RICE UNIVERSITY

Using Multiple Imputation, Survival Analysis, And Propensity Score Analysis In Cancer Data With Missingness

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 of Statistics

David Scott
Professor of Statistics

Houston, Texas
December, 2015

ABSTRACT

Using Multiple Imputation, Survival Analysis, And Propensity Score Analysis In Cancer Data With Missingness

by

Nathan Karmazin Berliner

In this thesis multiple imputation, survival analysis, and propensity score analysis are combined in order to answer questions about treatment efficacy in cancer data with missingness. While each of these fields have been studied individually, there has been little work and analysis on using all three together. Starting with an incomplete dataset, the goal is to impute the missing data, and then run survival and propensity score analysis on each of the imputed datasets to answer clinically relevant questions. Along the way, many theoretical and analytical decisions are made and justified. The methodology is then applied to an observational cancer survival dataset of patients who have brain metastases from breast cancer to determine the effectiveness of chemotherapeutic and HER2-directed therapies.

Contents

Abstract ii

List of Illustrations v

List of Tables vi

1 Introduction and Background Information 1

1.1 Motivation 1

1.2 Imputation 2

1.3 Survival 5

1.4 Causal Analysis 9

2 Methods 14

2.1 Multiple Imputation 14

2.1.1 Selecting the MI scheme 15

2.1.2 Setting and checking the model 20

2.1.3 Combining the MI estimates 22

2.2 Survival analysis 25

2.2.1 Kaplan-Meier Survival Curve 25

2.2.2 Median Survival Time 26

2.2.3 Log Rank Test 27

2.2.4 Cox Proportional Hazards Model 28

2.3 Propensity Score Analysis 30

3 Application 35

3.1 Data Explanation 35

3.2 Imputation 38
3.3 Survival Analysis 43
3.4 Causal Analysis 54

4 Discussion **58**

5 Conclusion **62**

A Appendix **63**
A.1 Missing data mechanisms 63
A.2 Cancer and Treatment Overview 66

Bibliography **68**

Illustrations

1.1 Visualization of MI data 5

2.1 Normal JM imputation pseudocode 16
2.2 MICE FCS imputation pseudocode 18

3.1 Visualization of missingness in the cancer dataset 38
3.2 Selected plots of continuous variable imputation mean and standard deviation by iteration 43
3.3 Selected plots of binary variable imputation mean and standard deviation by iteration 44
3.4 Landmarked Kaplan-Meier curves for chemotherapeutic drugs, AC and MI data 47
3.5 Landmarked Kaplan-Meier curves for HER2 drugs, AC and MI data . 48
3.6 AC Schoenfeld residuals over time 50
3.7 MI Schoenfeld residuals over time 51

Tables

3.1 Data Categories and Examples 36
3.2 Table of important covariates to be used in the analysis 37
3.3 Characteristics of available case data versus MI data 45
3.4 Chemo log rank tests 49
3.5 HER2 log rank tests 49
3.6 AC and MI baseline Cox model 52
3.7 AC and MI Cox model with chemo treatment 53
3.8 AC and MI Cox model with HER2 directed treatment, on HER2+ subjects 54

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 and causal inference from imputed data. While all three fields (imputation, survival, causal) 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 this thesis is motivated by cancer data, I believe that the methods used in this thesis are general enough to be applied to other types of data and situations.

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 new research is seldom produced. Propensity score analysis will help us determine causal relationships when we don't have a randomized controlled experiment. As one could imagine, all three of these fields are important to the applied statistician, as they will come across issues in each discipline 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 separately 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. To be specific, missing data is when a subjects data for a collected covariate was not obtained. Missing data is not when a covariate is never collected, and then suddenly the researcher wishes he had collected it. Imputation (specifically multiple imputation) is a way to "fill in missing data" with plausible values, and it forms the base of this thesis. All of the other analyses that will be used will follow from it, thus we need a solid understanding of it before we may proceed. Imputation itself has been around since the 1930's [1], but multiple imputation is a recent development, proposed in the 1970's and formalized in 1987 by Donald Rubin [2]. 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 paid no attention to missing data, and happily discarded records from their data that were incomplete. This procedure is known as complete case (CC) analysis. There are many problems with this paradigm. To begin with, you will lose a lot of statistical power when, because you are throwing away records and thus decreasing your sample size. In addition, this can be costly (in terms of time or money) to the researcher, because they had to actually obtain these records that are not used. Lastly, and most importantly, we will be biasing our estimates if we discard them. For example, suppose we have a random sample of people and are testing a drugs efficacy, and want to run a regression on some collected covariates. If men are known to not want to give all of their information, in the analysis, we will need to discard the male samples because they are incomplete, leaving us only with women.

Thus, we no longer have a random sample, and will get biased results because we have knowingly thrown away half of our data which we know to be different [1].

A slight improvement on this is called available case (AC) 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 between different analyses.

The next wave of statisticians in the early 20^{th} century wanted to improve upon available case analysis, so they developed what we now call today (single) imputation. Their goal was to fill in missing values with a single plausible replacement value. A single method (such as regression, taking the mean, resampling, etc.) is used one time to impute or fill in the missing value, and to account for the uncertainty, degrees of freedom are deducted in the following analysis. 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. 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 completely different results from analysis on the data. This is obviously not a desirable trait. In addition, one single imputed dataset 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 (MI) began in the 1970's, but it wasn't until 1987 when Donald Rubin formalized multiple imputation methodology in his seminal book *Multiple Imputation for Nonresponse in Surveys* did it start to gain acceptance [2]. 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 m substitute values,we can think of the data now as being m datasets, each dataset containing the observed data, and one value for each piece of missing data.

An example will make the MI process clear. Suppose that we had a dataset of age, weight, and height. We want to regress age on weight but we have missingness. First, we will impute our data (figure 1.1, the first two columns). Once we have a sufficient number of datasets (we will talk about how to pick the number later), we can run the regression analysis on each of the MI datasets, treating the dataset as if it was complete (horizontal lines and third column). After running the model on the m datasets, we can pool the results to get one single estimate with its associated variance (last column). We will speak of the exact details of this later.

The MI method is obviously much better than the first two methods because is permits us to keep the data we already have, as well as to quantify our uncertainty about imputing the missing values.

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

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. We then regress age on weight, get the results from the individual datasets, and then pool them together.

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 [1]. Thus, using multiple imputation to handle missing data will be advised because of its popularity and strength.

1.3 Survival

Survival analysis is a large and important field in statistics, 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 refrence [3].

Survival analysis on the whole can generally be described as the analysis of time

to event data, often in the presence of censoring or truncation (when we don't have complete information about the time of event). 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 software (and in this paper) we see terms like "death/failure" and "survivors". This is due to survival analysis being heavily influenced and intertwined with medical studies. A more general term for these would be "event" and "those who have not had an event yet". These terms are used because they are clear and concise, although it might not accurately describe the event at hand. For example, if we were tracking the time until a child loses all of their baby teeth, the term death would obviously not portray the event of interest, but may be used in the context to denote losing all of the teeth.

The Kaplan-Meier (KM) estimator is a nonparametric estimate of the true survival function (the probability of survival after time t, $S(t) = P(T > t) = \int_t^\infty f(u)du$, where $f(u)$ is the unknown probability density function). It is defined as

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

Where n_i is the risk set, defined as the number of people who have not had the event or been censored right before time t_i ,and d_i is the number of deaths observed at time t_i [4]. The Kaplan-Meier estimator is very commonly used as a measure to see how different treatments affect the survival of the population in question, and is helpful in seeing at what time points survival changes the most (i.e. early or late).

Knowing the shape of the Kaplan-Meier curve is interesting, but it would be wise to have a statistical test to compare two curves. 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 [5]. 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 whether 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".

The general log rank test is defined as:

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

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 subjects in the risk set at time j, composed from the sum of the number of deaths at time j in each group, $O_j = O_{1j} + O_{2j}$ 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 $V_j = \frac{O_j(N_{1j}/N_j)(1-N_{1j}/N_j)(N_j 0_j)}{N_j - 1}$

Typically, all of the weights are set to one, as this test 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 takes a long time to start working. We wouldn't care about early deaths, only about later times when we are comparing the survival. Putting more weight on the later deaths would help to answer this question better. It can be proven that the log rank test is equivalent to the score test on a Cox model (which will be discussed next) fit the same data with no tied event times, and very similar when there are ties [3].

Proportional hazards regression, often called Cox regression or Cox model 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 function 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:

$$\lambda(t) = \lim_{\Delta t \to 0+} \frac{P[t \leq T < t + \Delta t | t \leq T]}{\Delta t}$$

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

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

. where $h_0(t)$ is what's known as the baseline hazard, and can be any positive function , and often times a parametric model like the Weibull is chosen. Note how it only depends on time and not any covariates. Z is a vector of observed covariates, and does not depend on time. The βs are found by maximizing the partial likelihood function

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

Where $Z_{(i)k}$ is subject i's kth covariate, $R(t_j)$ is the risk set (set of those who have not died yet at the time just prior to t_j), D is the number of distinct death times and p is the total number of covariates in the model. The betas are maximized by the Newton-Raphson method [6].

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 individuals with different covariates. Often times, the interest lies in what happens when all covariates are held constant, and the covariate of interest is increased one unit. This ratio will be a constant, and should not vary over time; hence the name proportional hazards. This is so because the ratio does not depend on the baseline hazard (which cancels out when taking the ratio). Using Cox regression, statements such as "Increasing the drug by one *mg* will decrease the rate of death (compared to non-users) by 30%". Cox modelling is one of the most used models in survival and medical literature.

1.4 Causal Analysis

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 with B". However, the only way to get this interpretation is if we conduct a randomized controlled trial (RCT). Although we can analyze any observational data using survival analysis, unless a randomized controlled trial is conducted, we cannot make any claims about causality. In order to prove causality, we need experimental data in a randomized and controlled setting, not observational data. The reason for this is because in an RCT, we expect the groups to be similar at baseline, and thus any differences after treatment should only be related to the treatment. However, in an observational study (where treatment assignment is not random, and may even by chosen by the participant), we have no reason to believe the groups are similar at baseline, and the difference after treatment could be due to the treatment or something else.

Randomized controlled trial (RCT) is a term that is often thrown around, but I want to be precise with its definition. An RCT is a study design that "randomly assigns participants into an experimental group or a control group. As the study is conducted, the only expected difference between the control and experimental groups in an RCT is the outcome variable being studied" [7]. This is in stark contrast to a retrospective study of observational data, where historic data of people who chose what group/ treatment they wanted to be in are studied. When making judgment on a retrospective study, we cannot be sure if the differences between the groups are due to their treatment choice, or some other factor. RCT's are the gold standard for experiments, and should be used if possible when trying to study causality. But often times monetary, ethical, and other factors prevent us from doing so. In this case, the best data we may be able to get is retrospective observational study.

Luckily, we can still analyze observational data for causality. We can frame our problem in a framework known as the Rubin Causal model, which helps get causal understanding from non-experimental settings [8]. Rubins causal model is built upon the idea of a counterfactual, also known as a potential outcome. Put simply, the counterfactual is the result that we would have observed had the subject been in the other group. For example, if we were testing a weight loss drug versus a diet in a study to see which lead to more weight loss, if subject 1 took the drug, then his observed value would be the weight lost on the drug, whereas his counterfactual would be the weight lost had he dieted. Together, the observed value and the counterfactual what is known as the potential outcomes. For the i^{th} subject, the potential outcomes are denoted as $Y_i(0), Y_i(1)$, where the 0 means treatment and 1 means control. Often times, authors give the observed value concisely as $Y_i = Y_i(1)T + Y_0(1-T)$, where T is the treatment indicator (1 means treatment, 0 means control). In an ideal world, we would observe both potential outcomes, however, this is obviously not possible, since we can only observe one outcome. This issue is what is known as the fundamental problem of causal inference.

