RICE UNIVERSITY

Using multiple imputation, survival analysis, and propensity score analysis in cancer data with a large amount of missing data

by

Nathan Karmazin Berliner

A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree

Master of Arts Statistics

Approved, Thesis Committee:

Rudy Guerra, Committee Chair
Professor of Statistics

Kenneth Hess, Thesis Director
Professor, MD Anderson Cancer Center

Yu Shen
Professor, MD Anderson Cancer Center

Marina Vannucci
Professor and Chair of Statistics

Houston, Texas

April, 2015

ABSTRACT

Using multiple imputation, survival analysis, and propensity score analysis in cancer data with a large amount of missing data

by

Nathan Karmazin Berliner

In this thesis, we will use the theory of multiple imputation, survival analysis, and propensity score analysis in order to answer questions about cancer survival data. While each of these fields have been studied individually, there has been little work and analysis on using the three in trio. Starting out with an incomplete dataset, we aim to impute reasonable imputations, run survival analysis on each of the imputed datasets to get survival estimates, and then do propensity score analysis to observe causal effects. Along the way, many theoretical and analytical decisions are mode. I explain why each decision is made, but offer ample evidence for the other choices such that the interested reader may implement the methods if they so choose. I apply the methodology to a cancer survival dataset in a case study, but the methods used are general, and could be used for any type of data.

Contents

Illustrations

Tables

Chapter 1

Introduction and Background information

1.1 Motivation

The motivation of this thesis is to show the methodology that can be used both by applied researchers and clinicians to draw meaningful survival inference from data with a high amount of missingness. While both fields are well studied, their interaction is not. I want the methods to be easy enough to describe to someone with a limited statistical background, but meaningful and valid so that the results obtained can be used in publication. The desire to have it this way stems from working on a related project with both statisticians and clinicians. While we will be motivated by cancer data, I believe that the methods used in this thesis are general enough to be applied to other types of survival data.

Missing data is a major problem in both statistics and medicine, however, it has not received attention proportional to its need. Survival analysis is well studied, but is relatively complete, so not much new research comes out of this field. Propensity score analysis will help us determine causal relationships when we dont have a completely randomized experiment. As one could imagine, all three of these fields are important to the applied statistician, as they will come across at least one at some point in their career. The goal of this thesis is to demonstrate how to use all three in trio, a topic that has only received little interest in the literature. I will explain each of these three disciplines in detail before we dive into combining them.

1.2 Imputation

In an ideal world, we would have complete data with no missingness, however this is rarely ever the case. Imputation (specifically multiple imputation) is a way to fill in missing data with plausible values, and it forms the base of this paper. All of the other analyses that will be used will follow from it, thus we need a good understanding of it before we may proceed. Imputation itself has been around since the 1970s, but multiple imputation is a recent development, proposed formally in 1987 by Donald Rubin [1]. To understand the use and importance of multiple imputation, we need to understand the problem of missing data, and the previous attempts to deal with it.

At first, statisticians payed no attention to missing data, and happily discarded records for their analysis that were incomplete. This procedure is known as complete case analysis. There are many problems with this paradigm. To begin with, you will lose a lot of statistical power when doing this, because you are literally throwing away records and thus decreasing your sample size. In addition, this can be costly to the researcher. If it costs a set amount to collect a single record, and you dont use this record, you are literally wasting money. As well, in some rare cases, incomplete data might be the only type we can get (like if we have a machine that analyzes a blood sample chemical level, but cant detect it if the level is too high or low). Lastly, and most importantly, we will be biasing our estimates if we discard them. For example, if we have a random sample of people and are testing a drug, and want to run a regression on some collected covariates. Men are known to not want to give all of their information, so they leave them blank. In the analysis, we will need to discard the male samples because they are incomplete, leaving us only with women. Thus, we dont have a random sample anymore, and will get biased results because we have knowingly thrown away half of our data which we know to be different.

A slight improvement on this is called available case analysis. In this setting, a record is used in the analysis if it has all of the needed information for that analysis. So, a record could have missingness, but if the covariate with missingness is never used in the analysis, it will not be discarded. This paradigm is the standard analysis type for most statistical packages. It is better than complete case analysis, but is still flawed. We are still throwing away valuable data as we were with complete case analysis, although likely not as much. Available case analysis will still lead to bias in the same way that complete case did too. As well, new complications arise in available case analysis, namely that nonsensical situations like correlations outside of ± 1, and inconsistent sample sizes for different analyses can arrise.

The next wave of statisticians wanted to improve upon available case analysis, so they developed what we now call today imputation. Their specific incarnation was called single imputation, and their goal was to fill in missing values with a plausible replacement value. A single method (such as regression, taking the mean, resampling) is used one time to impute the missing value. While this is a little better than complete case analysis, it still has many drawbacks. Asserting that a single value is the true value is unjustified and foolish. There is always some amount of error and uncertainty involved, and we can in no way be 100% confident that our imputed value is correct. Furthermore, if I impute one value and you impute another, we may get totally different results from analysis on the data. This is obviously not a desirable trait. In addition, imputing one time and calling it your data will artificially increase your sample size. You are in effect treating the imputed values as if they were real, inflating your sample size with data that was not actually observed. This will give you unjustified statistical power and accuracy. While single imputation certainly has its drawbacks, the idea of actually trying to fill in the data is an important

one, and multiple imputation fills in the gaps that single imputation is not able to cover. Multiple imputation began in the 1970s, but it wasnt until 1987 when

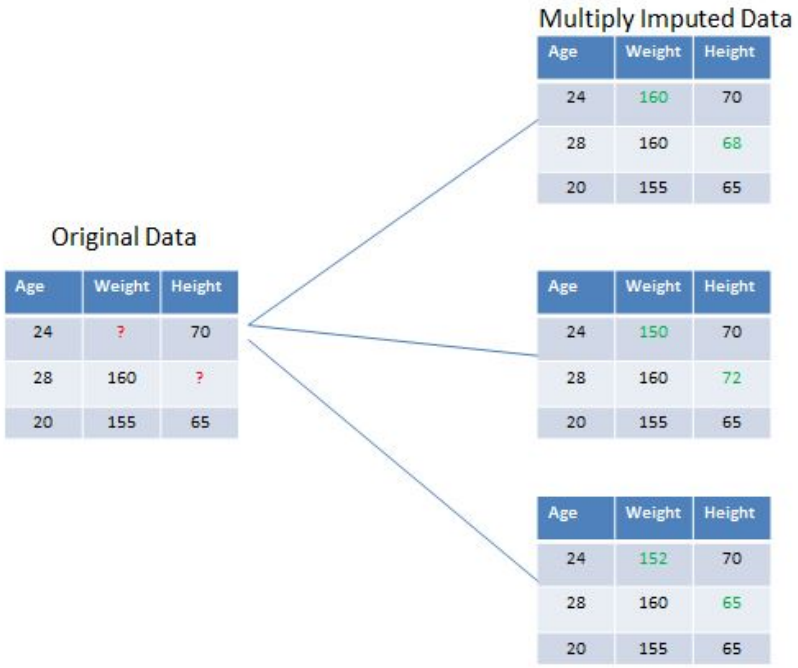

Figure 1.1 : Visualization of MI data

In the original data, missingness is displayed by ?s and the imputed data is shown in the multiply imputed data as #s

