RICE UNIVERSITY

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

by

Nathan Karmazin Berliner

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

Master of Arts Statistics

Approved, Thesis Committee:

Rudy Guerra, Committee Chair
Professor of Statistics

Kenneth Hess, Thesis Director
Professor, MD Anderson Cancer Center

Yu Shen
Professor, MD Anderson Cancer Center

Marina Vannucci
Professor and Chair of Statistics

Houston, Texas

April, 2015

ABSTRACT

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

by

Nathan Karmazin Berliner

In this thesis, we will unify the theory of multiple imputation, survival analysis, and propensity score analysis. While each of these fields have been studied individually, there has been little work on using the three in trio. Starting out with an incomplete dataset, we aim to impute reasonable imputations, run survival analysis on each of the imputed datasets to get survival estimates, and then do propensity score analysis to observe causal effects. Along the way, methods are proposed to check the validity of assumptions that are made, and different types of analyses are used in the multiple imputation setting which have not been previously studied in the literature. I apply the methodology to cancer survival in a case study, but the methods used are general, and could be used for any type of data.

Contents

Illustrations

Tables

Chapter 1

Introduction

1.1 Motivation

The motivation of this thesis is to develop methodology that can be used both by applied researchers and clinicians to draw meaningful survival inference from data with a high amount of missingness. 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.

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

1.2 Imputation

Imputation (specifically multiple imputation) is a way to fill in missing data with plausible values, and it forms the base of this paper. All of the other analyses that will be used will follow from it, thus we need a good understanding of it before we may proceed. Imputation itself has been around for some time, but multiple imputation is a recent development, proposed formally in 1987 by Donald Rubin [1]. At first, statisticians payed no attention to missing data, and happily discarded records for their analysis that were incomplete. This was known as complete case analysis. There are many problems with this paradigm. To begin with, you will lose a lot of statistical power when doing this, because you are literally throwing away records and thus decreasing your sample size. In addition, this can be costly to the researcher. If it costs a set amount to collect a single record, and you dont use this record, you are literally wasting money. As well, in some rare cases, incomplete data might be the only type we can get. Lastly, and most importantly, we will be biasing our estimates if we discard them. For example, if we have a random sample of people and are testing a drug, and want to run a regression on some collected covariates. Men are known to not want to give all of their information, so they leave them blank. In the analysis, we will need to discard the male samples because they are incomplete, leaving us only with women. Thus, we dont have a random sample anymore, and will get biased results. The next wave of statisticians wanted to improve upon this, so they developed what we now call today (single) imputation. In single method (such as regression, trees) is used one time to impute the missing value. While this is a little better than complete case analysis, it still has many drawbacks. Asserting that a single value is the true value is just foolish. There is always some amount of error 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 desirable. 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. If this was valid, we wouldnt ever need to collect samples, and we could just be armchair statisticians. 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 theoretical gaps. In 1987, Donald Rubin proposed multiple imputation as a solution [1]. The central idea is to make many datasets for the missing data, drawing the estimates for the missing data from the parameters posterior distribution. Once we have a sufficient number of datasets, we can run whatever analyses we would like on them, and then pool the results. We can get the standard errors by noting the within and between imputation variance. This is obviously much better than the first two methods because it allows is to not throw away data, as well as allowing us to quantify our uncertainty about imputing the missing values. The only real drawback of multiple imputation is that we still dont have true data, but we can be confident enough in our estimations to compensate for that. Multiple imputation use has been steadily increasing over the past 30 years, and it is now the standard for missing data.