Rubins causal model framework relies on two assumptions. The first is called the stable unit treatment value assumption (SUTVA), which states that the potential outcome for the i^{th} subject will be the same no matter what treatment the other subjects receive. The second is called ignorable treatment assignment (sometimes called no unmeasured confounders), which states that the potential outcome should be independent of the treatment assignment given the confounding factors. Put formally, this condition is $(Y_0, Y_1) \perp T|X$, where X are the pretreatment covariates confounding the potential outcomes and the treatment assignment [8]. This assumption is untestable, but can be reasonably assumed in most cases. This assumption is satis-

fied in an RCT (because knowing treatment assignment doesnt give any information about the potential outcomes). If we are able to meet these assumptions, then using Rubins causal model will help us make causal inference from non-RCT data, as we will discuss.

If we are able to get the counterfactual, we could compute the treatment effect of the treatment for each individual. However, this is not of much interest to us. Rather, we are interested in the treatment effect on the entire population, not the individual. There are two common measures of the causal effect in the population. The average treatment effect (ATE), defined as how the treatment effect changes when the population moves between treatment groups. It is given mathematically by $E[Y(1) - Y(0)]$, and the average effect on the treated (ATT), which is the treatment effect of one group moving to another treatment is given by $E[Y(1|T = 1) - Y(0|T = 1)]$.

It should be noted that unless we have ignorability $E[Y(1)|T = 1] \neq E[Y(1)]$, because $E[Y|T = 1] = E[Y_1T + Y_0(1 - T)|T = 1] = E[Y_1|T = 1] \neq E[Y(1)]$. And the same holds for $Y(0)$, such that $E[Y(0)|T = 0] \neq E[Y(0)]$. If we were to use $E[Y(1)|T = 1], E[Y(0)|T = 0]$ in our calculations, we would get biased estimation of the treatment effect. What this means is that if we are not in the causal frame work, then we cannot split the expectation up and just take the mean from each group to get the causal effect. However, if we are in the causal framework, and we assume ignorability, then $E[Y(1)|T = 1, X] = E[Y(1)]$ because the potential outcome is independent of the treatment assignment, and the same holds for the control group. If the potential outcomes were independent of the treatment assignment, we could break apart the expectations and calculate the estimands easily to get an unbiased estimate.

The goal of randomization in RCTs is to make the groups be as similar as possible at baseline, that is, the distribution of the pretreatment covariates should not be too different. However, in an observational study, the groups may be very different. It is quite likely that some baseline factor might influence how somebody picked their treatment. In the weight loss pill example, healthier people may choose the diet over the pill, whereas people motivated to lose weight might be inclined to take the pill, so the participants in the two groups will not be similar to start with. We should try to make them more similar (i.e. balance) if we wish to emulate an RCT.

In order to implement balancing in Rubin's causal model, *propensity score analysis* is used. The propensity score is the probability that the subject received the treatment given the covariates that are believed to be confounded with treatment choice. The propensity score for the i^{th} subject is given by $\hat{e}_i(X) = p(T_i = 1|X_i = x)$ [9], and is can be found by any method that gives group membership probabilities. We assume that some of the covariates play a role in deciding how the patient chooses the group, so controlling for this makes all of the patients seem similar at baseline. Because of the propensity score theorem, if we have ignorability, then we say that $(Y_0, Y_1) \perp Z|e(X)$ [10]. With the ignorability assumption, we must condition on a vector to get independence, however, the appeal of the propensity score is that we can get the same results by only conditioning on a scalar quantity, the propensity score.

Propensity scores lead to group balancing, that is, if propensity scores are controlled, then our groups will have the same distribution at the baseline (i.e. propensity scores remove the confounding between pretreatment characteristics and treatment). And thus, we can treat the data like it was an RCT. The most popular propensity score methods are propensity score matching (where we match those who picked the treatment to those who did not based on propensity score similarity) and propensity

score weighting (where each individual's contribution to the average treatment effect is dependent on the inverse of their propensity score). Propensity score matching has fallen out of favor recently because there will be some data thrown away, and the balancing that follows might not be that accurate [11]. However, propensity score weighting is still a popular method today [12].

Getting the propensity score was historically done by logistic regression, although recently newer machine learning methods have been employed. Any method that will lead to probability of treatment or control membership will suffice. There has been a lot of debate as to which covariates to include in the propensity score; either all variables, or only relevant variables, but I don't aim to settle this debate in this paper.

This paper is interested in weighting, specifically, inverse probability of treatment weight (IPTW). With IPTW, the sample is reweighted so that at baseline, the groups appear similar. These weights are given by

$$w_i = \frac{T_i}{\hat{e}_i(X)} + \frac{1 - T_i}{1 - \hat{e}_i(X)}$$

and thus the weighted response will be $\frac{TY}{\hat{e}(X)}$ for the treated and $\frac{(1-T)Y)}{(1-\hat{e}(X))}$ for the untreated.

The idea of IPTW weighting is to reweight the sample so that we get a population where there is no confounding, and the weighted averages reflect the true population averages. Its justification is due to the fact that it can be proven that $E\frac{TY}{\hat{(e)}(X)} = E[Y(1)]$ and $E\frac{(1-T)Y}{(1-\hat{(e)}(X))} = E[Y(0)]$ [12]. Because of this, the IPTW outcome from each of the groups can be treated as if it has been obtained from randomization, and we may treat it as such.

Chapter 2

Methods

I want the framework and methods used to be easy to use and understand, so that it can be discussed among clinicians and other people who don't 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. There are many different theories and implementations to choose from for the three topics covered in this thesis. I aim to pick the ones that optimize ease of understanding and power of results, with the motivating example being cancer survival data. Throughout this section, decisions must be made as to what methods and analyses to use. Whenever a decision is made, I explain why it was chosen as well as alternative methods that could also be used. It is my hope that I will provide enough clarity and detail so that the interested reader can intelligently apply these methods to their own data, even if the choice of methods is different than this thesis.

2.1 Multiple Imputation

It should be clear that multiple imputation is the preferred method to deal with missing data, so our first decision comes as to what paradigm to impute under. It should be noted that as long as valid imputations are produced, the choice of method to generate the imputations do 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 poor imputations will lead to poor results be it bias, high variability, or loss in statistical power.

2.1.1 Selecting the MI scheme

There are two main divisions in modern multiple imputation: 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. I suggest that everyone reads appendix A.1 (even if you are familiar with MI), so that the concepts and symbols that will be used in this paper are understood.

The first main imputation strategy is called joint modelling (JM). In JM, we assume that the missing data mechanism is ignorable and that the data can be described by a multivariate distribution specified by the user on the rows (missing data pattern) of the data. Then, we run a sampler that draws imputations from the specified model, and updates model parameters. Since the true model parameters are unknown, they need to be estimated. This is often done by a data augmentation algorithm [1]. A pseudocode example will help better clarify the steps, as can be seen in figure 2.1, where the assumed model is a multivariate normal.

There has been extensive programming and research on using the normal model for this, and research shows that it even performs well under situations where the data has strong non-normality [15]. Some research has been done for other types of

1. Sort the rows of Y into S missing data patterns $Y_{[s]}, s = 1, \ldots, S$.
2. Initialize $\theta^0 = (\mu^0, \Sigma^0)$ by a reasonable starting value.
3. Repeat for $t = 1, \ldots, T$:
4. Repeat for $s = 1, \ldots, S$:
5. Calculate parameters $\dot{\phi}_s = \text{SWP}(\dot{\theta}^{t-1}, s)$ by sweeping the predictors of pattern s out of $\dot{\theta}^{t-1}$.
6. Calculate p_s as the number missing data in pattern s. Calculate $o_s = p - p_s$.
7. Calculate the Choleski decomposition C_s of the $p_s \times p_s$ submatrix of $\dot{\phi}_s$ corresponding to the missing data in pattern s.
8. Draw a random vector $z \sim N(0, 1)$ of length p_s.
9. Take $\dot{\beta}_s$ as the $o_s \times p_s$ submatrix of $\dot{\phi}_s$ of regression weights.
10. Calculate imputations $\dot{Y}^t_{[s]} = Y^{\text{obs}}_{[s]}\dot{\beta}_s + C'_s z$, where $Y^{\text{obs}}_{[s]}$ is the observed data in pattern s.
11. End repeat s.
12. Draw $\dot{\theta}^t = (\dot{\mu}, \dot{\Sigma})$ from the normal inverted-Wishart distribution according to Schafer (1997, p. 184).
13. End repeat t.

Figure 2.1 : Normal JM imputation pseudocode

Taken from van Buuren's book Flexible imputation of missing data [1]

models, but by and large the normal is the most popular.

This is just one implementation of a JM approach. Another, like that used in the Amelia package, uses an EM algorithm with user specified priors to make draws [17]. Other distributions can be used, but the user will have to specify it, derive any relevant distributions, program it, and research to see if the method provides optimal properties, such as proper coverage and correct estimation of the parameters of interest.

An obvious issue arises when the data is categorical (either alone or mixed in with continuous). There has been much debate in the literature about how to handle this situation. Some authors argue that you should just impute under a continuous distribution and round imputations to the nearest class number, and others suggest using distributions that are more suited for categorical data [1]. A few of the most used R packages for joint modelling imputation include Amelia [17], norm [18] and cat [19].

It is my opinion that unless the user is very confident in the multivariate joint distribution, that JM should not be used. In the cancer example, there are many categorical variables and strictly positive variables to impute, so JM seems inappropriate.

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/covariates based off of the specification of the imputation model for each covariate, given by the user. Whereas JM imputes on the rows, FCS imputes on the columns. This theory goes by many names, including partially compatible MCMC, iterated univariate imputation, and chained equations [1]. These full conditionals should factor to specify the joint distribution. In the JM setting, the user must give a k dimensional model, however in the FCS setting, the user must give k one dimensional models. The goal is to sample from

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

By sampling from the full conditionals

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

In this notation, Y_{-j} means all of the columns with missing data except for j, and

Y is the fully observed columns (which could possibly be empty). A pseudocode example can be seen in figure 2.2. This method is called Multiple Imputation via Chained Equations (MICE), and draws from the posterior predictive density of the missing covariate. Note that the previous imputations enter the current imputation only through the relation with the other variables, and not directly. Because of this, convergence is often seen within 5 iterations [20].

1. Specify an imputation model $P(Y_j^{\text{mis}}|Y_j^{\text{obs}}, Y_{-j}, R)$ for variable Y_j with $j = 1, \ldots, p$.
2. For each j, fill in starting imputations $\dot{Y}_j^0$ by random draws from Y_j^{obs}.
3. Repeat for $t = 1, \ldots, T$:
4. Repeat for $j = 1, \ldots, p$:
5. Define $\dot{Y}_{-j}^t = (\dot{Y}_1^t, \ldots, \dot{Y}_{j-1}^t, \dot{Y}_{j+1}^{t-1}, \ldots, \dot{Y}_p^{t-1})$ as the currently complete data except Y_j.
6. Draw $\dot{\phi}_j^t \sim P(\phi_j^t|Y_j^{\text{obs}}, \dot{Y}_{-j}^t, R)$.
7. Draw imputations $\dot{Y}_j^t \sim P(Y_j^{\text{mis}}|Y_j^{\text{obs}}, \dot{Y}_{-j}^t, R, \dot{\phi}_j^t)$.
8. End repeat j.
9. End repeat t.

