.0243	-0.0020	-0.0024	0.2238	0.2238	0.6135	0.6108
	A 1	0.0362	0.0045	0.0214	0.0042	0.2282	0.2037	0.6088	0.7139
	A 2A	0.0440	0.0279	0.0214	0.0043	0.2283	0.2039	0.6088	0.7137
Scen 4	A 2B	0.0459	0.0315	0.0214	0.0043	0.2283	0.2039	0.6088	0.7137
	N 1	0.0337	0.0042	0.0203	0.0042	0.1889	0.2033	0.8284	0.7150
	N 2A	0.0400	0.0235	0.0204	0.0042	0.1892	0.2034	0.8266	0.7148
	N 2B	0.0422	0.0276	0.0207	0.0046	0.1892	0.2034	0.8266	0.7148

Table 18: Results of the simulation study for the statistics described in section 3.3 at 5 years of follow-up and using $M = 100$ multiple imputations. Results are reported separately for fully observed records and for those with missing values and referred to simulated datasets with MCAR values.

MAR		Variation		Bias		Brier Score		C-index	
M=10, 1 year		Missing	Fully obs.	Missing	Fully obs.	Missing	Fully obs.	Missing	Fully obs.
	A 1	-	-	-	-	0.0802	0.0801	0.5796	0.5889
	A 2A	-	-	-	-	0.0802	0.0801	0.5793	0.5884
Scen 1	A 2B	-	-	-	-	0.0802	0.0801	0.5793	0.5884
	N 1	-	-	-	-	0.0801	0.0801	0.5879	0.5891
	N 2A	-	-	-	-	0.0801	0.0801	0.5877	0.5886
	N 2B	-	-	-	-	0.0801	0.0801	0.5877	0.5886
	A 1	0.0511	0.0097	0.1362	-0.0028	0.0861	0.0880	0.5954	0.6874
	A 2A	0.0521	0.0271	0.1366	-0.0028	0.0861	0.0880	0.5936	0.6872
Scen 2	A 2B	0.0522	0.0273	0.1366	-0.0028	0.0861	0.0880	0.5936	0.6872
	N 1	0.0521	0.0094	0.1225	-0.0029	0.0814	0.0880	0.7711	0.6875
	N 2A	0.0536	0.0271	0.1225	-0.0029	0.0815	0.0880	0.7704	0.6873
	N 2B	0.0537	0.0272	0.1225	-0.0029	0.0815	0.0880	0.7704	0.6873
	A 1	-	-	-	-	0.0805	0.0781	0.5812	0.5865
	A 2A	-	-	-	-	0.0805	0.0781	0.5811	0.5861
Scen 3	A 2B	-	-	-	-	0.0805	0.0781	0.5811	0.5861
	N 1	-	-	-	-	0.0804	0.0781	0.5919	0.5884
	N 2A	-	-	-	-	0.0804	0.0781	0.5899	0.5883
	N 2B	-	-	-	-	0.0804	0.0781	0.5899	0.5883
	A 1	0.0504	0.0147	0.1344	0.0011	0.0849	0.0947	0.5969	0.6785
	A 2A	0.0520	0.0313	0.1345	0.0014	0.0849	0.0947	0.5967	0.6786
Scen 4	A 2B	0.0520	0.0326	0.1345	0.0014	0.0849	0.0947	0.5967	0.6786
	N 1	0.0517	0.0136	0.1210	0.0014	0.0805	0.0946	0.7679	0.6795
	N 2A	0.0527	0.0273	0.1211	0.0015	0.0806	0.0946	0.7670	0.6793
	N 2B	0.0529	0.0288	0.1210	0.0014	0.0806	0.0946	0.7670	0.6793

Table 19: Results of the simulation study for the statistics described in section 3.3 at 1 year of follow-up and using $M = 10$ multiple imputations. Results are reported separately for fully observed records and for those with missing values and referred to simulated datasets with MAR values.

