RICE UNIVERSITY

Using Multiple Imputation, Survival Analysis, And Propensity Score Analysis In Cancer Data With A Large Amount Of Missing Data

by

Nathan Karmazin Berliner

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

Master of Arts Statistics

Approved, Thesis Committee:

Rudy Guerra, Committee Chair
Professor of Statistics

Kenneth Hess, Thesis Director
Professor, MD Anderson Cancer Center

Yu Shen
Professor, MD Anderson Cancer Center

Marina Vannucci
Professor 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 A Large Amount Of Missing Data

by

Nathan Karmazin Berliner

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

Contents

Illustrations

Tables

Chapter 1

Introduction and Background Information

1.1 Motivation

The motivation of this thesis is to show the methodology that can be used both by applied researchers and clinicians to draw meaningful survival and causal inference from data with a high amount of missingness. While all three fields are well studied, their interaction is not. I want the methods to be easy enough to describe to someone with a limited statistical background, but meaningful and valid so that the results obtained can be used in publication. The desire to have it this way stems from working on a related project with both statisticians and clinicians. While we will be motivated by cancer data, I believe that the methods used in this thesis are general enough to be applied to other types of 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 not much new research comes out of this field. 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 at least one at some point in their career. The goal of this thesis is to demonstrate how to use all three in trio, a topic that has only received little interest in the literature. I will explain each of these three disciplines in detail before we dive into combining them.

1.2 Imputation

In an ideal world, we would have complete data with no missingness, however this is rarely ever the case. Imputation (specifically multiple imputation) is a way to "fill in missing data" with plausible values, and it forms the base of this thesis. All of the other analyses that will be used will follow from it, thus we need a good understanding of it before we may proceed. Imputation itself has been around since the 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 doing this, because you are literally throwing away records and thus decreasing your sample size. In addition, this can be costly to the researcher. If it costs a set amount to collect a single record, and you don't use this record, you are wasting money. As well, in some rare cases, incomplete data might be the only type we can get (like if we have a machine that analyzes a blood sample chemical level, but can't detect it if the level is too high or low). Lastly, and most importantly, we will be biasing our estimates if we discard them. For example, if we have a random sample of people and are testing a drug, and want to run a regression on some collected covariates. Men are known to not want to give all of their information, so they leave them blank. In the analysis, we will need to discard the male samples because they are incomplete, leaving us only with women. Thus, we don't have a random sample anymore, and will get biased results because we have

knowingly thrown away half of our data which we know to be different [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 for different analyses can arise.

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

has its drawbacks, the idea of actually trying to fill in the data is an important one, and multiple imputation fills in the gaps that single imputation is not able to cover.

Multiple imputation (MI) began in the 1970's, but it wasn't until 1987 when the 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 substitute values (m values), we can think of the data now as being m datasets, each dataset having the observed data, and one value of the missing data. This might be hard to think about, but seeing a visualization of it will make it clear. 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 out analyses 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).

This method is obviously much better than the first two methods because it allows is to not throw away data, as well as allowing us to quantify our uncertainty about imputing the missing values. The only real drawback of multiple imputation is that we still don't have true data, but we can be confident enough in our estimations to compensate for that. The only better option would be to not have any missing data.

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

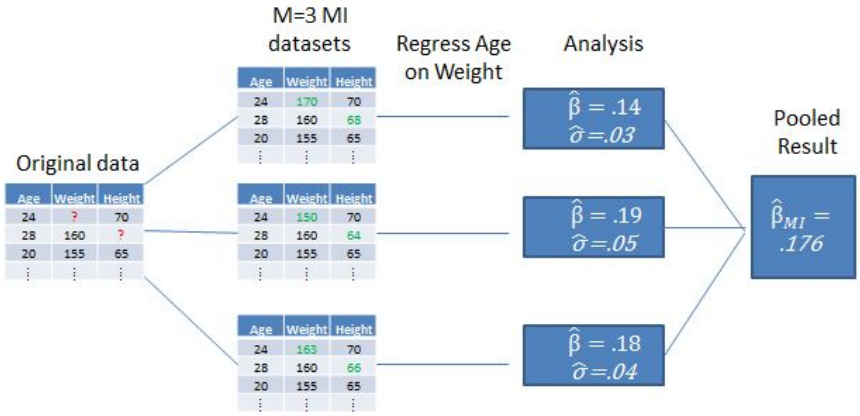

Figure 1.1 : Visualization of MI data

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 will be advised because of its popularity and strength.