the Donald Rubin formalized multiple imputation methodology in his seminal book multiple imputation for nonresponse in survey did it start to gain acceptance [1]. The central idea is to frame the problem in a Bayesian framework, and produce $m \geq 2$ values to substitute in for each missing value, drawing these values from the missing covariates posterior distribution. Using these substitute values (m values), we can think of the data now as being m datasets, each dataset having the observed data,

and one value of the missing data. This might be hard to think about, but seeing a visualization of it will make it clear !! put in figure here!! . Once we have a sufficient number of datasets (we will talk about how to pick the number later), we can run whatever analyses we would like on them individually, treating the dataset as if it was complete. Once we have run the model on the m datasets, we can pool the results to get one estimate. We can get the standard errors by noting the within and between imputation variance.

This method is obviously much better than the first two methods because it allows is to not throw away data, as well as allowing us to quantify our uncertainty about imputing the missing values. The only real drawback of multiple imputation is that we still dont have true data, but we can be confident enough in our estimations to compensate for that. The use of multiple imputation has been steadily increasing over the past 30 years, and it is now the standard for missing data. Stef van Burren, an influential author in multiple imputation did a study of academic papers, and concluded that the number of publications using or mentioning multiple imputation is growing at an exponential rate since about 1990 [2].

1.3 Survival

Survival analysis is a huge field, and there have been many textbooks written about it. I only plan to introduce the topics that are relevant to my case study. For a much more detailed account of survival analysis, please see [3]. Survival analysis on the whole can generally be described as the analysis of time to event data, often in the presence of censoring (when we dont have complete information about the time of failure). There are many techniques used in this field, but the main tools that we will be using are Kaplan-Meier estimates, log rank tests, and Cox regression.

Before we go on, it should be noted that often in the literature (and in this paper), we see terms like death and survivors. This is due to survival analysis being heavily intertwined with medical studies. A more general term for these would be event and those who have not had an event yet. You dont actually have to die to be considered a death, you just need to have had the event of interest. Survival analysis doesnt have to be gloomy!

The Kaplan-Meier estimator (sometimes called the product limit estimator) is a non-parametric estimate of the true survival function (the probability that you survive after time t). It is defined as

$$\hat{S}(t) = \prod_{t_i<t} \frac{n_i - d_i}{n_i}$$

Where n_i is the number of survivors minus the number of censored cases just prior to time t_i, and d_i is the number of deaths or events that you observe at time t_i. The Kaplan-Meier estimator is very commonly used as a measure to see how different treatments affect survival time The log-rank test is a popular nonparametric test that researchers often use to see if two or more survival curves come from the same distribution. This is a useful tool to have, because visualizing curves alone does not give us this information (we could have two curves that look radically different due to sampling error, yet still come from the same distribution). Knowing about if the survival curves come from the same or different distribution is useful because it allows us to make statements like drug a is associated with longer survival time than drug b. It is given by

$$\frac{\sum_{j=1}^{J} w_j(O_{1j} - E_{1j})}{\sqrt{\sum_{j=1}^{j} w_j^2 V_j}} \sim N(0,1)$$

Where w_j is the weight of each individual (must be ≥ 0, we will set all to be 1),and $N_j = N_{1j} + N_{2j}$ is the number of deaths at time j, composed from the sum of the

number of deaths at time j in each group, $O_j = O1j + O_{2j}$ is the observed number of deaths at time j, composed of the sum of deaths from either group at time j, which leads to the desired quantities $E_{1j} = \frac{O_j N_{1j}}{N_j}$, and $V_j = \frac{O_j(N_{1j}/N_j)(1-N_{1j}/N_j)(N_j 0_j)}{N_j - 1}$ Since we set all of the weights to be 1, this test as it is places equal weight to all of the deaths we observe. We could change these weights though to give more emphasis to certain death times This is useful for example if we have a drug that is supposed to improve your life expectancy, we wouldnt care about early deaths, only about later times. Putting more weight on the later deaths would help to answer this question. It can be proven that the log-rank test is equivalent to the score test on a Cox model of the same data with no ties [3]. We will speak about the Cox model next.

Proportional hazards regression, often called Cox regression is a modelling tool that allows us to analyze the hazard ratio of a covariate, assuming that each covariate acts to multiply the hazard ratio. The hazard is a survival tool that tells us the rate of events at time t, conditional on survivorship until time t. Mathematically, it is given by

$$h(x) = \lim_{\Delta x \to 0} \frac{P[x \leq X < x + \Delta x | X \geq x]}{\Delta x}$$

Cox regression is a maximum (partial) likelihood method estimator, given by

$$h(t|Z) = h_0(t) \exp(\sum_{k=1}^{p} \beta_k Z_k)$$

The $h_0(t)$ is whats known as the baseline hazard, and can be any function that we would like. Note how it only depends on time. Z is our observed covariates. The βs are found by maximizing the partial likelihood function

$$L(\beta) = \prod_{i=1}^{D} \frac{\exp(\sum_{k=1}^{p} \beta_k Z_{(i)k})}{\sum_{j \in R(t_i)} \exp(\sum_{k=1}^{p} \beta_k Z_{(i)k})}$$

Where $Z_(i)k$ is subject is kth covariate, $R(t_j)$ is the risk set (set of those who have not diet yet at the time just prior to t_j),D is the number of distinct death times and p

is the total number. The betas are maximized by the Newton-Raphson method. Our inference of interest is the hazard ratio, given by $\frac{h(t|Z)}{h(t|Z^*)} = \exp(\sum_{k=1}^{p} \beta_k(Z_k - Z_k^*))$ Where Z* is another set of covariates. The relative risk (or hazard ratio) describes how the hazard changes between individual with different covariates. This ratio will be a constant (it should not vary over time); hence the name proportional hazards, and does not depend on the baseline hazard.

Using Cox regression, we can make statements such as increasing the drug by one mg will decrease the rate of death (compared to non-users) by 30%.

1.4 Propensity Score Analysis

Observing data is obviously needed, but unless we conduct a randomized controlled trial, we cannot make any claims about causality. In an ideal world we would like to be able to do research and say that A causes B, rather than our study says that A is associated to B.