Figure 2.2 : MICE FCS imputation pseudocode

Taken from van Buuren's book Flexible imputation of missing data [1]

One of the major criticisms 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 multiple studies have shown that even when the models are highly incompatible, FCS methods are very robust and produce

proper imputations [21]. FCS allows us much more flexibility than JM does, and it handles discrete and categorical data much better than JM does [16]. Some popular R implementations of FCS methods include MICE [20], mi [22], and BaBooN [23], which go about drawing from the posteriors in different ways. These methods are good, but I will be using MICE in the applied section of this paper.

The user is going to have to specify something, there is no escaping that, but it is easier for the average person to be able to define a single distribution and model rather than to guess at a multivariate distribution, particularly if it is high dimensional. In addition, in the survival analysis setting, there will naturally be strictly positive and categorical variables. 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 full conditional specification certainly seems like the more appealing option.

In order to use FCS methods, the missingness in our data must be MCAR or MAR. It can work with MNAR missingness, but it requires some extra modelling assumptions that are beyond the scope of this thesis. T tests have been proposed to test if the data is MAR or MCAR, but this is of little use for us, because we only want to know if the data is MNAR, which is impossible to test since testing for MNAR would entail us using information that is impossible to get [24]. Luckily, we can safely assume MAR if there is reason to believe that some of the covariates collected account for the missingness [1].

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 or outcomes may follow a different

distribution than the observed times. This is cleared up by Zhao et. al in 2014 through Kaplan-Meier MI [25]. This is beyond the scope of this report though so I omit its details.

2.1.2 Setting and checking the model

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 method to use (regression, predictive mean matching, logistic regression, trees, etc.), and then what predictors will go into those models. We should choose predictor variables that help explain missingness, as well as those we are doing inference on, as to avoid bias [1]. For variables that are derived from others, we impute its components and then compute that variable, in a process known as passive imputation. Since our data is of manageable size, I include any reasonable predictor that doesn't induce collinearity for predictors in the imputation models. With large datasets, we may need to perform variable selection before specifying the models, so that it doesn't become unmanageable.

Since FCS is an iterative process, we must choose how many iterations to do until convergence. The older literature suggests 5 iterations are enough, but with modern computation, we can easily exceed this, even with large data [1]. A good way to assess how many iterations to run is to look at the diagnostic plots (which will be talked about later), then add 5 iterations to the number at which we assume convergence is achieved. Due to the nature of FCS methods, convergence is often very quick, often as soon as 5 iterations. As well, we need to decide how many datasets to impute. The early literature argued that 5 would suffice, but modern literature argues for more, since it will cut down on simulation error. Many different authors have different

criteria, but a popular criterion is to impute as many datasets as the 100 times the percentage of cases in the analysis with missingness. With the speed of computers and availability of storage, many authors now suggest using more imputations [1].

Once we have determined how many iterations and imputations to run, the FCS algorithm is ran. Depending on the model specifications, it should not take too long for small datasets (seconds or minutes), but may take hours for larger ones. FCS models often converge quickly, so after convergence, we are just taking draws from the missing covariates posterior. The first few iterations may be considered as burn in, and the rest as samples. The value at the last iteration of each chain is taken as the sample from the posterior that will be used as the value for the missing data.

We need to verify that our imputations are valid once we complete them. First, we need to see if our chains have converged. Since FCS methods are MCMC method, we should check the chain for irreducibility, aperiodicity, and recurrence. To determine when the chain has converged, van Buuren suggests that "Convergence is diagnosed when the variance between different sequences is no larger than the variance within each individual sequence" [1]. There is some research (but not from the MI perspective) about tests to check for convergence, and a popular test is Gelman and Rubin's $\hat{R}$ scale reduction factor test [26]. This test is often used in conjunction with visual tests. Assessing convergence by looking at of all of the values over m imputations and k iterations would be very hard to visualize because there would be so many chains, so often times, we will choose to observe a statistic (like the mean) of the chain and assess on that.

Once convergence is assessed, we need to check that the values imputed are valid and come from the correct posterior. This will serve as our model checking and validation. 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. This is a visual task and there is no statistical method to validate this.

This whole process can be very time consuming because every time we want to make a change in the methods used, we have to rerun the algorithm and reassess our results. But once we find the setup that works for us, we don't need to repeat it again. So, while it may take a lot of time now, setting up a proper model will save us even more time in the future.

2.1.3 Combining the MI estimates

Once the m datasets are imputed, we may run any valid analysis on each imputed dataset *individually*, treating each of the m datasets as if it was complete. I am a strong advocate of running the model with the available cases if possible before running it on the MI data, so that we can assess the appropriateness of the model. As well, running the model with the available cases will give us a clue as to what to expect from the MI analyses (such as sign of the coefficients or level of significance).

On the m imputed data sets, we may then apply Rubin's rules [2] to pool our estimates. Rubin's rules are essential for using multiply imputed datasets, so we need to investigate them thoroughly.

Rubin's rules are a set of rules that guide us in making inference from multiply imputed data. They will give us a single point estimate for the scientific estimand in mind, as well as the proper variance for it from the m imputed datasets.

Rubin's rules involve three parts. The first is getting an estimate of the population estimand Q, estimated with the MI datasets by $\bar{Q}$. To get $\bar{Q}$, define $\hat{Q}_i$ as the estimand evaluated from the data in the i^{th} dataset. Then, take the average over all m datasets

to get a single estimate $\bar{Q}$.

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

The estimates are not set, and there is variance associated with them. The first form of variance is the "within" variance, or the variance of each estimate $\hat{Q}_i$. Define the variance of the i^{th} imputed dataset as $\bar{U}_i$, then the overall within variance is computed as

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

The last 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

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

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 [2].

The theory of inference with Rubin's rules is rooted in the assumption that under repeated sampling, the complete data quantity of interest is asymptotically normally distributed with mean Q and variance U, where U is the variance of $(Q - \hat{Q})$ [1]. We don't have the true population variance, so we must use what we have from the sample, namely T, the MI total variance. With this assumption, we know that

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

And the degrees of freedom is proven to be

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

Where $\nu_{obs} = \frac{\nu_{com}+1}{\nu_{com}+3}\nu_{com}(1 - \frac{B+B/m}{T})$, ν_{com} is the hypothetical complete sample degrees of freedom, and $\nu_{old} = \frac{m-1}{(\frac{B+B/m}{T})^2}$ [27].

Rubin's rules assume normality, so if our statistic in mind is not asymptotically normal, we need to transform it towards normality before pooling. There have been some research about how to pool non normal quantities, but current research shows poor results and power when doing so [28]. It should also be noted that we have discussed the univariate case, but this easily extends the case with multivariate estimands. If we have multivariate estimands, we can still use Rubin's rules, but will need to replace the notation above with vectors for $\hat{Q}_i$ and variance-covariance matrices wherever we have variance. We now have a powerful framework to get valid inference from data with missingness via multiple imputation multiply imputed data, so let's now see how to fit in other types of analyses in the MI paradigm.

It should be noted that there is another way to work with the multiply imputed data. This method is colloquially called the stacked method. For the stacked method, we take all of our imputed data and stack them one on top of each other to get one huge dataset of size $(m * i)$ rows and j columns. Under the stacked method, unbiased estimates of quantities of interest can be produced, but the estimates of variance will be too small (since we are artificially increasing the sample size) [1]. Thus, the stacked method is a poor choice for running any hypothesis tests or quantifying uncertainty. It is not useless though. The stacked method is useful when we want to analyze just one plot instead of m for model checking. As well, the stacked method may be useful in situations where we partition categorical data on an imputed variable and then look at the percentage in each category. Under the averaging portion of Rubin's rules, we are not guaranteed that the percentages will sum to unity, but under the stacked method we are.

2.2 Survival analysis

Now that we have the multiple imputation datasets created, we may run our analysis on each of them. Since the goal of this paper is cancer survival data, it naturally follows that the models we would like to run are survival models. As a general rule of thumb, we should run our desired analyses on the original available case data to get the available case estimates, as well as to get an idea of what to expect and to check if the model assumptions are met. Following Rubin's rules, we run the individual analyses on each of the m datasets, and then pool our results. In this section we will discuss each type of survival analysis in the MI setting and the issues associated with it.

2.2.1 Kaplan-Meier Survival Curve

Analysis for the Kaplan-Meier estimate is quite simple in the non-MI setting, but special care should be taken in the MI setting. First, we need to clearly define what the groups and population are, and what constitutes an event of interest. A very common mistake that researchers make is to try to frame a competing risks problem as a Kaplan-Meier problem. We also need to be sure that we have non-informative censoring, that is, knowing that the individual is censored tells us nothing about their survival probability. Once we have checked all of these, we can compute the Kaplan-Meier curve on each of the m datasets. For simplicity, in this section we assume that the Kaplan-Meier curves are run on a dataset with only subjects who take a treatment or a control. 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 [28]. We can make

this transformation, pool our results, and then back transform to get the pooled KM estimate.

In the MI setting, an interesting situation may arise when the last event (and thus the range of survival time) differs between the imputed datasets. This is the result of a person with a long survival time being put in different groups via imputation. We can deal with this by either extending the last observed Kaplan-Meier estimate out until the last event time, or truncating all of the imputed curves at the minimum time of last event. Alternatively, we could use the stacked method instead to analyze it, but we must accept the fact that the variance at every time point will be too small. In traditional analysis, we would not extend out the Kaplan-Meier curve out past the last event, but if we have wildly varying imputations, this might be a good option, so that we can make inference between curves and are not hampered by one poor imputed dataset. The decision of what to do is up to the researcher and depends on the nature of the problem. Several other methods are also described by Klein about how to deal with this [3]

2.2.2 Median Survival Time

One of the main tasks that clinicians are interested in is the median survival time of each group, which is the smallest time that the survival probability function for the group is less than 0.5. The median is the preferred method of central tendency in survival analysis because often times the survival times are right skewed, making the mean a poor estimator of the truth. Finding the median in the MI case is quite simple, as it is the first time that the averaged Kaplan-Meier curve crosses below 0.5. We would also like to have the variance at the median. The first way to obtain it is to multiply impute the variance (i.e. make the variance the scientific estimand of

interest), and then use Rubins rules to get the variance associated with each time point and then take the average as the variance at that time point. However, taking the mean and variance of the variance doesn't make much sense, and under typical pooling of the Kaplan-Meier estimate, we already get an good estimate of the variance. A better 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 using the total variance obtained from Rubin's rules pooling. The median is defined as the first time when the pooled MI curve crosses 0.5 survival line, and the lower and upper bounds are the medians of the upper and lower confidence bands. This method is preferred because it uses the variance we actually have and is much more robust to poor imputations.

2.2.3 Log Rank Test

Now, we will have a pooled estimate of the true survival curve for each group. In the typical setting, we might want to look to see if these curves are similar to each other, so we can determine if the treatment really prolongs survival time. 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, which 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 log rank test on each of the datasets and then pool the results via Rubin's rules. This is the logical way to do it, but under this scheme we will run into a multiple comparison problem. As well, Marshall shows that this method is unreliable and will not give the proper p-values [28]. Another option is to run a Cox regression on just the group in

question, since this is equivalent to the log rank test under no tied failure times [3]. Under tied failure times, they are very similar to eachother. From the Cox regression, we can obtain the score test. However, in order to do this, we need to know who is in the risk set. But the concept of a risk set doesn't exist in the MI setting. Luckily, the Wald test is asymptotically equivalent to the score test, and the Wald test is very easy to obtain. So we can use the Wald test of the coefficient from the Cox model as a proxy for the log rank test. In this way, we get a statistic that calculates what we want, while still making sense in the MI context.