1.3 Survival

Survival analysis is a huge field, and there have been many textbooks written about it. I only plan to introduce the topics that are relevant to my case study. For a much more detailed account of survival analysis, please see [3].

Survival analysis on the whole can generally be described as the analysis of time to event data, often in the presence of censoring 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". We use these terms because it is 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 literally portray it, but may be used in the context to denote the event of interest. Survival analysis doesn't have to be gloomy, but it makes more sense when it is.

The Kaplan-Meier (KM) estimator is a nonparametric estimate of the true survival function (the probability that you survive 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 or events that you observe 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).

The log rank test is a popular nonparametric test that researchers often use to see if two or more survival curves come from the same distribution. This is a useful tool to have, because visualizing curves alone does not give us this information. We

could have two curves that look radically different due to sampling error, yet still come from the same distribution. Knowing about if the survival curves come from the same or different distribution is useful because it allows us to make statements like "drug A is associated with longer survival time than drug B".

The general log rank test is given by

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

Where w_j is the weight of each individual (must be ≥ 0, we will set all to be 1), and $N_j = N_{1j} + N_{2j}$ is the number of deaths at time j, composed from the sum of the number of deaths at time j in each group, $O_j = O_{1j} + O_{2j}$ 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 $V_j = \frac{O_j(N_{1j}/N_j)(1-N_{1j}/N_j)(N_j 0_j)}{N_j - 1}$

Since we set all of the weights to be 1, this test as it is places equal weight to all of the deaths we observe. We could change these weights though to give more emphasis to certain death times. This is useful for example if we have a drug that 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 we will discuss next) fit the same data with no 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 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

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

The $h_0(t)$ is what's known as the baseline hazard, and can be any function that we would like (often times we choose a parametric distribution, like the Weibull). 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 β s 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 [5].

Our inference of interest is the hazard ratio, given by $\frac{h(t|Z)}{h(t|Z^*)} = \exp(\sum_{k=1}^{p} \beta_k (Z_k - Z_k^*))$ Where Z* is another set of covariates. The relative risk (or hazard ratio) describes how the hazard changes between individual with different covariates. 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 (it should not vary over time); hence the name proportional hazards, and does not depend on the baseline hazard. Using Cox regression, we can make statements such as "increasing the drug by one *mg* will decrease the rate of death (compared to non-users) by 30%". Cox modelling is one of the most used models in survival and medical literature.

1.4 Causal Analysis

Readers: please note that this is the weakest section. I plan to make it more thorough as my research progresses.

Although we can analyze any observational data using survival analysis, unless we conduct a randomized controlled trial, we cannot make any claims about causality. In order to prove causality, we need experimental data, not observational data. In an ideal world we would like to be able to do research and say that A causes B, rather than "our study says that A is associated to B". However, the only way that we can get this interpretation is if we conduct a randomized controlled trial.

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

However, we can frame our problem in a framework known as the Rubin Causal model, which helps us to get causal understanding from non-experimental settings [6]. In this model, the goal is to balance the groups as much as possible, so the only difference is due to the treatment in question. In an ideal world, we would like to

observe both potential outcomes (i.e. how the subject would respond to the treatment and to the control, denoted $Y_i(0), Y_i(1)$ for subject i). This is obviously not possible, since we can only observe one outcome. But we can get a good idea of the treatment effect by taking the difference of the average treatment effect between the two groups, known as the average treatment effect (ATE).

There are many different ways that we can go about implementing Rubin's causal model, but the most popular is what is called propensity score analysis. The propensity score is the probability that the subject received the treatment given the subjects covariates. It is computed using the patient's baseline (pretreatment) information [7]. Basically, 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. Propensity scores lead to group balancing, that is, if we control for propensity score, then our groups will have the same distribution at the baseline. And thus, we can treat the data like it was an RCT. The two most popular propensity score methods are propensity score matching (where we match those who picked the treatment to those who did not based off of propensity score) and propensity score weighting (where each 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 takes place might not be that accurate [8]. However, propensity score weighting is still a popular method.