Randomized controlled trial (RCT) is a term that is often thrown around, but I want to be precise with its definition. In the simplest example, RCT is an experiment where subjects (with no knowledge of the experiment) are randomly assigned to either the treatment of the control. Thus, the only difference between the subjects should be due to the fact that they have or have not received the drug. This is in stark contrast to a retrospective study, where we analyze historic data of people who chose what group they wanted to be in. When making judgment on this type of study, we cannot be sure if the differences between the groups is due to their group choice, or some other factor. RCTs are the gold standard, and we should try to use them if possible, but often times monetary, ethical, or other factors prevent us from doing so. In this case, the best data we may be able to get is retrospective study. By using

propensity score analysis, we aim to take data from a non-RCT, balance the groups so that they are similar to an RCT, and analyze them. In this way, we intend to let the only differences between the groups to be because of treatment selection.

There are many different ways that we can go about implementing Rubins causal model, but the most popular is what is called propensity score analysis. The propensity score is the probability that the subject received the treatment, and it is computed using the patients baseline (pretreatment) information. Propensity scores lead to group balancing, that is, if we control for propensity score, then our groups will have the same distribution at the baseline. And thus, we can treat it as if it was an RCT. The two most popular propensity score methods are propensity score matching, where we match those who picked the treatment to those who did not based off of propensity score, and propensity score weighting, where each individuals contribution to the average treatment effect is dependent on their propensity score. Actually getting the propensity score is often done by logistic regression, although newer methods include regression trees and other binary classifiers. Its use is justified by the propensity score theorem, which states that if we assume conditional independence of the treatment given covariates on the outcomes, then we can also assume conditional independence of the treatment given the propensity score on the outcomes. Symbolically

$$(Y(0), Y(1)) \perp T|X \implies (Y(0), Y(1)) \perp T|p(X)$$

The proof can be found in[4]. So, now we have the probability or propensity of a subject being assigned to a specific treatment given their covariates. Its usefulness is immediately seen in the matching case, because now instead of matching on n different levels, we need only to match on one number, the propensity score.

Chapter 2

Methods

I want the framework and methodology we use to be easy to use and understand, so that it can easily be discussed among clinicians and other people who dont have a statistics or mathematics background. On the same token, I want the methods and theory to be sound from a statistical point of view. For the three parts that we are combining, there are a lot of differing theories and paradigms. I aim to pick the ones that optimize ease of understanding and power of results. Along the way, we will also develop new methods and validation tools, as well as apply the existing theory to situations that it has not previously been applied to.

2.1 Multiple Imputation

Our first decision comes as to what paradigm we should impute under. It should be noted that as long as we can produce valid imputations, the choice of method does not matter. However, since the base of our analysis starts with imputation, we need to make sure that we pick a good method. Everything that follows in the analysis is dependent on our imputed data, so it is necessarily the case that bad imputation will lead to poor results (be it bias, high variability, loss in statistical power, etc.). The methods we will discuss here are geared towards and motivated by cancer research, but can be easily adapted to other areas.

There are two main divisions in modern multiple imputation, and they are joint

modelling and full conditional specification. Both have their own flaws and advantages. I will describe both, and then explain why full conditional specification is better suited for cancer research.

Before we get in to the imputation models, we need to have a firm understanding of missing data concepts. They take up quite a bit of space to explain, but they are fundamental concepts. If you are unfamiliar with them, please read appendix A.1 before reading further.

In joint modelling (JM), we assume that the missing data mechanism is ignorable and that the data can be described by a multivariate distribution on the rows of the data (specified by the user). We then draw imputations from the joint distribution of the unknowns for the rows, given what we do know and their associated unknown parameter of the imputation model. !!!!Example here!! Since we dont know the true model parameters, we need to estimate them. This is often done by a data augmentation algorithm [2]. There has been extensive research on using the normal model for this, and research shows that it even performs well under data that has strong non-normality. An obvious issue arises when we have discrete or categorical data. There has been much debate in the literature about what to do with it. Some authors argue that you should just impute under a continuous distribution and round, and others suggest using distributions that are more suited for categorical data [2]. There are a few R packages for joint modelling imputation include Amelia [5], norm [6] and cat [7]. On the other hand, there is fully conditional specification (FCS). In this paradigm, missing data is imputed on a variable by variable case (on the columns), based off of a specification of the imputation model for each imputed variable. These full conditionals should factor to specify the joint distribution. In the JM setting, we must give a k dimensional model, however in the FCS setting, we must give k one

dimensional models. We are trying to sample from

$$P(Y, X, R|\theta)$$

By sampling from the full conditionals

$$P(Y_j|X, Y_{-j}, R, \phi_j)$$

In this notation, Y_j means all of the columns with missing data except for j, and X is the fully observed columns (which could possibly be empty). One of the major flaws of this method is that in order for there to be a guarantee that we are sampling from the correct distribution, we need to ensure that our full conditionals are compatible, i.e that they factor into the proper joint. This is very hard to check in practice, but studies have shown that even when the models are highly incompatible, FCS methods are very robust [8]. But despite this, FCS allows us much more flexibility than JM does. This is the framework for FCS, and there are many different implementations of it. The three most common ones are the additive linear regression approach implemented in the Harrell package, and package mi We are going to have to specify something, there is no escaping that, but I think that it is easier for the average person (especially a clinician) to be able to define a single distribution and model rather than to guess at a multivariate. In addition, in the survival analysis setting, we will naturally have time variables be only positive, and some binary indicators, whereas others can take any value. Trying to fit a parametric distribution with these stipulations will be very hard if not impossible, so we will be relegated to using a general distribution (like the normal), which will certainly elicit a poor fit. So, the fully conditional specification will be our choice. In an ideal world, we would have complete data, and would not need to resort to imputation. But since we dont have complete data, we must choose one method and accept its strengths and weaknesses.

Now that we have chosen the paradigm, we need to select an implementation of it. Many exist (such as MICE [9], mi [10], etc.). I wanted to select the implementation that combined ease of use, understanding, and programming. What I decided upon was a method called MICE- Multiple Imputation by chained equations [9] MICE is an FCS MCMC method that under compatibility, is a Gibbs sampler, where we obtain samples from the joint by sampling from the full conditionals. The user defines the full conditionals, so it is possible that the joint may only exist implicitly, and not actually have a functional form.

In order to use mice, we must have that the missingness in our data to be MCAR or MAR. It can work with MNAR data, but it requires some extra modelling assumptions. This is a seldom observed case in practice, so the interested reader may check [9] section 6.2 for a detailed look at this.