2.2.4 Cox Proportional Hazards Model

We would now like to investigate the hazard ratio of different baseline covariates and treatments 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. 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 isn't 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 looking to see if the Schoenfeld residuals are uncorrelated with time for each covariate in the Cox model. The Schoenfeld residuals are tedious to explain and derive, and add no value to this thesis. But knowing that they are partial residuals that that are formulated to be independent of time should suffice to understand this paper [3]. We can check a test

for correlation or observe a spline fit to the residuals to determine how uncorrelated they are, but the latter is usually preferred in practice. In cancer research, the most common test to look for proportional hazards is to plot the spline fit (often cubic) to the residuals along with the 95% confidence intervals, and see if any straight line could pass through the bounds. Because the residuals are independent of time, a straight line over time signifies that the hazards are proportional [29]. There isn't an official name for this method, but the straight edge method seems to be a fitting name since you can check it by placing a straight edge between the confidence bands. If this is the case, then we say that that the covariate in question follows the proportional hazards assumption. Another method to check for proportional hazards is to use a chi square test of independence between the Schoenfeld residuals and time, although this is rarely used in practice.

We have discussed how to check the proportional hazards assumption in the available case scheme, but how can we do this in the MI setting? 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 (especially with large m, but with graphical tools such as shiny, this isn't 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 analyze to see how they affect the hazards. An interesting thing to look at will be how the hazards change from the baseline once we factor in the treatment.

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 and appealing than conventional analysis. Although no statistical method will give us a causal relation, the results from causal analysis will be much more similar to the results we would have got from a RCT. Propensity score methods are an easy to understand yet powerful tool. The use of propensity scores justified in Rosenbaum and Rubin's 1983 paper [9]. Our overall goal is to estimate the average treatment effect in a setting where the initial study was not an RCT. We will use the propensity scores to balance the groups so that it seems more like an RCT.

Before we even begin using propensity scores, we should have in mind what estimand we wish to estimate (the ATE or the ATT). This should be done by addressing the problem and discussing the implications of each with subject matter experts. Once we have this, we can start out with the propensity score methods. We will need to make many 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 [13]. 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 treatment effect, but each goes about it in a different way. Covariate adjustment and matching have fallen out of favor in the literature recently, but weighting and stratification remain very popular [13]. We will be using weighting in our real data analysis, so the main focus will be on that, but the methods discussed will work with any propensity score method. Whatever method is chosen will need to be run on the MI datasets. The stacking method described before would obviously be inappropriate, as the sample size will be too large, and with high probability, subjects will be matched to themselves. Thus, the analysis will not be valid

Next we must decide how to use propensity scores on the multiply imputed data. Two methods are described in Mitra and Reiter about how use propensity scores in the MI setting [30]. In both methods, the propensity scores are first computed for each MI dataset, as we would normally do in the analysis part of the MI process. However how the two methods use the propensity scores are different. In the first method, known as the "within " method, the propensity score method is done within each MI dataset and then pooled. For example, if we wanted to find the ATE, we would estimate the ATE in each dataset, and then pool via Rubins rules. The other method, known as the "across" method, takes the m estimates of each individuals propensity scores between the MI datasets, and then averages them so that every subject has one averaged propensity score. Once the global averaged propensity scores are estimated, then the propensity score method is used in each dataset, and then pooled via Rubins rules. Both methods have their pros and cons, and are appropriate for different scenarios. However, 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 cannot determine if the subject is a case of a control, thus matching will be impossible. When this is the case, we must use the within method.

Now that we have discussed the method about how to use propensity score methods in the MI setting, lets discuss how to actually get the propensity score. From now on, propensity score weighting will be the only method discussed, because that is what will be used in the application section.

Different propensity scores will be obtained according to what pretreatment covariates we use in the propensity score model, so we need to be sure that model is fit with clinically relevant and meaningful predictors. Recall that our overarching goal is to account for any pretreatment covariate that confounds 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 don't plan to settle this debate, but I do believe that the user should account for everything that is believed to be different between the two groups in question, in order to make the ignorability assumption more plausible.

Propensity scoring using logistic regression is simple and has been historically used, but with newer machine learning methods, we are able to get better propensity score weights that induce more balance. For single analysis, logistic regression would be preferred, but in the MI setting, the model that fits well for one dataset might not fit well for another one. That is why machine learning methods are preferred. In recent years, boosting has become very popular method to compute propensity score weights, because they can be selected in a way to achieve optimal balance[31].

The goal of weighting is to create a pseudo-sample that is balanced, by up weight-

ing samples that look like treatment cases, and down weighting those who do not. Two methods have been suggested to check for balance. The first is known as standardized bias. For each pretreatment covariate we wish to balance, we see how similar the cases and controls are by observing $|\bar{X}_{k1} - \bar{X}_{k0}|/\hat{\sigma}_k$, where the barred Xs denote the weighted means of the covariate in the treatment and control groups, and the σ is the original standard deviation. Typically, standardized bias below .2 indicates good balance, although this number is subjective [31]. Another common method is the KolmogorovSmirnov test of distributional equality. In this nonparametric test, the empirical cumulative distribution function (ECDF) is computed for the two (weighted) groups. Then, the maximum distance between the two ECDFs is calculated. Its p-value is found via the permutation test for the null hypothesis of no differences in the distribution.

Both of these methods offer us tools to assess balance in the data when weighted by the propensity scores. We can use the tools to our advantage alongside of a boosting algorithm to get good balance. In boosting, a simple piecewise linear function is fit to the data, and after every iteration, a new piecewise linear function is fit to the previous models residuals. We can boost our data, weight each observation by its propensity score, and then check the balance at each iteration of the boosting algorithm. The goal is to find the propensity score weights that minimize either the standardized bias or the KS statistic, although minimization of one often leads to significantly better balance in the other metric.

Once we have obtained our optimal propensity scores, we weight the data by them. Then, we check to see if we have balance. As well as checking that we truly have achieved balance, we also need to look at the distribution of the propensity scores. We would like to see some overlap between the treatment and control propensity scores.

Overlap indicates that subjects in both groups are similar to each other, and can be compared. If the propensity scores dont have any common support, then we have not properly balanced. We should also be sure that all of the propensity scores are bound between zero and one. If a subject has a propensity score that is either one or zero, then they will always be or never be treated, and estimating the counterfactual will be ill-advised.

Once we have our propensity scores, we need to go about checking the treatment effect. We have no guarantee that the model we set for our propensity score is correct, so many authors advocate including troublesome variables in the analysis of the treatment effect (this is known as a doubly robust method) [12].

Now that the theory is set up, we may use it on the MI data. And now we may interpret the results with a causal lens, even though the data may be observational. We will be using the within method, so for each MI dataset, we compute the treatment effect from the weighted data. While we could check the treatment effect in terms of time difference in survival between the treatments, this doesnt work very well, because we have censored data. The better option is to observe the average treatment effect by looking at the change in the hazard between the treatment groups. Checking the hazard makes use of the fact that we have censoring in the data. We may then combine the estimates via Rubins rules to get the MI estimate of the average treatment effect and its variance.

Chapter 3

Application

3.1 Data Explanation

Now that the theory is 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, and Dr. Hess. The IRB protocol is RCR03-0931. This dataset has historical records of 1514 MD Anderson patients who have had breast cancer that has metastasized to the brain from October 2009 to December 2012. Metastases are often shortened in speech and in paper to the word "mets". The dataset consists of 111 covariates, with missingness ranging from 0 to 99%. Some predictors are metadata, and rare tests, so they will not be considered. Ignoring these, so the missingness in the useful data ranges from 0 to 65%. Included in these are 90 different covariates, a few different treatments, as well as survival endpoints (which are all observed). The data can be broken down into a few broad categories, as can be seen in table 3.1.There are too many covariates to completely explain here, but Ive listed the ones relevant to our models in table 3.2

To get a feel for how the data is missing, in figure 3.1 we see a plot of the missingness in the data. There are certainly large swaths and groups of covariates that were not collected, but overall, there seems to be no real pattern in the way the data is missing.

This data is exemplary for demonstrating thesis ideas because it is a large retro-

Type	Example
Subject data	Age range, race, date of birth
Breast Cancer data	TNM staging, type, receptor status
Pre brain mets data	Treatment types
Post brain mets clinical observations	Seizures, headache, nasuea
Post brain mets data	Treatment type, type of brain mets
Survival data	Survival time after brain mets, censoring indicator

Table 3.1 : Data Categories and Examples

spective study (pulled from a database), survival amenable, has missingness that is a prime candidate for imputation, and has treatment variables that are not given in an RCT.

Our first step is to clearly define what we would like to find. We will answer two separate questions in the section. In the first, we will explore the treatment effect of Capecitabine (a chemotherapeutic agent) versus other chemotherapeutic agents versus no treatment. In the second, we will look at the effect of two HER2 directed drugs (Lapatinib versus Trastuzumab) versus no HER2 targeted drugs in a subset of the patients who are HER2+.

It isn't vital to understand the entirety of cancer and its treatments for understanding this analysis, but the interested reader may want to look at appendix A.2 for a very basic overview of breast cancer and the methods of how different drugs

Name	Percent Missing	Meaning
capeothno	18%	Indicator: Capecitabine, other, or no chemotherapeutic treatment. Treatment variable 1
lapatrasno	18%	Indicator: Lapatinib, Trastuzumab, or no HER2 treatment. Treatment variable 2
controlled	12%	Indicator: Extracranial progression of brain mets
hrher2	5%	Categorical variable: The hormonal receptor and HER2 receptor status of the subject
braintype	4%	Categorical: Single, multiple, Leptomeningeal disease
timedx	1%	Indicator: Time (years) from breast cancer diagnosis to brain mets diagnosis greater or less than 6 years
site5	1%	Indicator: First metastasis was to brain
race2	0%	Categorical: White, Black, Hispanic, other
priorn	0%	Indicator: Number of prior treatments in metastatic setting before brain mets
os	0%	Overall survival (months)
dead	0%	Indicator: death indicator
agebrainmet	0%	Indicator: Age greater or less than 60 at time of brain mets

Table 3.2 : Table of important covariates to be used in the analysis

work. For a much more detailed analysis and other clinically relevant questions, see the upcoming paper by Bugano, Hess, and Berliner. This is the project that this research was forked off of.