Getting the propensity score is often done by logistic regression, although newer methods include regression trees and other binary classifiers. The propensity score for individual i is given by

$$\hat{e}_i(z) = p(T_i = 1|Z_i)$$

Where T_i is treatment (1 means the subject got treatment), and Z_i are their covariates. There has been a lot of debate as to which covariates to include in the propensity score (all variables, only relevant variables). I don't aim to settle this debate in this paper.

Propensity scores use is justified by the propensity score theorem, which states that if we assume conditional independence of the treatment given covariates on the outcomes, then we can also assume conditional independence of the treatment given the propensity score on the outcomes. Symbolically

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

Where the Y's denote the potential outcomes. If this is the case, then the potential outcomes are independent of the treatment given the covariates. We need this to hold, because it allows us to break the connection between the covariates and the treatment choice.

The proof of the propensity score theorem can be found in [9]. So, now we have the probability or propensity of a subject being assigned to a specific treatment given their covariates. Its usefulness is immediately seen in the matching case, because now instead of matching on n different levels, we need only to match on one number, the propensity score.

We are interested in weighting though, specifically, inverse probability of treatment weight (IPTW). These weights are given by

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

Where T is the treatment and Z is the covariate. This might be unstable as propensity scores approach 0 or 1, so alternate methods such as stabilized weights by Cole and Herman or trimmed weights have been presented that reduce this issue [10]. !! elaborate on this later!!

Once we have the IPTW calculated, we need to check how much balancing we actually achieved. We can do so by observing the change in standardized bias, or graphically on a histogram of propensity scores [11]. Once we do that though, we can add the IPTW weights to our model and get out causal results and inferences such as the ATE

Chapter 2

Methods

I want the framework and method we use to be easy to use and understand, so that it can easily 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. For the three parts that we are combining, there are many different theories and implementations to choose from. 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, we will need to make decisions as to what methods and analyses we will use. Whenever a decision is made, I explain why it was chosen, and 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 we should impute under. It should be noted that as long as we can produce valid imputations, the choice of method does not matter. However, since the base of our analysis starts with imputation, we need to make sure that we pick a good method. Everything that

follows in the analysis is dependent on our imputed data, so it is necessarily the case that 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.

In joint modelling (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 we don't know the true model parameters, we need to estimate them. 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 we are drawing imputations assuming MVN.

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. Some Research has been done for other types of models,

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]

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. Other distributions can be used, but the user will have to specify it, derive any relevant distributions, program it, and research the optimality properties.

An obvious issue arises when we have discrete or categorical data (alone or mixed in with continuous). There has been much debate in the literature about what to do

with in this case. 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]. However, it has been shown that FCS methods are superior to JM methods for mixed categorical data [16].A few of the most used R packages for joint modelling imputation include Amelia [12], norm [13] and cat [14].

It is my opinion that unless we are very confident in the multivariate joint distribution, that JM should not be used. In our cancer example, we have 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, we must give a k dimensional model, however in the FCS setting, we must give k one dimensional models. We are trying to sample from

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

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 X is the fully observed columns (which could possibly be empty). A pseudocode example can be seen here

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 [15]. 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 include MICE [17], mi [18], and BaBooN [19]. We will be using MICE in the applied section of this paper, and from now on, mice and FCS will be used interchangably.

We are 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, we will naturally have time variables that are strictly positive, and some binary indicators, whereas others can take any value. Trying to fit a parametric distribution with these stipulations will be very hard if not impossible, so we will be relegated to using a general distribution (like the normal), which will certainly elicit a poor fit. So, the full conditional specification certainly seems like the more appealing option. In an ideal world, we would have complete data, and would not need to resort to imputation. But since we don't have complete data, we must choose one method and accept its strengths and weaknesses.