The mice algorithm in pseudocode here!! Figure out how to do this!!

It should be noted that in the real data we will use, the response variable is fully observed, but the covariates have a lot of missingness. If it were the case that we had missingness in the survival time, then the methods described above might not work. They might fail because the unobserved times may follow a different distribution than the observed times. This is cleared up by Zhao et. al in 2014 through Kaplan-Meier MI [11]. This is beyond the scope of this report though so I omit its details. Once we have the correct assumptions, we need to set up our full conditionals imputation models. This may take a while for large datasets, but the extra time spent will ensure a better model. We choose what predictors will go into imputation, and what method to use (regression, predictive mean matching, logistic regression, etc.). We should choose predictor variables that are somewhat correlated with the missing data, as well as include the covariates that we are doing inference on, as to avoid bias.

For variables that are derived from others, we impute the others and then compute that variable, in a process known as passive imputation. Since mice is an iterative process, we must choose how many iterations we will do until convergence. The older literature suggests only 5 is enough (source), but with modern computation, we can easily exceed this, even with large data. As well, we need to decide how many datasets to impute. The early literature argued that 5 will due, but more is better, since it will cut down on simulation error (find where I wrote up the reasons why). Modern literature suggests X.

?? rhat test??We need to verify that our imputations are valid once we complete them. The overarching idea that we need to pay attention to is does the data look like it could have been real data. We can assess this in many ways, including density plots, box and whisker plots, etc. There is not much in the imputation literature about statistical tests to check for convergence, but !!!!!work on this!!

Once we have m imputed datasets, we may run any valid analysis (regression, computing any statistic) on each imputed dataset INDIVIDUALLY, treating each of the m datasets as if it was complete. We may then use Rubins rules [1] to pool our estimates. This will give us a point estimate, as well as the proper variance for the quantity we have in mind. Rubins rules are essential for using multiply imputed datasets, so we need to investigate them thoroughly.

Rubins rules are a set of rules that guide us in making inference from multiply imputed data. It involves three parts. The first is getting an estimate of the population estimand Q, we do so by taking the average of the MI sample estimands ($\hat{Q}_i$) to get the MI estimate $\bar{Q}$.

$$\bar{Q} = \frac{1}{m} \sum_{i=1}^{m} \hat{Q}_i$$

, Where $\hat{Q}$ is the estimand evaluated from the data in the i^{th} dataset. The estimates

are not set, and there is variance associated with them. The first form of variance is the within variance, or the variance or each estimate. We can get an MI estimate of this quantity by doing

$$\bar{U} = \frac{1}{m}\sum_{i=1}^{m}\bar{U}_i$$

Where $\bar{U}_i$ is the i^{th} datasets variance The other form of the variance is the between datasets variance. This is the variance associated with the fact that we have missing data. It is given by

$$B = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{Q}_i - \bar{Q})$$

The total variance for our estimand is given by

$$T = \bar{U} + B + \frac{B}{M}$$

The last term is our simulation variance, and its existence is proven by Rubin in [1] The theory is rooted in the assumption that whatever we are trying to pool is asymptotically normally distributed with mean Q and variance U. We dont have these population values, so we must use what we have from the sample, namely $\bar{Q}$ and $\bar{U}$. With this assumption, we know that

$$\frac{Q - \bar{Q}}{\sqrt{T}} \sim t_\nu$$

And the degrees of freedom is proven to be equal to !! !! Put source when I get the source for it from anderson

$$\nu = \frac{\nu_{old}\nu_{obs}}{\nu_{old} + \nu_{obs}}$$

Where $\nu_{obs} = \frac{\nu_{com}+1}{\nu_{com}+3}\nu_{com}(1 - \frac{B+B/m}{T})$ And $\nu_{old} = \frac{m-1}{(\frac{B+B/m}{T})^2}$ Rubins rules assume normality, so if our statistic in mind is not asymptotically normal, we need to transform it towards normality before we pool. It should also be noted that we have discussed

the univariate case, but this easily extends to the multivariate case. We now have a powerful framework to get valid inference from multiply imputed data.

2.2 Survival analysis

Now that we have the multiple imputation datasets created, we may run our analysis. As a general rule of thumb, we should run our desired analyses on the available data first, to get an idea of what to expect. Because we are working with cancer data, we are interested in some basic survival quantities (Kaplan-Meier survival estimates for survival function, log rank test to test for similarity of curves, Cox regression to determine hazard ratio), as well as some more advanced ones (cumulative incidence for survival in the competing risks setting). Following Rubins rules, we run the individual analyses on each of the m datasets, and then pool our results.

Before we begin, it should be noted that there is another way to work with the multiply imputed data. If we take all of our imputed data and stack them to get one huge dataset of size m*i x j. We can call this the stacked method. Under the stacked method, we can produce unbiased estimates of quantities of interest, but the estimates of variance will be too small (since we are artificially increasing the sample size). The stacked method is useful when we want to observe just one plot instead of m for model checking. As well, the stacked method may be useful in situations where we dichotomize factor data on an imputed variable and then look at the percentage in each sub group. Under the normal MI scheme, we are not guaranteed that the percentages will sum to one, but under the stacked method we are.

Analysis for the Kaplan-Meier estimate is easy individually. All we need to do is clearly define what groups are, and what constitutes an event of interest. As well, we should verify that we are not in a competing risks situation. This setting will

be discussed later. Once we have checked all of these, we can run the Kaplan-Meier curve on each of the m datasets. Now, we could pool these estimates , but that would be ill advised, because the Kaplan-Meier curve is not normally distributed. To get around this, it has been proposed by Marshall et. al to take the complimentary log log transformation of the survival estimates before pooling [12]. We can make this transformation, and then pool our results. An interesting situation may arise where some of the survival curves may end before others. This is the result of a person with a long survival time being put in different groups through the imputation. We can deal with this by either extending the last observed Kaplan-Meier estimate out until the last time, or truncating all of the imputed curves at the minimum time of last event. *** figure out which one, although I think the latter**. Once we have our pooled estimates, we can back transform them using inverse of the complimentary log log transformation.

One of the main tasks that clinicians are interested in is the median survival time (the smallest time that the survival function is less than .5), specifically, the variance at the median. The median is a much better estimate of the typical survival time than the mean is, because in survival analysis, the time to event is typically right skewed, so the mean survival time is almost certainly not relevant to the typical patient/user of the km plot. We could get an unbiased estimate of the median via the stacked method, but this will give us false confidence, because our sample size is greatly inflated. Having a confidence interval for the median is actually very important, so we will need a method that allows us to do this. We can go about it in two ways. The first way is to pool the greenwood variance associated with each time point and then take the average as the variance at that time point, but this would lead to too big of changes between the data sets (by construction of the greenwood estimator).