MAR		Variation		Bias		Brier Score		C-index	
M=10, 5 years		Missing	Fully obs.	Missing	Fully obs.	Missing	Fully obs.	Missing	Fully obs.
	A 1	0.0202	0.0088	0.0003	-0.0007	0.2260	0.2240	0.5981	0.6054
	A 2A	0.0300	0.0257	0.0004	-0.0006	0.2261	0.2241	0.5975	0.6050
Scen 1	A 2B	0.0301	0.0257	0.0004	-0.0006	0.2261	0.2241	0.5975	0.6050
	N 1	0.0199	0.0087	0.0003	-0.0007	0.2246	0.2240	0.6080	0.6056
	N 2A	0.0302	0.0255	0.0003	-0.0007	0.2247	0.2241	0.6077	0.6051
	N 2B	0.0302	0.0255	0.0003	-0.0007	0.2247	0.2241	0.6077	0.6051
	A 1	0.1164	0.0089	0.0298	0.0004	0.2273	0.2040	0.5904	0.7152
	A 2A	0.1177	0.0252	0.0299	0.0003	0.2273	0.2041	0.5900	0.7149
Scen 2	A 2B	0.1180	0.0258	0.0300	0.0004	0.2273	0.2041	0.5900	0.7149
	N 1	0.1091	0.0087	0.0270	0.0004	0.1874	0.2039	0.8114	0.7153
	N 2A	0.1108	0.0248	0.0270	0.0004	0.1874	0.2040	0.8108	0.7151
	N 2B	0.1110	0.0253	0.0270	0.0004	0.1875	0.2040	0.8108	0.7151
	A 1	0.0241	0.0126	0.0015	0.0019	0.2253	0.2228	0.5991	0.6051
	A 2A	0.0314	0.0289	0.0017	0.0021	0.2253	0.2230	0.5992	0.6045
Scen 3	A 2B	0.0314	0.0291	0.0015	0.0019	0.2253	0.2230	0.5992	0.6045
	N 1	0.0235	0.0122	0.0015	0.0019	0.2234	0.2225	0.6126	0.6075
	N 2A	0.0311	0.0260	0.0017	0.0020	0.2237	0.2226	0.6101	0.6072
	N 2B	0.0312	0.0262	0.0015	0.0018	0.2237	0.2226	0.6101	0.6072
	A 1	0.1124	0.0139	0.0263	0.0009	0.2250	0.2078	0.5979	0.7149
	A 2A	0.1143	0.0298	0.0265	0.0009	0.2251	0.2079	0.5978	0.7148
Scen 4	A 2B	0.1149	0.0327	0.0265	0.0010	0.2251	0.2079	0.5978	0.7148
	N 1	0.1045	0.0130	0.0246	0.0011	0.1871	0.2073	0.8068	0.7160
	N 2A	0.1062	0.0259	0.0247	0.0011	0.1873	0.2074	0.8055	0.7158
	N 2B	0.1070	0.0292	0.0247	0.0012	0.1874	0.2074	0.8055	0.7158

Table 20: Results of the simulation study for the statistics described in section 3.3 at 5 years of follow-up and using $M = 10$ multiple imputations. Results are reported separately for fully observed records and for those with missing values and referred to simulated datasets with MAR values.

MAR		Variation		Bias		Brier Score		C-index	
M=100, 1 year		Missing	Fully obs.	Missing	Fully obs.	Missing	Fully obs.	Missing	Fully obs.
	A 1	-	-	-	-	0.0801	0.0801	0.5787	0.5906
	A 2A	-	-	-	-	0.0802	0.0801	0.5783	0.5903
Scen 1	A 2B	-	-	-	-	0.0802	0.0801	0.5783	0.5903
	N 1	-	-	-	-	0.0800	0.0801	0.5869	0.5909
	N 2A	-	-	-	-	0.0800	0.0801	0.5863	0.5902
	N 2B	-	-	-	-	0.0800	0.0801	0.5863	0.5902
	A 1	-	0.0031	-	-0.0002	0.0840	0.0861	0.6059	0.6951
	A 2A	-	0.0270	-	-0.0001	0.0841	0.0861	0.6049	0.6949
Scen 2	A 2B	-	0.0272	-	-0.0001	0.0841	0.0861	0.6049	0.6949
	N 1	-	0.0030	-	-0.0001	0.0797	0.0861	0.7916	0.6952
	N 2A	-	0.0264	-	-0.0002	0.0797	0.0861	0.7912	0.6950
	N 2B	-	0.0267	-	-0.0002	0.0797	0.0861	0.7912	0.6950
	A 1	-	-	-	-	0.0808	0.0777	0.5946	0.5963
	A 2A	-	-	-	-	0.0808	0.0777	0.5943	0.5959
Scen 3	A 2B	-	-	-	-	0.0808	0.0777	0.5943	0.5959
	N 1	-	-	-	-	0.0806	0.0777	0.6053	0.5982
	N 2A	-	-	-	-	0.0807	0.0777	0.6033	0.5977
	N 2B	-	-	-	-	0.0807	0.0777	0.6033	0.5977
	A 1	0.0164	0.0048	0.1324	-0.0034	0.0844	0.0949	0.5990	0.6877
	A 2A	0.0210	0.0299	0.1322	-0.0032	0.0844	0.0949	0.5990	0.6875
Scen 4	A 2B	0.0216	0.0314	0.1322	-0.0033	0.0844	0.0949	0.5990	0.6875
	N 1	0.0168	0.0044	0.1186	-0.0036	0.0797	0.0948	0.7894	0.6885
	N 2A	0.0210	0.0255	0.1187	-0.0034	0.0798	0.0948	0.7881	0.6883
	N 2B	0.0217	0.0273	0.1187	-0.0035	0.0798	0.0948	0.7881	0.6883