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 introduction, please see X Y Z Survival analysis on the whole is analysis of time to event data, often in the presence of censoring. The two main tools that we will be using are Kaplan-Meier estimate and Cox regression. The Kaplan-Meier

estimate is a non-parametric estimate of the true survival function (the probability that you survive after a time t). Cox regression is a modelling tool that allows us to analyze the hazard ratio of a covariate. Using cox regression, we can make statements such as increasing the drug by one mg will increase its hazard by 30

1.4 Propensity Score Analysis

¡¡THIS NEEDS A LOT MORE WORK¿¿ In an ideal world we would like to be able to do research and say that A causes B, not something along the lines of our study says that A is related to B. Using propensity score matching, we are able to get this interpretation, via Rubins causal model. Rubins causal model is a framework that allows us make causality statements, and propensity score is a tool that grants us access to the framework. The way propensity score analysis works is that we will match patients on their propensity to be in the treatment group. This is often done by logistic regression, although newer methods include regression trees and such.

Chapter 2

Methods

2.1 Multiple Imputation

I want the framework we use to be easy to use and understand, so that it can easily be discussed among clinicians and other non statistically minded people. 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, I try to pick the optimal theory and software packages to achieve this. Our first decision comes as to what paradigm we should impute under. It should be noted that as long as we can produce valid imputations, the choice of method does not matter. However, since the base of our analysis starts with imputation, we need to make sure that we pick a good method. Everything that follows in the analysis is dependent on our imputed data, so it is necessarily the case that bad imputation will lead to poor results (be it bias, high variability, loss in statistical power). There are two main divisions in multiple imputation, and they are joint modelling and full conditional specification. In joint modelling, we assume that the data can be described by a multivariate distribution (specified by the user), and draw imputations from a joint distribution of the unknowns for the rows, given what we do know and their associated hyperparameters. There has been extensive research on using the normal model for this, and research shows that it even performs well under non-normality. R packages for joint modelling imputation include Amelia (source), NORM, and cat (Schafer source) On the other hand, there is

fully conditional specification. In this paradigm, missing data is imputed on a variable by variable case on the columns, based off of a specification of the imputation model for each imputed variable. These full conditionals should factor to specify the joint distribution. We are going to have to specify something, there is no escaping that, but I think that it is easier for the average person (especially a clinician) to be able to define a single distribution rather than to guess at a multivariate. In addition, in the survival analysis setting, we will naturally have time variables be only positive, whereas others can take any value. Trying to fit a parametric distribution with these stipulations will be very hard, so we will be relegated to using a general distribution (like the normal), which will certainly elicit a poor fit. So, the fully conditional specification will be our choice. Now that we have chosen the paradigm, we need to select an implementation of it. Many exist (such as MICE, mi, etc.). I wanted to select the implementation that combined ease of use, understanding, and programming. What I decided upon was a method called MICE- Multiple imputation via chained equations. (SOURCE). MICE is a MCMC method that under compatibility, is a Gibbs sampler, where we obtain samples from the joint by sampling from the full conditionals. Compatibility means that knowing the full conditionals allows you to know the joint. Specifically ¡¡See 4.5.4 in van Burren¿¿ [2]The user defines the full conditionals, so it is possible that the joint may only exist implicitly, and not actually have a functional form. This is an obvious disadvantage of this method, but multiple studies have shown that MICE is robust to this flaw, even when using highly incompatible models¡¡sources for this¿¿. Mice has been shown in simulation studies to yield unbiased estimates with proper coverage. In order to use mice, we must have that the missingness in our data is missing at random (MAR). What this means is that the probability of a covariate being missing is related to another covariate that