The solution for this issue is to derive the variance of the median by the reflection method. In this method, we first fit the MI Kaplan Meier curve, and then construct a 95% confidence interval for all time points with the total variance obtained from Rubins rules pooling. The median is defined as the first time when the pooled MI curve crosses .5 survival line, and the lower and upper bounds are the points where the lower and upper bands cross the .5 survival line respectively.

Now, we will have a pooled estimate of the true survival curve. In the typical setting, we might want to look to see if these curves are similar to each other. We would do this with a log-rank test under the regular setting. However, we should not be deceived. We have an averaged survival curve. It is not constructed in the same way that a regular Kaplan-Meier curve is, so we cannot get the quantities that we would need to compute the log rank test. However, we can still get the pooled log rank test. To do so, we can do one of two things. The first is to run the logrank test on each of the datasets and then pool via Rubins rules. This is the logical way to do it, but the log rank statistic is not normally distributed, and no obvious transformation comes to mind. Another option is to run a cox regression on just the group in question. From the cox regression, we can obtain the score test, which is in fact the log rank test, but that again is chi square distributed. We know that the Wald test is asymptotically equivalent to the score test, so we can use the Wald test of the coefficient as a proxy for the log rank test. In this way, we get a quantity that is normally distributed, so we may use Rules.

We would now like to model the hazard ratio via the Cox proportional hazards model. The overall goal will be to fit a Cox model with baseline covariates, check to see if it passes the proportional hazards assumption, and then add in the treatment variables to see how they affect the hazard. It is known that the cox regression

coefficients are normally distributed, so there is no issue in pooling, but we do need to be careful about checking the proportional hazards assumption !! source or known?!! The very first thing that we need to do is check to check the available case model to assess if we have proportional hazards. If one of the covariates truly is dependent on time, adding imputed data isnt going to change that, so checking the available case analysis is a good sanity check. The way we go about checking to make sure that we have proportional hazards is checking to see if the schoenfeld residuals are correlated over time for each covariate. We can check a test for correlation or observe a spline fit to the residuals. In cancer research, the most common test to look for proportional hazards is to plot the spline fit to the residuals along with the 95% confidence intervals, and see if any straight line could pass through the bounds. There isnt an official name for this method, but the straight edge method seems to be a fitting name. If this is the case, then we say that that the covariate in question follows the proportional hazards assumption.

We can take our imputed data and fit a cox model on each of the m datasets, and pool them easily. But how is the best way to check the proportional hazards assumption. We can go about this in a few different ways. The first is to check the assumptions on each individual model fit to each dataset. This may prove to be an arduous task, but with graphical tools such as shiny, this isnt too bad. We could also superimpose all of the spline fits on one plot, and see how the shape and general trend compare to the available case analysis. We can also use the stack data to get just one set of plots, but the straight edge method will not work here since the errors are too low. Rather, we would just need to assess the shape of the spline fit in comparison to the available case method. Once we have verified that the model follows the proportional hazard assumption, we may trust its results. We can now add

in our treatment covariates, and see how they affect the hazards. The last thing that we might be interested in is the cause specific hazard, and the cumulative incidence function. Work on this a lot, and decide if I even want to put this in.

2.3 Propensity Score Analysis

Now that we have laid down the theory for analyzing the survival section for clinical relevance, we can move on to the causal analysis part. While there is a lot of preparatory work that goes into the theory of it, the results that can be obtained using causal analysis framework and propensity scores is much stronger than conventional analysis. As well, with causal analysis, we get a cause and effect result, which is in tune with what the general population believes that results should be. Propensity score methods are an easy to understand yet powerful tool. The use of propensity scores justified in Rosenbaum and Rubins 1983 paper [13]. Our overall goal is to estimate the average treatment effect in a setting where the initial study was not a completely randomized experiment. Propensity score analysis helps us to do this by balancing the groups out so that it becomes more like a randomized controlled experiment.

We will need to make a few decisions along the way. Our very first decision comes when deciding how to use the propensity score. There are four main uses in the survival literature: Matching, stratification, weighting, and covariate adjustment [14]. The goal is to balance the treatment and control groups, such that the only difference between the groups is due to the treatment received, and not any underlying factor. All of these methods will help us to examine the average causal effect, but each goes about it in a different way. Propensity score matching is the easiest to understand, yet it is still quite powerful, so it will be the method that we will use

Our next decision comes as to how to use matching in the multiple imputation

setting. The stacking method described before would obviously be inappropriate, as we would have spurious and repetitive matches due the falsely inflated sample size. Matching on the stacked set would give us much more power to detect a difference, but the results from it would not be valid.

There is hope though for matching with multiple imputation data. Two methods are described in Mitra and Reiter about how to do this [15]. In the first method, propensity score matching is done within each MI dataset (known as within matching). This, we will get m estimates of the average treatment effect to which we will average. The other method, known as the across method takes the average propensity score for each individual among the m imputed data sets, matches on the averaged propensity scores, and estimates the treatment effect in that manner. Both methods have their pros and cons, and are appropriate for different scenarios. However, as we will see in the applied example, the treatment variable may itself have missingness, and thus needs to be imputed. In this situation averaging across datasets does not make sense (since there is no guarantee that a given subject in imputation i has the same treatment in imputation j), so we must use the within method. Propensity scoring using logistic regression is the most widely used method, because it is computationally simple and is well understood by both clinicians and statisticians. Thus, we will use it to compute our propensity scores. Different propensity scores will be obtained according to what predictors we use in our model, so we need to be sure that we fit a model with clinically relevant and meaningful predictors. Our overarching goal is to account for any covariate that leads to the selection of the treatment.There has been significant debate among statisticians about how to set up these models, by either throwing in every possible variable into it, or only include ones known to affect treatment selection. I dont plan to settle this debate, but since we are in a setting

where simplicity is a goal, I plan to use only pretreatment covariates deemed to be important (through consultation with subject matter experts) to treatment selection. This is the method that Peter Austin (a well-known researcher of propensity score analysis) recommends [14].

There exists research on how to test if we have properly set up the propensity score model (do this if mao sends the code). It is not clear how this could be applied in the multiple imputation setting, but I think that if used on the available cases, it would give us a good indication that we are heading in the correct direction.

The matching can be done in many different ways like a nearest neighbor, caliper, or mahalanobis distance matching. It shouldnt really matter which one we use, so long as the groups are similar in distribution after matching in each dataset. Once an acceptable propensity score model is selected, we will run it on each of the m imputed datasets, match within each dataset, obtain the average causal effect, and then pool. We will also be interested in the reduction of the probability of death. To observe this, we will need to check it in each dataset and then pool.