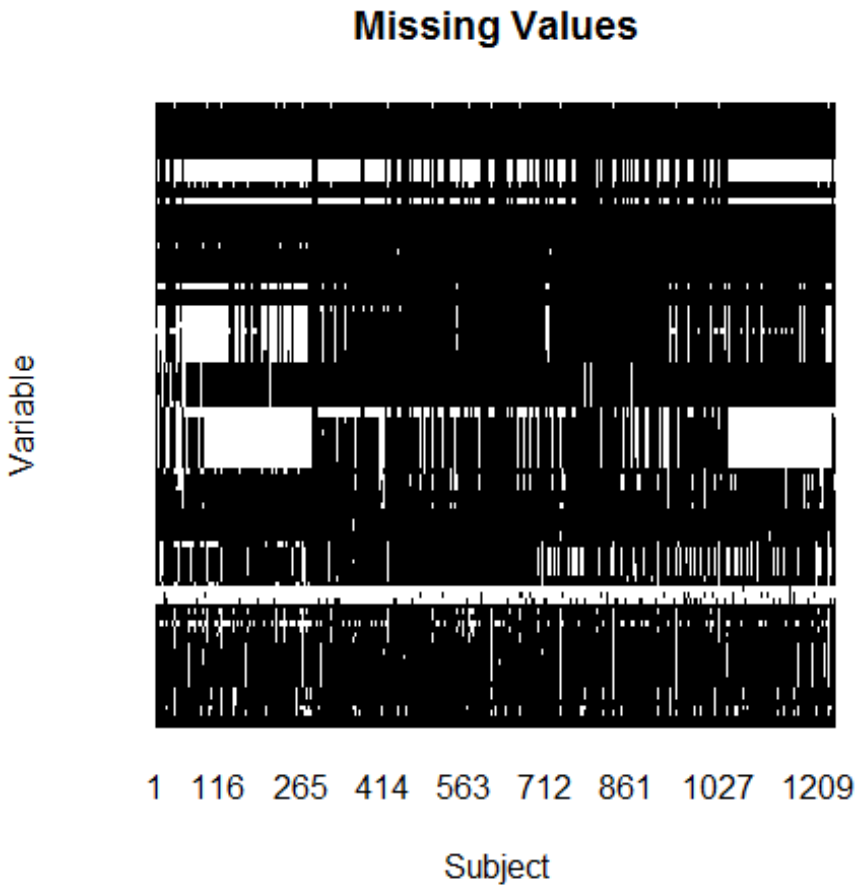

Figure 3.1 : Visualization of missingness in the cancer dataset

Along the horizontal axis is the subject, and the vertical axis are the covariates. Black denotes observed values whereas white is missing. The covariates with the highest missingness are three genetic measures, as well as some clinical assessments.

3.2 Imputation

We first need to impute the missing data. This is a challenging task, because of the attention and care that needs to be given to imputing about 90 covariates with missing data. Although we will not be using all 90 covariates, it is important that we impute

them, and do so properly. We need them because; they have the potential to be useful as predictors for other covariates, they might be something we are actually analyzing (now or later), 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, variable selection can be performed if there are too many covariates.

Our data is quite high dimensional, and there are a many binary variables as well as a handful of strictly positive covariates, thus JM imputation seems inappropriate. Instead, FCS models seem better suited. We will be using the R package mice [20] because it is easy to use yet powerful. There are other software implementations in different languages (such as PROC MI in SAS, ICE in Stata, and package mi in R), but I found mice to be the most flexible while also being powerful and easy to use.

The first task we need to do is to assess the missing data mechanism. As we have discussed before, there is no formal statistical test to determine what the mechanism is. It is very unlikely that the data is MCAR (which we typically associate with random/accidental deletion), so it is between MAR and MNAR. We have so many different covariates, and it could reasonably be assumed that the missing data we have could be explained by the type of disease, its stage, the subject's age, their standardized assessment, and their survival time, among other things which we have collected. So it would be reasonable to assume that the missing data mechanism is MAR, and thus imputation can be confidently used.

Now that the missing data mechanism has been assessed, the imputations need to be set up. For each covariate with missingness, we need to decide the form of the imputation model that will be used for imputation, and what predictors will be used in it. Most of the covariates that needed imputation were either binary or cate-

gorical, so the most popular model chosen were logistic regression, multinomial logit regression, or predictive mean matching. The method that was most appropriate for each situation was chosen. The continuous variables were often selected via regression or predictive mean matching. I decided to be very forgiving, and use nearly every reasonable predictor for each missing covariate. I did this to bolster the MAR claim, and avoid variable selection. Van Buuren proposes measures called influx and outflux to determine how worthy and connected each covariate will be as a predictor [1]. This information was used to remove 11 covariates that were very poor for prediction. Once the model and predictor choices were made, the imputations were run and tuned, and checked by trial and error. This took a considerable amount of time, because after every change made, the algorithm needed to be reran and the convergence and imputations needed to be assessed. As well, changes in model specification were rarely localized to that variable, and often affected others.

It took about three weeks to set up and check the models. This was because the number of covariates was large, and checking the imputations after a change was time consuming. It would not take this long for a smaller dataset. Creating valid imputations is a skill that lies somewhere between an art and a science, so it takes the theory to know what to do, and trial and error to see if you've done it correctly.

For the final MI dataset, it was decided to impute $m = 50$ datasets and 40 iterations for each. Research by White et. al says that you should choose m to be about 100 times the percentage of incomplete cases (for the analysis at hand) [33]. The data used in our analyses had about 30% missingness, so imputing 50 datasets was the chosen number. Mice generally converges quickly (within 5 or 10 iterations), but by setting the number of iterations so high (40), it is as if we are setting a burn in period, and then taking our sample.

After the final imputation model for each covariate with missingness has been set up, we need to run it and save the results. For 50 datasets, 40 iterations, the algorithm runs in about 4 hours, and for 50 datasets with 100 iterations, it took 11.5 hours on a computer with 4 GB of ram and 4 cores. 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. The imputations were run for 100 iterations to see how the run time scaled, as well as to check how the chains behaved and to see how the analyses differed. The results between 50 and 100 iterations were very similar. As well, there is hardly any confidence to be gained going from 50 to 100, and having such large objects in memory can be harder to work with. This is why 50 MI datasets, iterated 40 times each were chosen.

We need to check our final imputations for convergence, reliability, and validity. Convergence is assessed by looking at the trace plots of the imputed value 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" [20]. This may be hard to visualize all at one time because for each variable, there will be 50 time the number of missing data values for that covariate, so many authors suggest checking the plots of the chain mean and standard deviation by iteration instead. For continues variables, the chains are checked to see how the variance between and within changes at each iteration, as well as to ensure the chains show no pattern.

Diagnosing convergence for binary and categorical variables is a bit tricky, because the mean of a categorical variable is nonsensical, but we can get a good idea of convergence by looking at if the variance remains constant. And for binary variables,

knowing how the groups are coded and looking at where the means are can give us an idea of healthy convergence. For example, in the cancer dataset, the variable HER2 is coded as 1 for being HER2 negative and 2 for positive. In the available cases, the split between being HER2 negative and positive is about 60 to 40. Thus, is we were to look at the average of the chains groups, we would expect it to be around 1.4. Seeing something radically different than this (perhaps a mean of 1.9) might suggest the imputation was done wrong, because the group percentages are not near what they should be. We can assess the convergence of the binary variable also by checking how its variance changes by iteration. Large changes from groups by iteration will be noted as large changes in the variance plots, whereas staying constant indicates that the draws are remaining relatively stable.

There are 80 plots to check, they cannot all be shown in this paper, but all of them have been checked. Some interesting ones may be seen in figure 3.2 and 3.3. Looking at these plots, convergence certainly seems to be the case. Other authors suggest using a more formal statistical tests such as $\hat{R}$, but since most of our interesting variables are categorical, this does not make much sense.

Once the convergence is diagnosed, the validity of the imputations needs to be inspected. Diagnostic plots are viewed to ensure that the imputed data is similar enough to the real data. Common plots include density plots of each MI dataset compared to the complete cases, as well as bivariate scatterplots for imputed variables. ¡ A few of the plots have been replicated here!!! Still need to do this¿. Once again, all of the plots cannot be displayed in this paper, but some important ones are recreated. As we can see, not all of the imputed data follows the distribution of the observed data exactly, but for the majority of the plots, the data look like they could have been real data had they not been missing. Now that we have verified the data for

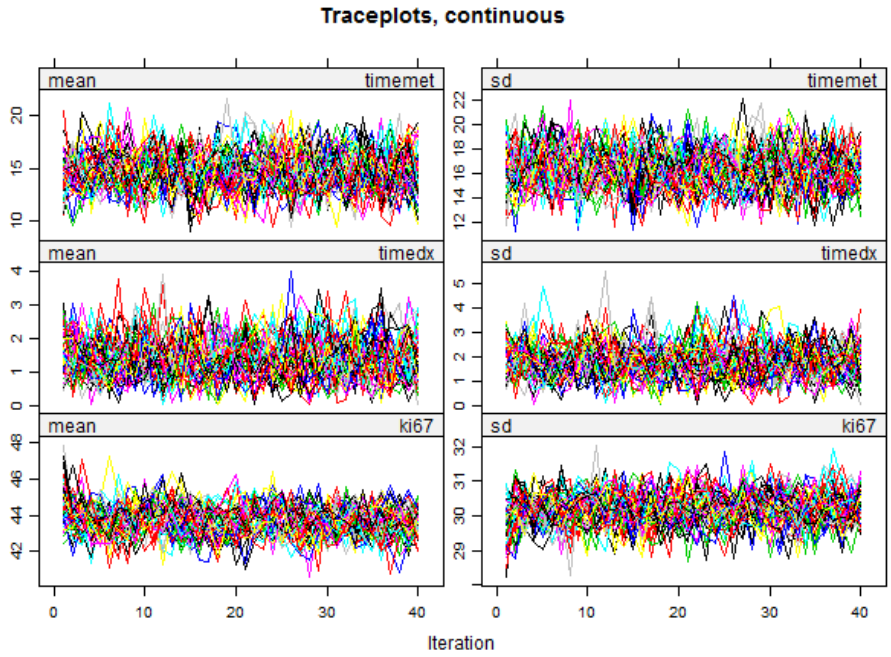

Figure 3.2 : Selected plots of continuous variable imputation mean and standard deviation by iteration

convergence and validity, we may run the analyses on the $m = 50$ datasets, which we will do in the coming sections. Before we move on to the analyses though, lets have a look at the breakdown of the data. In table 3.3 we can see how the MI data compares to the available case data, broken down by if the subject had any systemic therapy. This was computed via the stacked method.

3.3 Survival Analysis

Now that the datasets are imputed, we are ready to run our models on them. Before we begin though, we should check the models on the available cases to make sure that the model assumptions are met, and to get an idea of how the importance of each part of the model. The available case models (Kaplan-Meier, Cox) were run and the

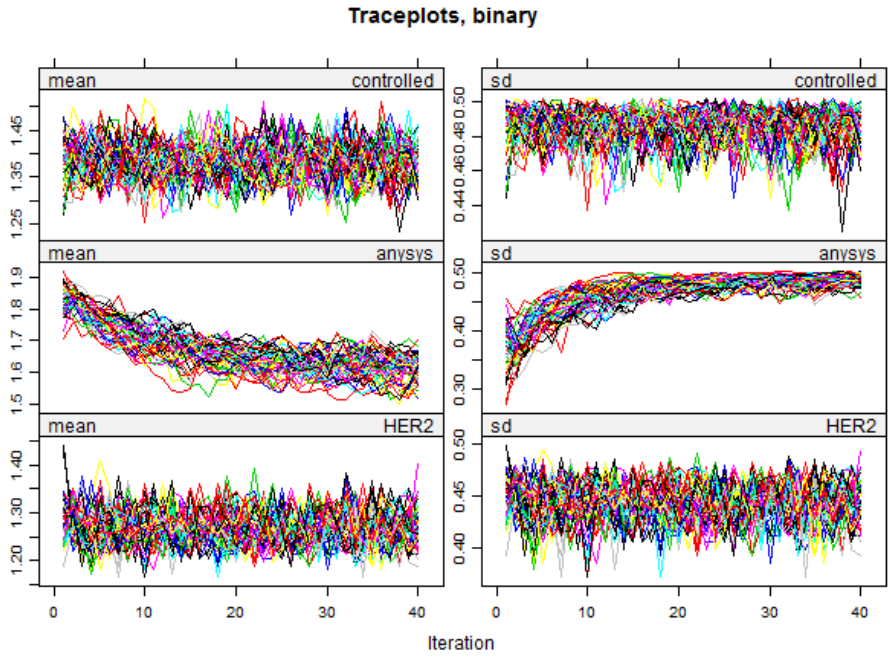

Figure 3.3 : Selected plots of binary variable imputation mean and standard deviation by iteration

assumptions were checked, and all of them passed. You can see the results, alongside the MI values throughout this section.