we have collected (for example, obese teens may not want to volunteer their weight, but knowing their sex and age will alert us to this). The opposite of this is MNAR-missing not at random. This is when the probability of a covariate being missing is dependent on another covariate that we dont have, for example, if obese male teens dont want to volunteer their weight, but we dont collect age or gender. MICE will work under both MAR and MNAR, but performs much better under MAR. There are tests for testing ??, but they arent very widely used) Once we have the correct assumptions, we need to set up our full conditionals. This may take a while for large datasets, but the extra time spent will ensure a better model. We choose what predictors will go into imputation, and what method to use (regression, predictive mean matching, logistic regression, etc.). For variables that are derived from others, we impute the others and then compute that variable, in a process known as passive imputation. Since mice is an iterative process, we must choose how many iterations we will do until convergence. The literature suggests only 5 is enough (source), but with modern computation, we can easily exceed this, even with large data. As well, we need to decide how many datasets to impute. Once again, 5 will due, but more is better, since it will cut down on simulation error (find where I wrote up the reasons why). We need to verify that our imputations are valid once we complete them. The overarching idea that we need to pay attention to is does the data look like it could have been real data. We can assess this in many ways, including density plots, box and whisker plots, etc. Once we have m imputed datasets, we may run any valid analysis (regression, computing any statistic) on each imputed dataset INDIVIDUALLY. We may then use Rubins rules (source) [put the steps in later], to pool our estimates. This will give us a point estimate, as well as proper variance for the quantity we have in mind. Rubins rules assumes normality, so if our statistic in mind is not

asymptotically normal, we need to transform it towards normality before we pool.

2.2 Survival analysis

Now that we have the multiple imputation datasets created, we may run our analysis. The analyses, as stated in the previous section are heavily based on Rubins pooling rules. We are interested in some basic survival quantities (Kaplan-Meier survival estimates for certain groups, cox regression to determine hazard ratio), as well as some more advanced ones (survival in the competing risks setting). Following Rubins rules, we run the individual analyses on each of the m datasets, and then pool our results. This works fine for the cox model, as the regression coefficients are asymptotically normally distributed. In many of the MI packages, there are functions to model and pool with cox regression. Care needs to be taken to ensure that our analysis is valid (i.e. that we do in fact have proportional hazards). We can go about this in two ways. The first is to check the assumptions on each individual model fit to each dataset. This may prove to be an arduous task, but with graphical tools such as shiny, this isnt too bad. We may also stack our multiple imputations on top of each other, and run one huge model (thus our MI data are acting as replicates). This will produce unbiased estimates, but the standard errors will be too low. So, this tool may be used as a graphical check, but the results from the inference of this cannot be trusted(source for all of the above). Once we have verified that the model follows the assumptions, we may trust its results, and perform inference on the parameters using the obtained total variance. We now would like to calculate the Kaplan-Meier curve for a subset of the population. While this is a pretty simple task in the complete case setting, it is a little more difficult in the multiple imputation setting. One of the main tasks that clinicians are interested in is the median survival time (the

smallest time that the survival function is less than .5), specifically, the variance at the median. The median is a much better estimate of survival than the mean is, because in survival analysis, the time to event is typically right skewed, so the mean survival time is almost certainly not relevant to the typical patient/user of the km plot. We could get an unbiased estimate of the median via stacking the imputations and then running a km model, but this will give us false confidence, because our sample size is greatly inflated. The correct way to go about this is to run a Kaplan Meier curve on each of the imputed datasets, and then pool the estimates. We will run into two problems here though. The first is that the Kaplan Meier curve is not normally distributed, so we will need to pretransform the data before pooling towards normality. [source] claims the complimentary log log transformation helps with this, and this transformation was implemented in [prostate] paper successfully. The second issue is that the variance at a point is usually computed under greenwoods formula, which depends on the order of the data. We can stack, but this will lead to too small of an estimate. As well, we could pool on greenwoods formula, but this would lead to too big of changes between the data sets (by construction of the greenwood estimator). The solution for this issue is to derive the variance of the median by the reflection method, is look at all of our Kaplan Meier curves, and take the median to be the average median, and take the confidence interval to be the time the first and last curve cross the 50 line * *work on this*. The last thing that we might be interested in is the cause specific hazard, and the cumulative incidence function. Work on this a lot.