Now we would like to put the propensity scores into our cox model. Doing so will allow us to assess the hazard ratio of covariates while controlling for group imbalances. Thus, the results ought to mimic the situation where the groups were chosen by random assignment. Although it has been shown that the results from using it in the cox model are biased, it can still be used as a useful explanatory tool [14] .We are still in the multiple imputation framework, so what we will need to do is to fit our models on each dataset, get the results, and then pool them via Rubins rules.

Chapter 3

Application

3.1 Data Explanation

We have now laid down the theory of what we want to do, so now we will put it in action. The data that we will apply it to is a dataset from M.D Anderson. The data is a collection of about 1500 patients at MD Anderson who have breast cancer that has metastasized to the brain, about 100 clinically relevant covariates, along with their survival status and time.

Unfortunately, as of October 5th, we are still waiting for IRB approval to use the data . I want to get the proposal out to yall though. So, what I will do in this section is describe what I will do, and once I get approval, I will fill in the tables and plots. In the meantime, I will put in ¡whatever¿ where something that uses the data should be.

Now that we have the theory in place, we can apply it to some real data. The dataset that I chose to analyze is a dataset from MD Anderson Cancer Center, with permission from Dr. Bugano, Dr. Ibrahim, Dr. Hess and ¡whoever else needs to be thanked¿. This dataset has historical records of X MD Anderson patients who have had breast cancer that has metastasized to the brain. There are lots of covariates recorded (about 90, with some missingness), a few different treatments, as well as survival endpoints (which are all observed). This data is exemplary for this task because it is large, survival amenable, has missingness and is a prime candidate for

imputation, and has treatment variables that are not given in an RCT. ¡plot or table of missingness¿

Our first step is to define what we would like to find. There are many interesting questions we could ask and answer from this data because of the amount of data available, but the question I will focus on here is the effect on survival and treatment of two HER2 (breast cancer grown protein) therapeutic drugs-Lapatinib and Trastuzumab. It isnt vital to understand what these drugs do, but the interested reader may want to look at appendix !! b?!! for a more detailed look at these. For a much more detailed analysis and other clinically relevant questions, see !!Hess, Bugano, Berliner!! This is the project that this research was forked off of, although it will probably not be published by the time this thesis is.

3.2 Imputation

We first need to impute the missing data. This is a little challenging just because of the sheer number of covariates that we have (about 90? with missingness). But we need these covariates, because they have the potential to be useful as predictors for other covariates, they might be something we are actually analyzing, we have spent the money to collect the data, and it strengthens the MAR assumption. As well, it is my opinion (and probably a consensus among applied statisticians) that is better to have too many covariates than not enough. After all, we can always use variable selection if we have too much data.

Our data is quite high dimensional, and there are a many binary variables, thus JM modeling seems inappropriate. Instead, FCS models seem better suited. We will be using the R package mice [9] because it is easy to use yet powerful.

The model is set up by hand. For each covariate with missingness, we need to

decide what method will be used for imputation, and what predictors will be used in it. I decided to be very forgiving, and use nearly every predictor for each missing covariate. I did this to bolster the MAR claim, and avoid variable selection. As well, the appropriate methods need to be selected for each datatype. The majority of the data is categorical, so decisions need to be made about whether to impute them via predictive mean matching or logistic regression. This decision was made by observing density plots after the algorithm was run to see if the imputations for each kind were valid. As well, derived variables were coded in as passive imputation, so that they would not be imputed, rather they would be computed.

After the model has been set up, we need to run it and save the results. For 50 datasets, 40 iterations, the algorithm runs in about X hours, and for 50 datasets with 100 iterations, it took Y hours on a computer with Z ram and Q processors. While this seems like a long time, this process only needs to be done once and requires no human interaction, so it can be run overnight and then never need to be touched again. For the rest of the analysis, I choose to use the 50 datasets, because there is hardly any confidence going from 50 to 100, and having such large objects in memory can be harder to work with.

Once the mice algorithm has run, we need to check for convergence and reliability. Convergence is assessed by looking at ¡plots of covariate mean and sd by iteration¿. According to van Buuren the different streams should be freely intermingled with each other, without showing any definite trends. Convergence is diagnosed when the variance between different sequences is no larger than the variance with each individual sequence [9]. Looking at these plots, this certainly seems to be the case. Other authors suggest using a more formal statistical tests such as ¡rhat¿ to assess convergence, so I also display that (values near 1.00 mean ok and values greater

than 1.1 indicate we should run longer). Diagnostic plots are viewed to ensure that the imputed data is similar enough to the real data.¡ A few of the plots have been replicated here¿. To see all of the plots, go to the shiny app/ R package (do this if enough time, also see about security. Might just need to make it be available upon request). As we can see, not all of the imputed data follows the distribution of the observed data exactly, but we obviously dont expect this to happen always. For the majority of the plots though, the data look like they could have been real data.

3.3 Survival Analysis

Now that the datasets are imputed, we are ready to run our models on them. As a sanity check, we may compare the fitted models to the available case analysis. Since the imputed values we generate ought to be quite similar to what data we have (unless there is reason to believe that the missing data is significantly different than the observed data), we should expect our estimates to be similar.

It should be noted that in all of our survival analyses, we will be doing a landmark analysis. Landmark analysis means that we dont start the analysis at time 0, rather, we start it at a different time. In Dr. Hesss words, Since the brain met treatment data was necessarily determined after the diagnosis of the brain met, it is not appropriate to use this data as baseline covariates in the analyses. Only covariates known at the time of diagnosis can be used in this fashion we can do a landmark analysis by estimating when the vast majority of patients would had their brain met treatment choices started and starting our analyses at this point. After speaking with subject matter experts (Dr. Bugano and Dr. Ibrahim), this landmark time was determined to be 2 months.

The first result that we will check is the Kaplan Meier curves for the imputed

data. The available case analysis shows that lapatinib and trastuzumab are quite close to each other, while having no her2 directed treatment being much lower. The logrank test statistic is X. The pooled KM estimate was found using Rubins rules, but under a complimentary log-log transform as suggested by [12] to get the survival curves towards normality. The results from MI look quite similar ¡AC analysis and MI analysis¿. We can also get an approximation for the log-rank test on the MI data via the Wald test on the pooled Cox model (NOT the Kaplan-Meier model, Kaplan Meier is not normally distributed, and no obvious transformation exists to make it so, so we must use Cox). We are not able to get the exact log-rank test because in doing so, we would need to compute either the likelihood ratio test or score test, both of which would include calculating the risk set, which is not possible in the MI case. Another suggestion is to pool the chi square statistics via methods presented in Marshall et al 2009, but even they say that this method is poor [12]. So, our only real option is to use the wald test (which is very easy to compute), and use that value as a proxy for the log rank test (they are asymptotically equivalent).