Table 21: Results of the simulation study for the statistics described in section 3.3 at 1 year of follow-up and using $M = 100$ multiple imputations. Results are reported separately for fully observed records and for those with missing values and referred to simulated datasets with MAR values.

MAR		Variation		Bias		Brier Score		C-index	
M=100, 5 years		Missing	Fully obs.	Missing	Fully obs.	Missing	Fully obs.	Missing	Fully obs.
	A 1	0.0063	0.0028	0.0002	-0.0006	0.2258	0.2239	0.5979	0.6061
	A 2A	0.0236	0.0253	0.0002	-0.0006	0.2258	0.2240	0.5979	0.6058
Scen 1	A 2B	0.0236	0.0253	0.0002	-0.0006	0.2258	0.2240	0.5979	0.6058
	N 1	0.0063	0.0027	0.0002	-0.0006	0.2244	0.2239	0.6077	0.6064
	N 2A	0.0236	0.0250	0.0002	-0.0006	0.2246	0.2240	0.6071	0.6059
	N 2B	0.0236	0.0250	0.0002	-0.0006	0.2246	0.2240	0.6071	0.6059
	A 1	0.0371	0.0028	0.0279	0.0019	0.2230	0.2029	0.6033	0.7173
	A 2A	0.0439	0.0251	0.0279	0.0020	0.2232	0.2030	0.6024	0.7171
Scen 2	A 2B	0.0441	0.0257	0.0280	0.0020	0.2232	0.2030	0.6024	0.7171
	N 1	0.0347	0.0028	0.0251	0.0020	0.1837	0.2029	0.8357	0.7174
	N 2A	0.0409	0.0246	0.0251	0.0020	0.1838	0.2030	0.8350	0.7172
	N 2B	0.0412	0.0251	0.0251	0.0019	0.1837	0.2030	0.8350	0.7172
	A 1	0.0078	0.0040	0.0025	0.0031	0.2249	0.2220	0.6007	0.6058
	A 2A	0.0239	0.0270	0.0027	0.0033	0.2250	0.2221	0.6005	0.6052
Scen 3	A 2B	0.0241	0.0272	0.0025	0.0031	0.2250	0.2221	0.6005	0.6052
	N 1	0.0075	0.0037	0.0025	0.0031	0.2230	0.2216	0.6146	0.6080
	N 2A	0.0237	0.0240	0.0027	0.0033	0.2235	0.2217	0.6119	0.6077
	N 2B	0.0239	0.0241	0.0025	0.0031	0.2235	0.2217	0.6119	0.6077
	A 1	0.0360	0.0044	0.0299	0.0036	0.2227	0.2066	0.6060	0.7173
	A 2A	0.0433	0.0278	0.0299	0.0036	0.2228	0.2067	0.6055	0.7170
Scen 4	A 2B	0.0450	0.0314	0.0299	0.0036	0.2228	0.2067	0.6055	0.7170
	N 1	0.0336	0.0041	0.0276	0.0035	0.1831	0.2062	0.8359	0.7183
	N 2A	0.0396	0.0238	0.0276	0.0036	0.1834	0.2063	0.8343	0.7181
	N 2B	0.0415	0.0278	0.0276	0.0035	0.1834	0.2063	0.8343	0.7181

Table 22: Results of the simulation study for the statistics described in section 3.3 at 5 years of follow-up and using $M = 100$ multiple imputations. Results are reported separately for fully observed records and for those with missing values and referred to simulated datasets with MAR values.

Bibliography

Blanche, Paul, Jean-Francois Dartigues, and Helene Jacqmin-Gadda (2013). "Estimating and Comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks". In: *Statistics in Medicine* 32.30, pp. 5381–5397.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2009). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction.* Second edition. Vol. 1. Springer series in statistics New York. Chap. 7, pp. 241–247.

Graf, Erika et al. (1999). "Assessment and comparison of prognostic classification schemes for survival data". In: *Statistics in Medicine* 18.17-18, pp. 2529–2545.

Harrell, Frank E., Kerry L. Lee, and Daniel B. Mark (1996). "Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors". In: *Statistics in medicine* 15.4, pp. 361–387.