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 and propensity scores is much stronger than conventional analysis. As well, with causal analysis, we get a cause and effect result, which is in tune with what the general population believes that results should be. Propensity score methods are an easy to understand yet powerful tool. Our overall goal is to estimate the average treatment effect in a setting where the initial study was not a completely randomized experiment. We will need to make a few decisions along the way. Our very first decision comes when deciding how to use the propensity score. We can either choose to match or stratify on the propensity score. The use of doing either is justified in Rubins paper !!Rubin 1983!!. Both will help us, but matching is easier for the layperson to understand, and easier to implement, so we shall use that. Our next decision comes as to how to use matching in the multiple imputation setting. The stacking method described before would obviously be inappropriate, as complete records would always be matched to themselves, and imputed values would often be matched with their selves from another imputation. As well, we would have a falsely inflated data, giving us confidence in our matches where we should not have any. Two methods are described in Robin and Mitra !!sauce!! about how to do this. In the first method, propensity score matching is done within each MI dataset (known as within matching). This, we will get m estimates of the treatment effect to which we will average. The other method, known as the across method takes the average propensity score for each individual and estimates the treatment effect in that manner. Both methods have their pros and cons, but Mitra and Reiter show that the across method

limits bias more than the within one does. I need to determine what to use, because I might want to use inverse PS weighting in the cox model.

Chapter 3

Application

3.1 Data Explanation

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 (get the permission!!). This dataset has historical records of X MD Anderson patients who have had breast cancer that has metastasized to the brain, and it records many covariate, treatments, as well as survival endpoints. This data is exemplary for this task because it is large, survival amenable, has missingness that is easy to impute on, and has treatment variables. Our first step is to define what we would like to find. There are many interesting questions we could ask from this data because of the amount of data available, but the question I will focus on here is the effect on survival and treatment of two HER2 therapeutic drugs-Lapatinib and Trastuzumab. For a much more detailed analysis and other clinically relevant questions, see !!Hess, Bugano, Berliner!!. So, we will want to check out survival curves and cox regression, as well as analyze the treatment effect. We first need to impute the missing data. This is a little challenging just because of the sheer number of covariates that we have. But we need these covariates. With more covariates, the more sure we can be in the assumption of MAR missingness. As well, it is better to have too many covariates than not enough. The model is set up, and the appropriate methods are selected for each datatype. The mice algorithm from the R package mice

is run. For 50 datasets, 40 iterations, the algorithm runs in about X hours. While this seems like a long time, this only needs to be done once. Convergence is assessed, and 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 (do this if enough time). Not all of the imputed data follows the distribution of the observed data exactly, but we obviously dont expect this to happen. Now that the datasets are imputed, we are ready to run our models on them. As a sanity check, we may compare them to available case analysis. Since the imputed values we generate ought to be quite similar to what data we have, we should expect our estimates to be similar. The first result that we will check is the Kaplan Meier curves for the imputed data. The available case analysis seems to show that lapatinib and trastuzumab are quite close to each other, with no treatment being much lower. The results from MI look quite similar. [put the stuff in]. The pooled KM estimate was found using Rubins rules, but under a cloglog transform as suggested by [3] to get towards normality. We can also run a log-rank test on the MI data. This was implemented by [4] using another form of imputation called kmmi, but it has not ever been used on regular MI (kmmi works on missing censoring times). Log rank test is a normally distributed quantity asymptotically, so we can just pool it as normal and use the degrees of freedom from Rubin and Barnard to get our inference. !! put the analysis here!! !!!Do I want to do competing risks analysis?!!! Now that we have estimate of survival, we may set up a model to observe how changes in some baseline covariates change the hazard. To do this, we need to run a Cox proportional hazards model. The original available case model is as follows. We need to make sure that the proportional hazards assumption is met, so we may check the cox zph command to look at the schoenfeld residuals over time, and check the test stat. Overall, it looks