!!!Do I want to do competing risks analysis? I will if I have time!!!

Now that we have estimate of survival, we may set up a model to observe how changes in some baseline covariates change the hazard. We will do this with the Cox Proportional Hazards model. Once we have a baseline model fit and the assumptions met, we can add our treatment variable to see how this affects the hazard. The original available case model is as follows. We need to make sure that the proportional hazards assumption is met, so we may check the cox zph command to look at the schoenfeld residuals over time, and check the test stat. Overall, the assumption of proportional hazards over time seems reasonable, and the test statistic affirms this ¡AC cox.zph plots¿. Although the splines fit the points may not look straight, it certainly seems

reasonable that a straight line could be fit (denoting a hazard that is constant over time) between the 95

Now that we have verified that the available case analysis seems reasonable, we can work with the MI data. We fit that same cox model on all of our imputed data sets, and pool our results via Rubins rules (no transformation needs to be done since the Cox model coefficients are assumed to be asymptoticaly normal). We need to verify that we still have proportional hazards though. This is not an easy task, since we dont actually have a model, rather, we have the average of multiple models. We are no longer estimating the parameters by maximizing the partial likelihood; rather we are estimating them based on the average of the coefficients from the MI datasets. There are two ways we can go about this. The first is to check the proportional hazards assumptions on the stacked dataset. This will give us a good visualization about the shape of the splines fit to the residuals over time, but when running the chi square test to check for the correlation between the coefficient and time, the sample will be artificially too big, and thus we cannot trust the results. The correct way to do this is to observe each plot and statistic generated from the m datasets to see if the assumptions hold. This may seem like an arduous task when the number of imputed datasets is large, but we can circumvent it by writing a shiny app to view them, or ¡plot all of the splines on one plot¿. We can also look at all of the chi square tests for the 50 datasets and 13 variables for each, although there is bound to be some overlap between significance and non significance due to the multiple testing problem. Overall though, our imputed plots are very similar to the plots produced by complete case analysis, to which we have deemed to be acceptable for the proportional hazards assumption. We may now look at the cox regression coefficients and exponentiate them in order to obtain the hazard ratios. Looking at !! table whatever!! , we can see

that some factors force a larger hazard ratio than others. We can take the reciprocal of it to look at the protective effects of each covariate. We may then add in our treatment variable to see how it effects the hazard, and see how it changes other factors.

3.4 Causal analysis

Lastly, we will want to draw causal inference, and see what the average treatment effect of each drug is. This is necessary because the data was collected from a database, and we did not have a completely randomized experiment. As well, this piece of information is what clinicians and laypeople really wantit answers the question of which drug is better. There are many interesting questions that we may ask with this dataset, but here we will only focus on lapatinib vs trastuzumab vs no treatment. The interested reader may read !!my paper!! Upon its publication. The idea for this part of the analysis is to use propensity scores to match subjects and then compare them. As we saw earlier the best way to match is using the X method. There are several R packages to do propensity score matching in R, including X Y Z . I chose to use the X package because of its ease of use. Do a lot more work on this part!!!!!

Chapter 4

Discussion

We have discussed a number of tools and methods to analyze survival data with missingness. There are lots of decisions to be made along the way, and I am in no way advocating that my exact choices are the right ones, I am only claiming that the decisions made were proper for the type of data and questions that we had. There will certainly be many disagreements about the multiple imputation portion. And since the multiple imputation serves as the root of the analysis, the concerns should be addressed. The first concern comes from people who dont understand or believe in imputation of missing values. Multiple imputation is a tool to help us find plausible values for missing data. We will make no claim that the imputed values are right, but when used correctly, the results will be unbiased. We arent using multiple imputation to create data where there is none, rather we are using it to fill gaps in places that we already do have data. We actually need to impute in certain cases if we want to get valid results, as analysis without imputation will lead to severely biased results. For example, if teenage males who are obese dont want to self-report their weight, then classic available case analysis will yield biased results because we have knowingly left out part of the population who are systematically different [2]. We need to impute to make sure we have included all of the information and not to bias our estimate.

The next and more substantial critique will come from statisticians who may not believe that the distribution that the imputations is being drawn from is valid. Multiple imputation is inherently a parametric procedure. No matter what method

we use to impute, we have to make a parametric assumption, be it the joint model for JM or the full conditionals for FCS. For our case, using the normal model is certainly wrong because we have so many categorical and strictly positive variavles (which is proven to be suboptimal in [16]), so we are left only with using FCS. And FCS alone has weak theoretical justification. But as we have discussed before, many studies have shown that FCS is robust to non-compatibility. As well, there was no formal model validation (such as cross validation), only ad hoc checks. In the literature there is hardly any mention of validation, because if we were to cross validate, we would be drawing from different models, and comparison between the folds would be like comparing apples to oranges. We already have missing data, there is no reason to destabilize it to try to compare it, as the standard methods seem to work fine [2].

An interesting extension to this project would be to use a non parametric approach to multiple imputation, such as the one suggested by Long et all in [17]. But at the time of publication, there is not much literature or software on this subject, so I felt that it was not appropriate to use its results.

To summarize about multiple imputation, I would say that it is a necessary evil. In the process of using multiple imputation, we lose predictive power, and are forced to use a distribution that may not fit the data to a t. But we need to use imputation techniques if we want to make any sense of our data. The advice I would give to those who are hesitant to use multiple imputation would be to not have missing data, but this is a task that is easier said than done. Multiple imputation is becoming the standard for missing data techniques, especially in the medical field. There are lots of pros to it, but there are certainly some conns. Much research has already gone in to it, but much more needs to be done. It is my hope that this thesis has shown a powerful example of why multiple imputation should be used.

Next we can critique the survival section. We decided to use standard Kaplan-Meier and Cox analyses because they are very standard in practice, and answer the questions well. However, some lesser known methods could have been used. A popular theoretical model is called the accelerated failure time model, which describes how covariates affect the hazard, assuming that it acts in a multiplicative fashion. This is useful for clinicians, but not really good for patients, because the conclusions drawn from it are drug x will make you live 50

There are two concepts in survival analysis that I find interesting, but our data did not allow for it. The first is variable selection. The clinicians knew what they wanted to test, so this was not needed, but variable selection in the context of MI is an interesting question, and van Buuren covers it in his book [2]. This would be very useful if our dataset had covariates that we were unsure of their predictive power or wanted to examine. Another interesting addition would be using multistate data. In this setting, subjects can transfer from one group to another, ie have cancer, get in to remission, and then relapse. We model the states as a stochastic process. This would be really interesting, and I would have liked to implement it because I think it would have been interesting from a multiple imputation perspective, but unfortunately our data was not conducive to that. !!!Work on this a lot more!!! Lastly, we move on to the causal analysis part. While there are many other binary classifiers that could be used to make propensity scores, we chose to use logistic regression. This choice was based solely on tradition and ease of understanding from non statisticians. As well, there has been some new research recently saying that propensity score matching is not as powerful as it was once thought (? King?). Lastly, our choice of how to combine propensity scores was solely based off of the Mitra paper, and no more studies have been done to show that this is in fact the optimal way to do it.