Hoke, Ulas et al. (2017). "Usefulness of the CRT-SCORE for Shared Decision Making in Cardiac Resynchronization Therapy in Patients With a Left Ventricular Ejection Fraction of 35". In: *The American Journal of Cardiology* 120.11, pp. 2008–2016.

Kartsonaki, Christiana (2016). "Survival analysis". In: *Diagnostic Histopathology* 22.7, pp. 263–270.

Kleinbaum, D. G. and M. Klein (2012). *Survival Analysis. A Self-Learning Text*. Ed. by Statistics for Biology and Health. Third Edition. Springer-Verlag New York.

Kohavi, Ron (1995). "A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection". In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*. IJCAI'95. Montreal, Quebec, Canada: Morgan Kaufmann Publishers Inc., pp. 1137–1143.

Lesaffre, Emmanuel and Andrew B Lawson (2012). *Bayesian biostatistics*. John Wiley & Sons.

Little, Roderick J. A. and Donald B. Rubin (2014). *Statistical analysis with missing data*. Vol. 333. John Wiley & Sons.

Mogensen, Ulla B., Hemant Ishwaran, and Thomas A. Gerds (2012). "Evaluating Random Forests for Survival Analysis Using Prediction Error Curves". In: *Journal of Statistical Software* 50.11, pp. 1–23.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Raghunathan, Trivellore E. et al. (2001). "A multivariate technique for multiply imputing missing values using a sequence of regression models". In: *Survey methodology* 27.1, pp. 85–96.

Rahman, M. Shafiqur et al. (2017). "Review and evaluation of performance measures for survival prediction models in external validation settings". In: *BMC Medical Research Methodology* 17.1, pp. 60–75.

Rubin, Donald B. (1976). "Inference and missing data". In: *Biometrika* 63.3, pp. 581–592.

— (1978). "Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse". In: *Proceedings of the survey research methods section of the American Statistical Association*. Vol. 1. American Statistical Association, pp. 20–34.

— (1996). "Multiple imputation after 18+ years". In: *Journal of the American statistical Association* 91.434, pp. 473–489.

— (2004). *Multiple imputation for nonresponse in surveys*. Vol. 81. John Wiley & Sons.

Rubin, Donald B. and Nathaniel Schenker (1986). "Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse". In: *Journal of the American Statistical Association* 81.394, pp. 366–374.

Schetelig, J. et al. (2017a). "Centre characteristics and procedure-related factors have an impact on outcomes of allogeneic transplantation for patients with CLL: a retrospective analysis from the European Society for Blood and Marrow Transplantation (EBMT)". In: *British journal of haematology* 178.4, pp. 521–533.

Schetelig, J. et al. (2017b). "Risk factors for treatment failure after allogeneic transplantation of patients with CLL: a report from the European Society for Blood and Marrow Transplantation". In: *Bone marrow transplantation* 52.4, pp. 552–560.

Steyerberg, Ewout W. et al. (2010). "Assessing the performance of prediction models: a framework for some traditional and novel measures". In: *Epidemiology (Cambridge, Mass.)* 21.1, pp. 128–138.

Uno, Hajime et al. (2007). "Evaluating prediction rules for t-year survivors with censored regression models". In: *Journal of the American Statistical Association* 102.478, pp. 527–537.

Van Buuren, S. and C. G. M. Oudshoorn (2000). *Multivariate imputation by chained equations: MICE V1. 0 user's manual*. Leiden: TNO.

Van Buuren, Stef (2007). "Multiple imputation of discrete and continuous data by fully conditional specification". In: *Statistical methods in medical research* 16.3, pp. 219–242.

— (2012). *Flexible imputation of missing data*. CRC press.

Van Buuren, Stef and Karin Groothuis-Oudshoorn (2011). "mice: Multivariate Imputation by Chained Equations in R". In: *Journal of Statistical Software* 45.3, pp. 1–67.

Van Buuren, Stef et al. (2006). "Fully conditional specification in multivariate imputation". In: *Journal of statistical computation and simulation* 76.12, pp. 1049–1064.

Van Houwelingen, Hans and Hein Putter (2011). *Dynamic prediction in clinical survival analysis*. CRC Press.

White, Ian R. and Patrick Royston (2009). "Imputing missing covariate values for the Cox model". In: *Statistics in medicine* 28.15, pp. 1982–1998.

White, Ian R., Patrick Royston, and Angela M. Wood (2011). "Multiple imputation using chained equations: issues and guidance for practice". In: *Statistics in medicine* 30.4, pp. 377–399.