It should be noted that in all of our survival analyses, we will be doing a landmark analysis. Landmark analysis means that we don't start the analysis at time 0, rather, we start it at a different time later than 0. In Dr. Hess's words, "Since the brain mets 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 have had their brain met treatment choices started, and start our analyses at this point ". After speaking with cancer experts (Dr. Bugano and Dr. Ibrahim), this landmark time was determined to be 2 months. This allows us to be sure that the treatment

	Sys therapy available case	Sys therapy MI	No Sys therapy available case	No Sys therapy MI
Age (mean,sd)	51.4(10.8)	51.2(10.9)	52.7(11.9)	52.9(11.4)
Breast Cancer subtype				
HR+/HER2-	27%	31%	28%	33%
HR+/HER2+	19%	18%	12%	13%
HR-/HER2+	22%	20%	15%	12%
Triple negative	32%	32%	45%	42%
Prior therapies for stage 4	1(0-3)	2(0-4)	2(0-4)	2(0-4)
Single brain lesion	25%	23%	23%	20%
Controlled extra-cranial	40%	40%	35%	36%
ECOG 0-1	84%	70%	53%	40%
Local Therapy				
Resection Alone	5%	5%	9%	7%
SBRT alone	13%	12%	9%	8%
WBRT	60%	59%	52%	53%
Resection/SBRT+WBRT	12%	14%	10%	8%
no local therapy	10%	10%	20%	23%

Table 3.3 : Characteristics of available case data versus MI data

was actually administered for the subjects in the analysis.

The first result that we will check is the Kaplan-Meier curves for the imputed data. Non-informative censoring seems to be a valid assumption, as knowing the censoring status seems to provide no information about the censoring time. The pooled KM

estimate was found using Rubin's rules, but under a complimentary log-log transform as suggested by [28] to get the survival curves towards normality. The median and confidence interval about it was computed as the median of the upper and lower confidence bands.

For the comparison of the chemotherapeutic drugs, we would like to look at is the survival curves for them. This can be seen in 3.4. The available case and MI analyses look similar, but the MI data seems to give a lower median survival time for the chemotherapeutic drugs. For not taking any chemotherapeutic drugs though, the median increased.

For the HER2 targeted drugs 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 MI analysis says about the same, but once again gives lower values for the median survival time, as can be seen in figure 3.5.

Both Kaplan-Meier curves for both questions seem to show the same thing. As compared to the available case analysis, the estimates of survival are a little pessimistic for the treatments, and a little more optimistic for the no treatment group. This is likely due to the fact that a larger proportion of the missing treatments were assigned to no treatment (which echoes the data), causing their survival to go up.

Now that we have a visual of the curves, we would like to see if there is actually a difference between them. Although visually, we can see that in both cases, no treatment seems to be much worse for survival than treatment, we need to formalize it. To do so, the log rank test needs to be run on them. We can also get an approximation for the log rank test on the MI data via the Wald test on the pooled Cox model fit only on the treatment. Recall that 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

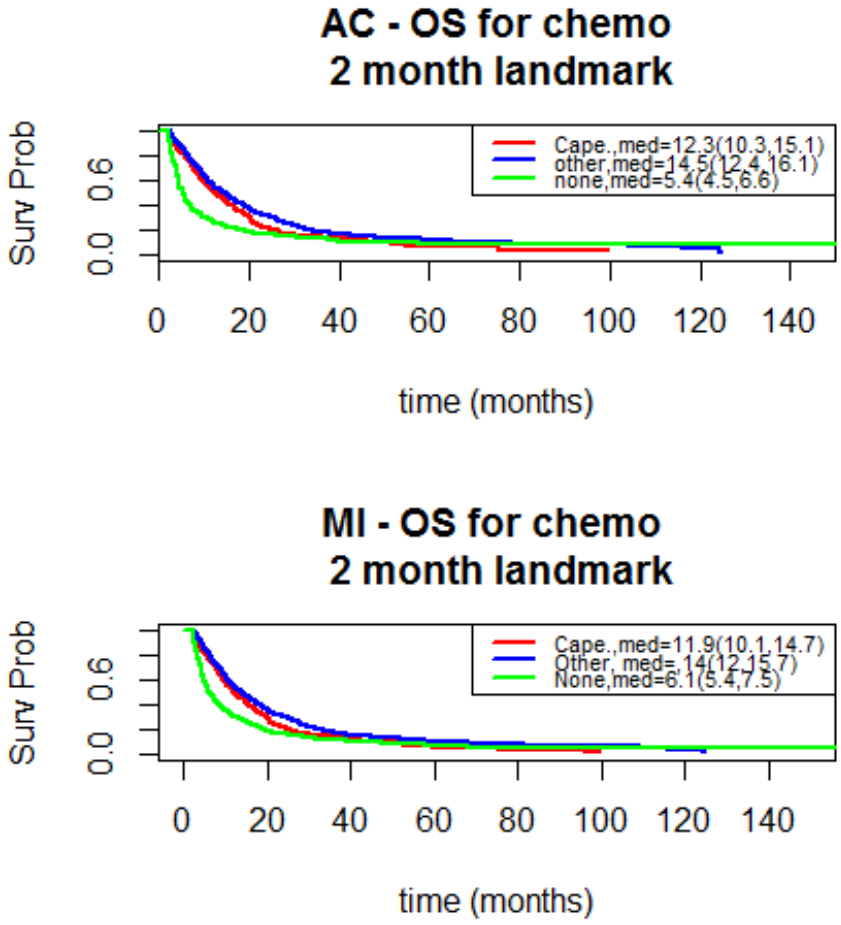

Figure 3.4 : Landmarked Kaplan-Meier curves for chemotherapeutic drugs, AC and MI data

score test, both of which would include calculating the risk set, which is not possible in the MI setting. It has been suggested to pool the chi square statistics via methods presented in Marshall et al 2009, but even they say that this method is poor [28]. 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).

The results for the overall test and for each comparison can be seen in table 3.4 and 3.5. The AC and MI results are quite similar. For Capecitabine vs other

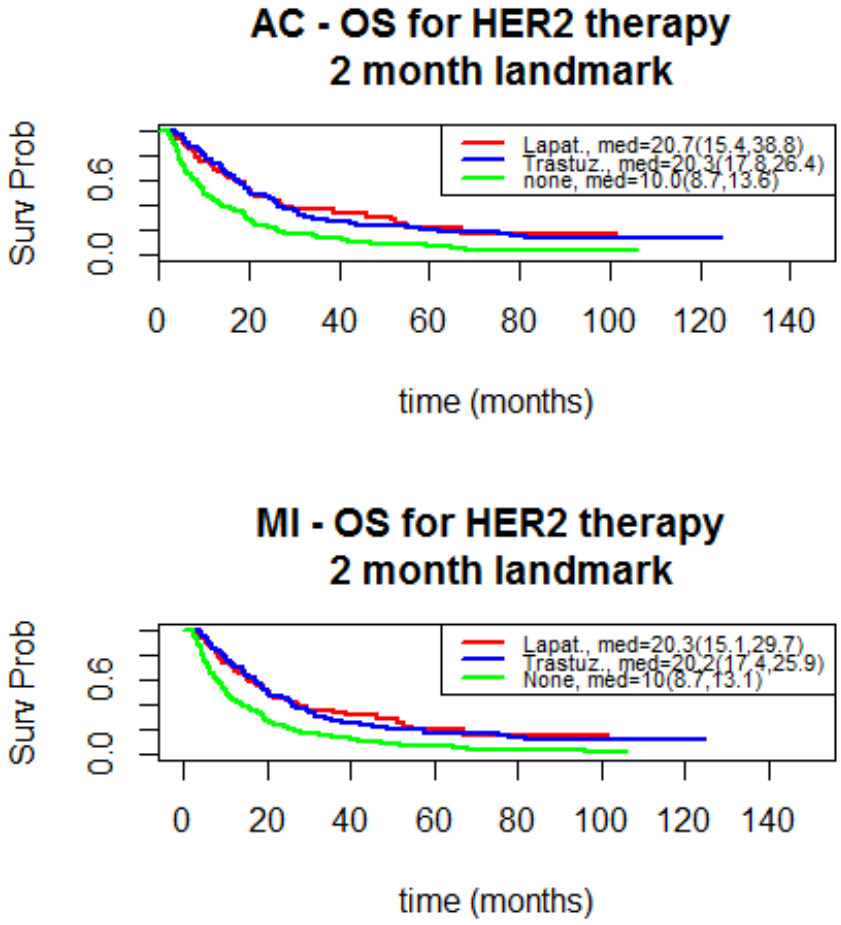

Figure 3.5 : Landmarked Kaplan-Meier curves for HER2 drugs, AC and MI data

chemo drugs, we see a relatively high p value. Depending on the acceptable rate of significance, different conclusions might be drawn between the two analyses. In this study, the level is $\alpha = .05$, so changing conclusions is not an issue, however, since we are doing multiple comparisons, it may be prudent to use a Bonferroni correction for multiple testing. For the HER2 directed drugs, there certainly seems to be a difference between the treatments and not having any, but there is no evidence that either of the treatments is better than the other.

	Chemo	
	AC	MI
cape/other/none	<.0001	<.0001
cape/other	0.0321	0.033
cape/none	0.00039	.0016
other/none	<.0001	<.0001

Table 3.4 : Chemo log rank tests

	HER2	
	AC	MI
Lapat/Traztuz/none	<.0001	<.0001
Lapat/Trastuz	.87	.81
Lapta/none	.00017	.00018
Trastuz/none	<.0001	<.0001

Table 3.5 : HER2 log rank tests

Now that we have estimate of the survival curve, 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. We first need to fit a reasonable model on the available cases. Speaking with the clinicians, they determined that the covariates listed in table 3.2 were clinically relevant for the baseline model.

We need to make sure that the proportional hazards assumption is met in the available case model to determine if the use of the Cox model is justified. To check,

we visually inspect the spline fit to the Schoenfeld residuals over time. The available cases can be seen in figure 3.6. In the available cases, the assumption of proportional hazards over time seems reasonable, as it a straight line could reasaonably be fit between the 95% confidence bands.