In order to use FCS methods, we must have that the missingness in our data to be MCAR or MAR. It can work with MNAR data, but it requires some extra modelling assumptions. This is a seldom observed case in practice, so the interested reader may check [17] section 6.2 for a detailed look at this. Enders proposes using t tests 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 have [20]. 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 may follow a different distribution than the observed times. This is cleared up by Zhao et. al in 2014 through Kaplan-Meier MI [21]. 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 predictors will go into imputation, and what method to use (regression, predictive mean matching, logistic regression, trees, etc.). 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 the others 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 we will do until convergence. The older literature suggests only 5 is 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 we will talk about later), then add 5 iterations to the number at which we assume convergence is achieved. Due to the nature of FCS, 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's, 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, we actually

run mice. 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. We can think of the first few iterations as burn in, and the rest as samples. We take the value at the last iteration of each chain as the sample from the posterior.

We need to verify that our imputations are valid once we complete them. First, we need to see if our chains have converged. Since mice is an 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 [22]. 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 we have assessed convergence, we need to actually 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 we have m imputed datasets, 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, as well as the proper variance for the quantity we have in mind from the m imputed datasets.

Rubin's rules involve three parts. The first is getting an estimate of the population estimand Q. To get this, we define $\hat{Q}_i$ as the estimand evaluated from the data in the i^{th} dataset. Then, we 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 or each estimate $\hat{Q}_i$. If we 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

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

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 that we are trying to pool 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 $\nu_{old} = \frac{m-1}{(\frac{B+B/m}{T})^2}$ [23].

Rubin's rules assume normality, so if our statistic in mind is not asymptotically normal, we need to transform it towards normality before we pool. There have been

some research about how to pool non normal quantities, but current research shows poor results and power when doing so [24]. 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

2.2 Survival analysis

Now that we have the multiple imputation datasets created, we may run our analysis on each of them. As a general rule of thumb, we should run our desired analyses on the original data to get the available case estimates, to get an idea of what to expect and to check if the model assumptions are met. Because we are working with cancer data, we are interested in some basic survival quantities, such as Kaplan-Meier survival estimates, log rank test to test for similarity of curves, and Cox regression. 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.

At the beginning of the last section, we learned about Rubin's rules, but before we begin, it should be noted that there is another way to work with the multiply imputed data. This method is colloquially called the stack method. For the stack 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, we can produce unbiased estimates of quantities of interest, but the estimates of variance will

be too small (since we are artificially increasing the sample size) [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.

2.2.1 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 noninformative 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 we run the Kaplan-Meier curve 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 [24]. 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. 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.

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 we just take 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 pool the Greenwood variance associated with each time point and then take the average as the variance at that time point. But the Greenwood estimator tends to underestimate the true variance [3], and taking the mean of the variance doesn't make much sense. 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 points where the lower and upper bands cross the .5 survival line respectively. 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, they are 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. In addition, many statistical packages only report the chi square version of the statistic, meaning we would have to reprogram it in order to use Rubin's rules for pooling. 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 times [3]. 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, we know that 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 checking 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 [25]. 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.

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. As well, with causal analysis, we get a cause and effect result, which is in tune with what the general population believes that results should be. Propensity score methods are an easy to understand yet powerful tool. The use of propensity scores justified in Rosenbaum and Rubin's 1983 paper [7]. Our overall goal is to estimate the average treatment effect in a setting where the initial study was not an RCT. Propensity score analysis helps us to do this by balancing the groups out so that it becomes more like an RCT.

We will need to make a few decisions along the way. Our very first decision comes when deciding how to use the propensity score. There are four main uses in the survival literature: Matching, stratification, weighting, and covariate adjustment [10]. The goal is to balance the treatment and control groups, such that the only difference between the groups is due to the treatment received, and not any underlying factor. All of these methods will help us to examine the average causal effect, but each goes about it in a different way. Covariate adjustment and matching have fallen out of favor in the literature recently, but weighting and stratification remain very popular. We will be using weighting in our real data analysis, but I want to first spend some time discussing matching because of its popularity in practice.

If we choose to match propensity scores in the multiple imputation setting, we will need to decide how. The stacking method described before would obviously be inappropriate, as we would have spurious and repetitive matches due the falsely inflated sample size. Matching on the stacked set would give us much more power to detect a difference, but the results from it would not be valid.