Chapter 5

Conclusion

This paper details how to use multiply imputed data to answer survival analysis questions. The motivation for the methods used is cancer data, although sufficient detail is given so that the methods can be applied towards other areas. Along the way, we analyze and visualize the results, and discuss alternative methods of use. We test the methods out on a large cancer dataset, trying to draw meaningful inference from a dataset with substantial missingness.

Appendix A

Appendix

A.1 Missing data mechanisms

There are three mechanisms of missing data. It is important to understand what type of missing data we have so that we can use methods that are suited for that type. Before we begin, we will need some notation. It is not constant throughout the literature, so I caution you to look at the authors notation before reading any other literature. I will give the symbols I will be using along with words to describe them to make it easy to explain

- Y is our whole dataset. It will have i rows and j columns
- Y_j is a specific column of Y. Y_j is actually composed as $Y_j = (Y_{obs}, Y_{mis})$, where
  - Y_{obs} is the data we have observed
  - Y_{mis} is the missing data
- R is a binary matrix the same size as Y where a 1 indicates we observed the data, and 0 means it is missing ψ is a vector of parameters for the missing data model, and the missing data model is given as $p(R|Y_{obs}, Y_{mis}, \psi)$

As well, we have a concept called ignorability, which is defined as

$$p(Y_{mis}|Y_{obs}, R) = p(Y_{mis}|Y_{obs})$$

That is, we may ignore the R. The probability of the data being missing does not depend on how the data is missing. Equivalently, we may write this as

$$p(Y_{mis}|Y_{obs}, R = 1) = p(Y_{mis}|Y_{obs}, R = 0)$$

Being ignorable makes it justified to model our missing data from our observed data, without needing to worry about how it was missing. The opposite of ignorable data is called non-ignorable data, in this case,

$$p(Y_{mis}|Y_{obs}, R = 1) \neq p(Y_{mis}|Y_{obs}, R = 0)$$

So we must take into account the missing data structure for imputation. We often times see ignorable missing data in practice, although one should certainly check the sensibility of ignorability, as some instances will certainly be non-ignorable (like censored data, or when we know that the missing data is systematically different than the observed. !!! Need to work on this more!! Now, we may discuss the three main types of missing data mechanisms. I will give the technical definition, a laymans definition, and an example.

- MCAR: Missing completely at random: $P(R = 0|Y_{obs}, Y_{mis}, \psi) = P(R = 0|\psi)$. The missingness in the data is not at all related to any of the data that we do or dont have. If a lab technician drops 5 vials of blood, the missingness caused by this would be MCAR

- MAR: Missing at random: $p(R = 0|Y_{obs}, Y_{mis}, \psi) = p(R = 0|Y_{obs}, \psi)$. The missingness we have is related to something in the data. If we collect the gender of the subject and we know that males tend to not give blood, we can attribute the missingness to the gender

- MNAR: Missing not at random $p(R = 0|Y_{obs}, Y_{mis}, \psi)$. We cannot get simplification, and the missingness depends on data that we have as well as have not collected. For example if a full moon causes the blood testing machine to break more often, but we dont have the moon phase as a variable.

A.2 HER2 and Cancer Drugs

Put some stuff about HER2 and cancer drugs here later.

Bibliography

[1] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*. No. JOHN WILEY & SONS, 1987.

[2] S. Van Buuren, *Flexible Imputation of Missing Data*. 2012.

[3] J. Klein and M. Moeschberger, *Techniques for Censored and Truncated Data*, vol. 19. 1984.

[4] J. Angrist and J. Pischke, *Mostly harmless econometrics: An empiricist's companion*. No. March, 2008.

[5] J. Honaker, G. King, and M. Blackwell, "AMELIA II : A Program for Missing Data," *Journal Of Statistical Software*, vol. 45, no. 7, pp. 1–54, 2011.

[6] A. A. Novo and J. L. Schafer, "Package norm ," *CRAN*, 2015.

[7] F. Tusell, "Package cat ," *CRAN*, p. 23, 2015.

[8] S. Van Buuren, J. P. Brand, C. G. Groothuis-Oudshoorn, and D. Rubin, "Fully conditional specification in multivariate imputation," *Journal of Statistical Computation and Simulation*, vol. 76, no. 12, pp. 1049–1064, 2006.

[9] S. Van Buuren and K. Groothuis-Oudshoorn, "Multivariate Imputation by Chained Equations," *Journal Of Statistical Software*, vol. 45, no. 3, pp. 1–67, 2011.

[10] Y.-S. Su, A. Gelman, J. Hill, and M. Yajima, "Multiple Imputation with Diagnostics (mi) in R: Opening Windows into the Black Box," *Journal of Statistical Software*, vol. 45, no. 2, pp. 1–31, 2011.

[11] Y. Zhao, A. H. Herring, H. Zhou, M. W. Ali, and G. G. Koch, "ANALYSES OF TIME-TO-EVENT DATA WITH POSSIBLY," vol. 24, no. 2, pp. 229–253, 2014.

[12] A. Marshall, D. G. Altman, R. L. Holder, and P. Royston, "Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines.," *BMC medical research methodology*, vol. 9, p. 57, 2009.

[13] P. Rosenbaum and D. Rubin, "The Central Role of the Propensity Score in Observational Studies for Causal Effects," vol. 70, no. 1, pp. 41–55, 1983.

[14] P. C. Austin, "The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments," *Statistics in Medicine*, vol. 33, no. 7, pp. 1242–1258, 2014.

[15] R. Mitra and J. P. Reiter, "A comparison of two methods of estimating propensity scores after multiple imputation," *Statistical Methods in Medical Research*, pp. 1–17, 2012.

[16] J. Kropko, B. Goodrich, A. Gelman, and J. Hill, "Multiple Imputation for Continuous and Categorical Data: Comparing Joint Multivariate Normal and Conditional Approaches," *Political Analysis*, pp. 497–519, 2014.

[17] Q. Long, C.-H. Hsu, and Y. Li, "Doubly robust nonparametric multiple imputation for ignorable missing data," *Statistica Sinica*, vol. 22, no. 1, pp. 1–22, 2012.