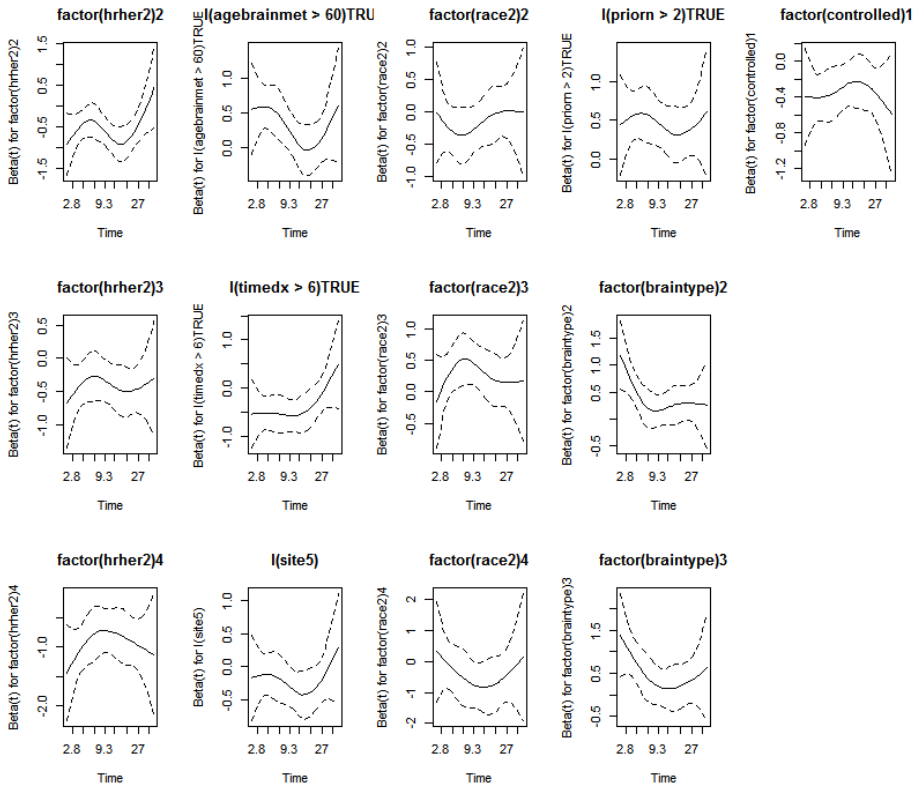

Figure 3.6 : AC Schoenfeld residuals over time

Now, to check if we have proportional hazards in the MI data, there are two options. The first is to fit a Cox model on the stacked data, but this will not give us accurate confidence bands. A better option is to collect all of the spline fits, and then superimpose them onto one plot, and assess the shape. This is what is done for the MI data, as can be seen in figure 3.7. The shapes are very similar to the available

cases, so it is reasonable to assume that the proportional hazards assumption holds on the MI data.

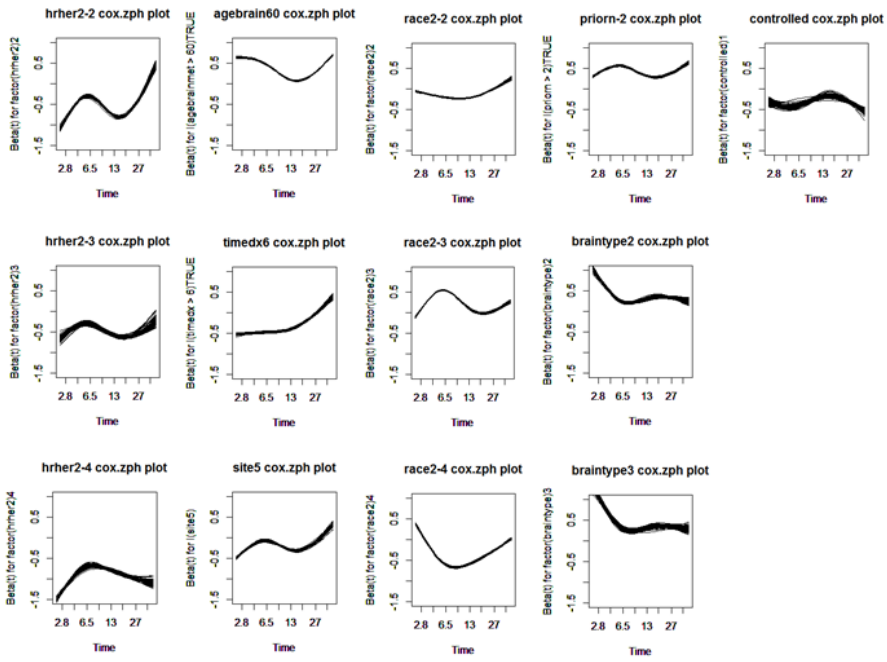

Figure 3.7 : MI Schoenfeld residuals over time

Now that the assumptions are met, we can run the Cox model to obtain the hazard ratios. A table of the AC and MI analyses can be seen in table 3.6.

We may then add in our treatment variable to see how it affects the hazard, and see how it changes other factors. Adding chemotherapy can be seen in 3.7, and for the HER2 positive patients, the HER2 directed drugs can be seen in 3.8. As it has been posited before, the use of chemo and HER2 drugs seem to greatly reduce the hazard ratio of death from cancer. The effect of other covariates in the presence of these drugs seems to change though.

			AC n= 845				MI	
Variable	Contrast	HR	95% CI	pvalue		HR	95% CI	pvalue (t test)
HR/HER2	-/+ vs. -/-	0.57	(0.46,0.71)	<0.0001		0.59	(0.48,0.72)	<0.0001
	+/- vs. -/-	0.66	(0.54,0.81)	<0.0001		0.63	(0.52,0.76)	<0.0001
	+/+ vs. -/-	0.4	(0.31,0.50)	<0.0001		0.4	(0.32,0.50)	<0.0001
Age	>60 vs. <60	1.37	(1.13,1.65)	0.0011		1.45	(1.22,1.72)	<0.0001
Dx to BM	>6 vs. <6	0.66	(0.54,0.82)	0.00013		0.71	(0.59,0.86)	0.0002
First DM	Brain vs. Oth	0.8	(0.66,0.97)	0.026		0.83	(0.70,0.99)	0.02
Race	Hisp. Vs. White	0.85	(0.68,1.07)	0.17		0.88	(0.71,1.08)	0.11
	Black vs. White	1.31	(1.06,1.63)	0.014		1.25	(1.02,1.52)	0.015
	Other vs. White	0.65	(0.40,1.04)	0.075		0.7	(0.45,1.07)	0.05
# prior Rx	>2 vs. 0-2	1.58	(1.31,1.91)	<0.0001		1.53	(1.29,1.82)	<0.0001
BM type	Mult. Vs. Single	1.45	(1.20,1.76)	<0.0001		1.48	(1.24,1.76)	<0.0001
	LMD vs. Single	1.6	(1.21,2.13)	0.001		1.58	(1.25,2.00)	<0.0001
Sys. Cont.	Yes vs. No	0.71	(0.61,0.83)	<0.0001		0.73	(0.63,0.85)	<0.0001

Table 3.6 : AC and MI baseline Cox model

			AC n=745				MI	
Variable	Contrast	HR	95% CI	p-value		HR	95% CI	p-value (t test)
HR/HER2	-/+ vs. -/-	0.62	(0.49,0.79)	<.0001		0.63	(0.51,0.77)	<.0001
	+/- vs. -/-	0.65	(0.53,0.81)	0.00011		0.64	(0.53,0.78)	<.0001
	+/+ vs. -/-	0.41	(0.31,0.53)	<.0001		0.42	(0.34,0.53)	<.0001
Age	>60 vs. <60	1.34	(1.10,1.64)	0.0041		1.44	(1.21,1.72)	<.0001
Dx to BM	>6 vs. <6	0.72	(0.58,0.90)	0.0032		0.71	(0.58,0.86)	0.00039
First DM	Brain vs. Oth	0.77	(0.63,0.95)	0.014		0.81	(0.68,0.96)	0.016
Race	Hisp. Vs. White	0.77	(0.61,0.98)	0.034		0.86	(0.69,1.06)	0.15
	Black vs. White	1.29	(1.02,1.63)	0.032		1.23	(1.01,1.51)	0.043
	Other vs. White	0.76	(0.47,1.25)	0.28		0.7	(0.45,1.08)	0.11
# prior Rx	>2 vs. 0-2	1.61	(1.32,1.98)	<.0001		1.53	(1.28,1.82)	<.0001
BM type	Mult. Vs. Single	1.46	(1.20,1.78)	0.00017		1.51	(1.27,1.81)	<.0001
	LMD vs. Single	1.45	(1.04,2.03)	0.029		1.41	(1.11,1.80)	0.0049
Sys. Cont.	Yes vs. No	0.57	(0.48,0.68)	<.0001		0.69	(0.59,0.80)	<.0001
Chemo	Cape. vs. none	0.69	(0.53,0.89)	0.0046		0.75	(0.60,0.95)	0.018
	other vs. none	0.52	(0.42,0.65)	<.0001		0.58	(0.47,0.71)	<.0001

Table 3.7 : AC and MI Cox model with chemo treatment

			AC n=292				MI	
Variable	Contrast	HR	95% CI	p-value		HR	95% CI	p-value (t test)
HR/HER2	+/+ vs. -/+	0.65	(0.49,0.87)	0.0036		0.66	(0.51,0.85)	0.0015
Age	>60 vs. <60	1.38	(0.95,2.01)	0.092		1.58	(1.15,2.18)	0.0054
Dx to BM	>6 vs. <6	0.64	(0.43,0.97)	0.033		0.69	(0.49,0.99)	0.041
First DM	Brain vs. Oth	0.84	(0.58,1.20)	0.34		0.86	(0.62,1.17)	0.34
Race	Hisp. Vs. White	0.69	(0.46,1.02)	0.064		0.76	(0.53,1.09)	0.14
	Black vs. White	1.41	(0.94,2.11)	0.1		1.43	(1.00,2.04)	0.047
	Other vs. White	0.7	(0.32,1.53)	0.38		0.83	(0.46,1.52)	0.55
# prior Rx	>2 vs. 0-2	1.88	(1.34,2.63)	0.00028		1.71	(1.28,2.28)	0.00028
BM type	Mult. Vs. Single	1.3	(0.92,1.86)	0.14		1.25	(0.91,1.70)	0.16
	LMD vs. Single	2.15	(1.20,3.88)	0.011		1.77	(1.10,2.83)	0.018
Sys. Cont.	Yes vs. No	0.73	(0.55,0.97)	0.029		0.78	(0.60,1.01)	0.063
HER2 therapy	Lapat vs. none	0.47	(0.32,0.69)	0.00015		0.52	(0.37,0.75)	0.00036
	Trastuz vs. none	0.45	(0.33,0.61)	<.0001		0.51	(0.38,0.68)	<.0001

Table 3.8 : AC and MI Cox model with HER2 directed treatment, on HER2+ subjects

3.4 Causal Analysis

Lastly, we will want to draw causal inference, and see what the average treatment effect of each drug is. There are two options to choose from in what type of inference

to do. There is the ATE (average treatment effect), and the ATT (average treatment effect on the treated). Since we have two treatments, the ATT is not of much interest, but the ATE would be clinically useful.

The idea for this part of the analysis is to use propensity score weighting to create a balanced sample, and to be able to treat it as if it was an RCT. Traditional methods to obtain the propensity score (like logistic regression) cannot be optimized for balance. However, machine learning algorithms can be optimized to achieve balance. The method chosen to fit the propensity scores is generalized boosting. In this ensembling method, The model consists of many simple regression trees iteratively combined to create an overall piecewise constant function. The iterative fitting algorithm begins with a single simple regression tree, and at each new iteration, another tree is added. The new tree is chosen to provide the best fit to the residuals of the model from the previous iteration [31] . Once a sufficient number of iterations are run, the weights at each iteration are taken, and the balance of the covariates (according to standardized bias or KS statistic) is measured. The iteration that minimizes the bias while not overfitting the data is the model where our final propensity scores are taken from.

In traditional propensity score analysis, there is only one treatment and control, however in our situation, there are actually three groups in each analysis. So we can no longer use methods tailored to two classes. However, if we make treatment status an indicator (for example 1 if treatment is capecitabine, and 0 if it is other or none), then we may treat it as if it were binary. Because we have three groups though, we will need to check the balance between all groups (i.e 1 vs 2, 1 vs 3, 2 vs 3) to ensure balance is achieved.