There is hope though for matching with multiple imputation data. Two methods are described in Mitra and Reiter about how to do this [26]. In the first method, propensity score matching is done within each MI dataset (known as within match-

ing). Thus, we will get m estimates of the average treatment effect to which we will average. The other method, known as the across method takes the average propensity score for each individual among the m imputed data sets, matches on the averaged propensity scores, and estimates the treatment effect in that manner. Both methods have their pros and cons, and are appropriate for different scenarios. However, 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. Propensity scoring using logistic regression is the most widely used method, because it is computationally simple and is well understood by both clinicians and statisticians. Thus, we will use it to compute our propensity scores.

Different propensity scores will be obtained according to what predictors we use in our model, so we need to be sure that we fit a model with clinically relevant and meaningful predictors. Our overarching goal is to account for any pretreatment covariate that leads to the selection of the treatment. There has been significant debate among statisticians about how to set up these models, by either throwing in every possible variable into it, or only include ones known to affect treatment selection. I don't plan to settle this debate, but since we are in a setting where simplicity is a goal, I plan to use only pretreatment covariates deemed to be important (through consultation with subject matter experts) to treatment selection. This is the method that Peter Austin ,a well-known propensity score researcher recommends [10]. There exists research on how to test if we have properly set up the propensity score model, however, It is not clear how this could be applied in the multiple imputation

setting [27]. We should however, check that we have balance between the groups, by examining the standard bias between the groups and looking at balancing tables.

Once an acceptable propensity score model is selected, we will need to pick what type of weighting we will do. Many exist, but the most popular are X,Y,Z . We will then run our Cox model again (as in the survival section), but we will factor in the weights. Then, the results that we receive can be viewed as if they were from a RCT. As well, we can go ahead and analyze them as such, and draw causal inference. I need a lot more work here.

Chapter 3

Application

3.1 Data Explanation

Unfortunately, as of October 23rd, we are still waiting for final approval from the PI of the project and the IRB. I want to get the proposal out to y'all though. So, what I will do in this section is describe what I will do, and once I get approval, I will fill in the tables and plots. In the meantime, I will put in ¡whatever¿ where something that uses the data should be. The IRB protocol is RCR03-0931, but neither Dr. Hess nor myself are on the protocol, because it is out of date. We are in the process of getting it updated, and plan to have it done by the time I defend this thesis

Now that we have the theory in place, we can apply it to some real data. The dataset that I chose to analyze is a dataset from MD Anderson Cancer Center, with permission from Dr. Bugano, Dr. Ibrahim, Dr. Hess and ¡whoever else needs to be thanked¿. This dataset has historical records of 1514 MD Anderson patients who have had breast cancer that has metastasized to the brain. There are lots of covariates recorded (about 90, most with some missingness), a few different treatments, as well as survival endpoints (which are all observed). The data can be broken down into a few parts; subject data (like age range and race), cancer type data(like stage, type, status), pre metastases data (like treatment types), post metastases data (like treatment, type of brain mets), and survival (survival of different events and censoring indicators). Below we can see a ¡plot and table of the missingness¿. This data is

exemplary for this task because it is large, survival amenable, has missingness and is a prime candidate for imputation, and has treatment variables that are not given in an RCT.

Our first step is to clearly define what we would like to find. There are many interesting questions we could ask and answer from this data because of the amount of data available, but the question I will focus on here is survival in two different settings. 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 what these drugs do, 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 work. For a much more detailed analysis and other clinically relevant questions, see !!Hess, Bugano, Berliner!! This is the project that this research was forked off of, although it will probably not be published by the time this thesis is.

3.2 Imputation

We first need to impute the missing data. This is a challenging task, because of the attention and care that needs to be given to imputing about 90 covariates with missing data. But we need these covariates to be imputed properly 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, we can always use variable selection if we have too much

data.

Our data is quite high dimensional, and there are a many binary variables 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 [17] because it is easy to use yet powerful.

The model is set up by hand, following the advice from [17]. It took a about three weeks to set up and check. This was because the number of covariates was huge, 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.

The first task we need to do is to assess the missing data mechanism. As we have discussed before, there is no 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

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 categorical, so the most popular model chosen were logistic regression and predictive mean matching. 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]. An ideal variable to use as a predictor would be one with an influx and outflux near one. I used all predictors except those that had very poor influx and outflux. ¡influx and outflux plot¿

Once the choices had been made, I needed to run the imputations by trial and error. This took a considerable amount of time, because after every change made, I had to rerun the imputation and reassess convergence and validity of the imputation. I did trial and error for model and predictor specification until the data seemed like it could have reasonably been observed.

For the final imputation, I decided to impute $m = 50$ datasets and 40 iterations for each. Mice generally converges quickly (within 5 or 10 iterations), but by setting the number of iterations so high, it is as if we are setting a burn in period, and then taking our sample. The number of datasets was selected to be 50 because in general, according to Reiter the number of datasets should be more than the imputations [28], and while there is no consensus about how many imputations to do, the modern research argues that the more the better. 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) [29]. We have about 30% missing data, but coupled with Reiter's advice, I chose 50.

After the final 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 Z ram and Q processors ¡Get this info from MDACC comps¿. 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. I ran it 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 imputations were very similar. As well, there is hardly any confidence gained going from 50 to 100, and having such large objects in memory can be harder to work with.

We need to check our final imputations for convergence and reliability. Convergence is assessed by looking at ¡plots of covariate mean and sd by iteration¿. According to van Buuren "the different streams should be freely intermingled with each other, without showing any definite trends. Convergence is diagnosed when the variance between different sequences is no larger than the variance with each individual sequence" [17]. Looking at these plots, this certainly seems to be the case. Other authors suggest using a more formal statistical tests such as $\hat{R}$¡ rhat¿ to assess convergence by looking at the within versus the between variance, so I also display that (values near 1.00 are ok and values greater than 1.1 indicate we should run longer). Diagnostic plots are viewed to ensure that the imputed data is similar enough to the real data.¡ A few of the plots have been replicated here¿. To see all of the plots, go to the shiny app/ R package (do this if enough time, also see about security. (Might just need to make it be available upon request). As we can see, not all of the imputed data follows the distribution of the observed data exactly, but for the majority of the plots, the data look like they could have been real data.

3.3 Survival Analysis

Now that the datasets are imputed, we are ready to run our models on them. 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 were run and the 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 metastases treatment data was necessarily determined after the diagnosis of the brain met, it is not appropriate to use this data as baseline covariates in the analyses. Only covariates known at the time of diagnosis can be used in this fashion we can do a landmark analysis by estimating when the vast majority of patients would had their brain met treatment choices started and starting our analyses at this point ". After speaking with subject matter experts (Dr. Bugano and Dr. Ibrahim), this landmark time was determined to be 2 months.

The first result that we will check is the Kaplan-Meier curves for the imputed data. The available case analysis shows that lapatinib and trastuzumab are quite close to each other, while having no HER2 directed treatment being much lower. For the chemotherapeutic drugs, available case was X,Y,Z. The log rank test statistic is X, with pvalue Y, indicating Z ,and these results can be seen on ¡table whatever¿. The pooled KM estimate was found using Rubin's rules, but under a complimentary log-log transform as suggested by [24] to get the survival curves towards normality. The results from MI look quite similar ¡and can be seen on the plot ofAC analysis

and MI analysis¿.

We can also get an approximation for the log rank test on the MI data via the Wald test on the pooled Cox model fit only on the treatment ¡with results in table here¿. 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 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 [24]. 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).

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 these X variables were clinically relevant for the baseline model. The available case model ¡will be seen in the table below¿. We need to make sure that the proportional hazards assumption is met in the available case model so we can apply it to the MI data. And seems to mostly pass the straight edge test for proportional hazards. To check, we visually inspect the Schoenfeld residuals over time, and check the test of correlation between the residuals and time. Overall, the assumption of proportional hazards over time seems reasonable, and the test statistic affirms this ¡AC Cox.zph plots¿. Although the splines fit the points may not look straight, it certainly seems reasonable that a straight line could be fit between the 95

We have verified that the available case analysis seems reasonable, now we can work with the MI data. We fit that same Cox model on all of our imputed data sets, and pool our results via Rubin's rules. We need to verify that we still have proportional hazards though. This is not a straight forward task anymore, since we don't actually have a single model, rather, we have the average of multiple models. We are no longer estimating the parameters by maximizing the partial likelihood; rather we are estimating them based on the average of the coefficients from the MI datasets. There are two ways we can go about verifying the proportional hazards assumption.