Once the propensity scores have been computed, we need to check that they are in fact valid. McCaffrey et Al. describe two conditions that need to be checked, and

graphical checks to do so. The first check is the check of no unmeasured confounders. This is impossible to check, but should be reasonable if we include any and all factors we believe to be confounding. The next condition is to ensure that all of the propensity scores are between 0 and 1. If any propensity score is 0 or 1, then there is no chance that the subject looks like that of another group. If this is the case, then the results will be poor, because the counterfactual framework and propensity score analysis was not meant to handle cases like this [31]. To check for balance, we can check ¡¡histograms of the propensity¿¿ scores, as well as¡¡ tables and plots¿¿ of how the balance changes between pretreatment and propensity score weighting.

Once propensity score validity is checked, the weights are put into the Cox model. We cannot be sure that our propensity scores are the truth (this is a known drawback of propensity score analysis), but we can be more confident that it is right by also including the pretreatment covariates as adjustments to the cox model (along with weighting). This is known as being doubly robust estimator [12].

This is what will happen for one dataset, but since we are in the MI setting, we have 50. The within method discussed before will be our combination plan. This method is chosen since our treatment variable is itself imputed, so this method makes more sense. The idea is to calculate the weighted cox model for each dataset, and then pool. In order to do this though, we need to check the ¡¡balance achieved on each dataset¿¿. However, once the balance is assessed, we may just pool the results via rubins rules.

After verifying and running the propensity score analysis, the ¡¡results may be seen here¿¿. As we can see, adjusting for the covariates ¡¡does something¿¿

!!!EVERYTHING BELOW THIS IS OLD AND WILL PROBABLY GO!! treatment status, with the Q variables to get our propensity score. We can look at the

¡standardized bias¿ before and after the weighting to ensure that we have controlled properly, and to see if we may go forward. Assuming that we have removed the confounding factors and now have two groups that we can treat like it was an RCT, we may now run our Cox model again, but weight by the IPTW. Once we have done this, we can ¡observe the results from the AC analysis¿.

Now we need to apply this propensity score weighting to the MI data. We discussed before the within and across method, and remarked that we were confined to use the within method since our treatment variable (lapat/cape) was itself imputed. So the plan will be to fit the Cox models with the inverse propensity score weights discussed in the AC analysis. We need to be sure that the IPTW weighting is still valid in the MI setting though, so we check ¡standard biased, other things¿. Now we may then pool the results via Rubin's rules and analyze is through the Rubin causal model framework. ¡the results can be seen here¿. The results that we can draw from this are X,Y,Z

Chapter 4

Discussion

We have discussed a number of tools and methods to analyze survival data with missingness and make causal inference. There are lots of decisions to be made along the way, and I am in no way advocating that my exact choices will be proper for all situations, I am only claiming that the decisions made were proper for the type of data and questions that we had. I hope that I have given the reader enough information to run their own analysis, even if they dont choose the options that I did.

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 from subsequent analyses 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 [1]. 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 variables (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 [1].

An interesting future extension to this project would be to use a non parametric approach to multiple imputation, such as the one suggested by Long et al in [34]. 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 cons. 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 made a lot of assumptions about how

our subjects were censored. We assumed that all of our subjects who were censored were right censored and non-informative. This seems to be a valid assumption, but there certainly exists left truncation. It may have been the case that there were some left truncated subjects, but once we landmarked, we certainly incurred some left truncation.

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

There are three 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 [1]. 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, i.e. have cancer, metastasize to the brain, go 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.

The last addition would be survival in the competing risks setting. I think that

modeling death in the presence of other factors would be very interesting. Our data structure supported this type of analysis, but it was not requested by the clinicians.

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 boosting to get our propensity scores. This method was chosen because it allowed for optimal balance within each of the IPTW weighted MI datasets. We chose to model the propensity scores only on the variables that the clinicians thought were useful, and we did not do any variable selection. There is a possibility that we omitted an important discriminating variable. The results we obtained through propensity are causal, but some people still may argue about its validity because we did not have a RCT.

Chapter 5

Conclusion

This paper details how to use multiply imputed data to answer survival and causal 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. The first section gives background information. In the second, we discuss the methods and theory used as well as alternative methods of use. In the third section, we test the methods out on a large cancer dataset, trying to draw meaningful inference from a dataset with substantial missingness. We model some basic survival quantities and draw causal inference from it.

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 understand and explain.

- Y is our whole dataset. It will have i rows and j columns. Some of the covariates in the dataset will be completely observed, and others will have missingness.
- Y_j is a specific column of Y. Y_j is composed as $Y_j = (Y_{j,obs}, Y_{j,mis})$, where
  - $Y_{j,obs}$ is the data we have observed for covariate j
  - $Y_{j,mis}$ is the missing data covariate j
- Y_{obs} is all of the data that we have observed
- Y_{mis} is all the data that we have not observed
- 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.

- The missing data model is given as $p(R|Y_{obs}, Y_{mis}, \psi)$
- θ is a vector of the parameters for the full model of Y

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, for example censored data, or when we know that the missing data is systematically different than the observed. If we have strongly nonignorable data, we should either try one of two things. The first is to expand the data (collect something else similar to the covariate with missingness) so that it becomes ignorable and the second is to formulate two imputation models, one for the observed and one for the missing.

Now, we may discuss the three main types of missing data mechanisms. I will give the technical definition, a laymans definition, and an example. For the example, suppose there is a study that takes down subject information as well as records the level of a protein in the blood.

- 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. For example, if a lab technician slips and 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. For example, 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. In general, MAR models are ignorable [1].

- MNAR: Missing not at random: $p(R = 0|Y_{obs}, Y_{mis}, \psi)$ does not simplify, and the missingness depends on data that we have as well as have not collected. For example if the blood testing machine breaks when the protein level is either too high or too low.

A.2 Cancer and Treatment Overview

Cancer is a disease in mammals where cells in the body begin to grow in an uncontrolled manner [35].There are many different types of cancers for the many different types of tissues we have. This paper focuses on breast cancer.

Breast cancer is a common type of cancer with many different subclassifications, that affects both men and women, although women much more [35]. It can be inherited, but often can be detected early on with screening and self-examination. Most cases of breast cancer are sporadic, but studies have shown that a womans risk to get breast cancer is doubled if a first degree relative has breast cancer (i.e. it is genetic). When breast cancer is inherited, it often presents itself earlier in life [36]. One of the major risks of breast cancer is that it will metastasize, that is, the cancerous cells move from the breast to another area of the body. It should be noted that when a cancer metastasizes to another part of the body, the patient is said to have the original cancer metastasized to a new area, not a new cancer. For example, in this paper we study breast cancer patients who have metastases to the brain, not brain cancer patients. The reason for this is because the makeup of the cancer cell is the same type of tissue as the original location, not the new one.

Luckily, there are lots of different types of treatments for breast cancer and its metastases. I will list the major types here.

- Chemotherapy is a class of drugs given to cancer patients. These types of drugs target fast growing cells (like cancer) and kill them. Capecitabine is a common chemotherapeutic drug

- HER2 directed therapy: HER2 is a human protein that is associated with cell growth. If it is determined that the patient has the HER2 protein, then HER2

directed therapies can be used. In HER2 directed therapies, a drug is given that targets the HER2 protein and tries to stop its effect . Common HER2 directed therapies include Lapatinib and Trastuzumab, which we discuss in this paper.

- Radiation therapy: When the cancer tumor is radiated by precisely located beams in hopes of killing or disturbing the cancer growth process by destroying the cancer DNA.

- Surgery: A doctor goes in and physically removes the cancerous cells.

- Hormone therapy: Some cancers have hormone receptors in the tumor cells. If this is the case, then drugs that interfere with these receptors can be used

Often times, a combination of these treatments is used. The exact course of treatment is very dependent on the type of cancer and the type of person. For example, chemotherapy is very difficult on the body, so it is not often used on the very elderly and frail. The course of treatment given should be determined by a subject matter expert (the oncologist), and is highly individual and cancer dependent. There have been many books and articles written about these, and for more information, you can check out [35] and [36].

Bibliography

[1] S. van Buuren, *Flexible imputation of missing data*. 2012.

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

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

[4] E. L. Kaplan and P. Meier, "Nonparametric Estimation from Incomplete Observations," *Journal of the American Statistical Association*, vol. 53, no. 282, pp. 457–481, 1958.

[5] J. M. Bland and D. G. Altman, "The logrank test.," *BMJ (Clinical research ed.)*, vol. 328, no. 7447, p. 1073, 2004.

[6] D. Cox, "Regression Models and Life-Tables," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 34, no. 2, pp. 187–220, 1972.

[7] T. H. H. S. Library, "Study Design 101," 2011.

[8] S. Guo and M. W. Fraser, "Propensity score analysis. Statistical methods and application," 2010.

[9] 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.

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

[11] G. King and R. Nielsen, "Why Propensity Scores Should Not Be Used for Matching," *Working paper*, 2015.

[12] J. K. Lunceford and M. Davidian, "Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study," *Statistics in Medicine*, vol. 23, no. 19, pp. 2937–2960, 2004.

[13] 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.

[14] P. C. Austin, "The performance of different propensity score methods for estimating marginal hazard ratios.," *Statistics in medicine*, vol. 32, no. 16, pp. 2837–49, 2013.

[15] H. Demirtas and D. Hedeker, "Imputing continuous data under some non-Gaussian distributions," *Statistica Neerlandica*, vol. 62, no. 2, pp. 193–205, 2008.

[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] 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.

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

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

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

[21] 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.

[22] 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.

[23] A. Florian and M. F. Meinfelder, "BaBooN: Bayesian Bootstrap Predictive Mean Matching - Multiple and Single Imputation for Discrete Data," 2015.

[24] C. K. Enders, *Applied Missing Data Analysis*. 2010.

[25] 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.

[26] A. Gelman and D. Rubin, "Inference from iterative simulation using multiple sequences," *Statistical Science*, vol. 7, no. 4, pp. 457–511, 1992.

[27] J. Barnard and D. Rubin, "Small-sample degrees of freedom with multiple imputation," *Biometrika*, vol. 86, no. 4, pp. 948–955, 1999.

[28] 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.

[29] D. Schoenfeld, "Partial Residuals for the Proportional Hazards Regression-Model," *Biometrika*, vol. 69, no. 1, pp. 239–241, 1982.

[30] 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.

[31] D. F. McCaffrey, B. A. Griffin, D. Almirall, M. E. Slaughter, R. Ramchand, and L. F. Burgette, "A tutorial on propensity score estimation for multiple treatments using generalized boosted models," *Statistics in Medicine*, vol. 32, no. 19, pp. 3388–3414, 2013.

[32] J. P. Reiter, "Selecting the number of imputed datasets when using multiple imputation for missing data and disclosure limitation," *Statistics & Probability Letters*, vol. 78, no. 1, pp. 15–20, 2008.

[33] I. White, P. Royston, and A. M. Wood, "Multiple imputation using chained equations: Issues and guidance for practice," *Statistics in Medicine*, vol. 30, no. 4, pp. 377–399, 2011.

[34] 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.

[35] G. Cooper, *Elements of Human Cancer*. Boston: Jones and Bartlett Learning, 1992.

[36] D. Morris, J. Kearsley, and C. Williams, *Cancer: A Comprehensive Clinical Guide*. 1998.