The first is to check the proportional hazards assumptions on the stacked dataset. To do so, we plot the Schoenfeld residuals over time, and observe the spline fit to it (which should be independent of time). This is a good visual tool, but when running the chi square test to check for the correlation between the coefficient and time, the sample will be artificially too big, and thus we cannot trust the results.

The correct way to do this is to observe each plot and statistic generated from the 50 datasets to see if the assumptions hold. This may seem like an arduous task when the number of imputed datasets is large, but we can circumvent it by writing ¡a shiny app to view them¿, or ¡plot all of the splines on one plot¿. We can also look at all of the chi square tests for the 50 datasets and 13 variables for each, although there is bound to be some overlap between significance and non-significance due to the multiple testing problem. Overall, our imputed plots are very similar to the plots produced by complete case analysis, to which we have deemed to be acceptable for the proportional hazards assumption. We may now look at the Cox regression coefficients and exponentiate them in order to obtain the hazard ratios, and obtain corresponding 95% confidence intervals. Looking at ¡!! table whatever!!¿ , we can see

that some factors (such as X, Y, Z) force a larger hazard ratio than others.

We may then add in our treatment variable to see how it affects the hazard, and see how it changes other factors. After doing so, we can see that X does Y. The results can be seen here in a table ¡of AC vs MI estimates¿.

3.4 Causal Analysis

This is the section that needs much more work. I'll only lay the outline here.

Lastly, we will want to draw causal inference, and see what the average treatment effect of each drug is. This is necessary because the data was collected from a database, and we did not have an RCT. As well, this piece of information is what clinicians and laypeople really wantit answers the question of which drug is better. There are many interesting questions that we may ask with this dataset, but here we will only focus on lapatinib /trastuzumab /no treatment and capecitabine/other/none.

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. Let's do this first on the available cases. To do this though, we need to get an actual propensity score. In talking with the cancer professionals on the project, it was determined that covariates X,Y,Z, were important pretreatment variables that needed to be controlled for. In the available case analysis, we see that the standardized bias is ¡this¿. Now, we fit a logistic regression / multinomial logistic model on the 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 [30]. But at the time of publication, there is not much literature or software on this subject, so I felt that it was not appropriate to use its results.

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

Next we can critique the survival section. We 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, get in to remission, and then relapse. We model the states as a stochastic process. This would be really interesting, and I would have liked to implement it because I think it would have been interesting from a multiple imputation perspective, but unfortunately our data was not conducive to that.

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.

¡This needs more work¿Lastly, we move on to the causal analysis part. While there are many other binary classifiers that could be used to make propensity scores, we chose to use logistic regression. We chose to weight, using IPTW. 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 choice of using logistic regression IPTW was based solely on tradition and ease of understanding from non-statisticians. I included a section about how we could use propensity scores to match, and while this is used by a lot of researchers, new statistical studies argue about its validity. 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.

- MCAR: Missing completely at random: $P(R = 0|Y_{obs}, Y_{mis}, \psi) = P(R = 0|\psi)$. The missingness in the data is not at all related to any of the data that we do or dont have. If a lab technician 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. 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 a full moon causes the blood testing machine to break more often, but we dont have the moon phase as a variable.

A.2 Cancer and Treatment Overview

Cancer is a disease in humans where cells in the body begin to grow in an uncontrolled manner [31].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 [31]. 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 [32]. 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 [31] and [32].

Bibliography

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

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

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

[7] 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.

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

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

[10] 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.

[11] V. S. Harder, E. A. Stuart, and James C. Anthony, "Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research," *Psychological Methods*, vol. 15, no. 3, pp. 234–249, 2011.

[12] 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.

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

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

[15] 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.

[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] 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.

[18] 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.

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

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

[21] 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.

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

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

[24] 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.

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

[26] 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.

[27] L. Li and T. Greene, "A Weighting Analogue to Pair Matching in Propensity Score Analysis," *The International Journal of Biostatistics*, vol. 9, no. 2, pp. 215–234, 2013.

[28] 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.

[29] 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.

[30] 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.

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

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