to be proportional hazards over time, and the test statistic affirms this. Then, we fit that same cox model on all of our imputed data sets, and pool our results via Rubins rules (no transformation needs to be done since the cox model coefficients assume asymptotical normality). We need to verify that we still have proportional hazards though. This is not an easy task, since we dont actually have a model, rather, we have the average of multiple models. We are no longer estimating the parameters by maximizing the partial likelihood, rather we are estimating them based on the average of the coefficients from the MI datasets. There are two ways we can go about this. The first is to check the proportional hazards assumptions on the stacked dataset. This will give us a good visualization about the shape of the proportional hazards over time, but when running the chi square test to check for the correlation between the coefficient and time, the sample will be artificially too big, and thus we cannot trust the results. The correct way to do this is to observe each plot and statistic generated from the m datasets to see if the assumptions hold. This may seem like an arduous task when the number of imputed datasets is large, but we can circumvent it by writing a shiny app to view them, or plot all of the loess curves on one plot. We can also get the average of the chi square test results if we need a little more information than looking at the plots. Overall though, our imputed plots are very similar to the plots produced by complete case analysis, to which we have deemed to be acceptable for the proportional hazards assumption. We may now look at the cox regression coefficients and exponentiate them in order to obtain the hazard ratios. Looking at !! table whatever!! , we can see that some factors force a larger hazard ratio than others. We can take the reciprocal of it to look at the protective effects of each covariate. Lastly, we will want to draw causal inference, and see what the average treatment effect of each drug is. This is necessary because the data was collected

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

Chapter 4

Discussion

We have discussed a number or tools and methods to analyze survival data with missingness. There are lots of decisions to be made along the way, and I am in no way advocating that my exact choices are the right ones, I am only claiming that the decisions made were proper for the type of data that we had. There will certainly be many disagreements about the multiple imputation portion. And since the multiple imputation serves as the root of the analysis, the concerns should be addressed. The first concern comes from people who dont understand or believe in imputation of missing values. Multiple imputation is a tool to help us find plausible values for missing data. We will make no claim that the imputed values are right, but when used correctly, the results will be unbiased. We arent using multiple imputation to create data where there is none, we are using it to fill gaps. In fact, there exist situations where not imputing could lead to biased results due to sampling bias (for example, if teenage males who are obese dont want to self-report their weight, then classic complete or available case analysis will yield biased results because we have knowingly left out part of the population). We need to impute to make sure we have included all of the information and not to bias our estimate. The next and more substantial critique will come from statisticians who may not believe that the distribution that the imputations are being drawn from are 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, 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. An interesting extension to this project would be to use a non parametric approach to multiple imputation, such as the one suggested by Long et all in [5]. 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. 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. Next we can critique the survival section. We decided to use standard Kaplan-Meier and cox analyses because they are very standard in practice, and answer the questions well. However, some lesser known methods could have been used. A popular theoretical model is called the accelerated failure time model, which describes how covariates effect the hazard, assuming that it acts in a multiplicative fashion. This is useful for clinicians, but not really good for patients, because the conclusions drawn from it are drug x will make you live 50Next, since there is no well established method to validate the model, we had to be creative and define our own methods to check them. The methods are reasonable and both the stacked and individual methods are very similar. More research should be done though to verify if this will always be the case, specifically in the presence of pathological data. There are two concepts that are interesting, but our data did not allow for it. The first is variable selection. The clinicians knew what they wanted to test, so this was not needed, but variable selection in the context of MI is an interesting question, and van Buuren covers it in his book [2]. This would be very useful if our dataset had covariates that we were unsure of or wanted to examine.

Another interesting addition would be using multistate data. In this setting, subjects can transfer from one group to another, ie have cancer, get in to remission, and then relapse. We model the states as a stochastic process. This would be really interesting, and I would have liked to implement it because I think it would have been interesting from a multiple imputation perspective, but unfortunately our data was not conducive to that. Lastly, we move on to the causal analysis part. While there are many other binary classifiers that could be used to make propensity scores, we chose to use logistic regression. This choice was based solely on tradition and ease of understanding from non statisticians. As well, there has been some new research recently saying that propensity score matching is not as powerful as it was once thought (? King?). Lastly, our choice of how to combine propensity scores was solely based off of the Mitra paper, and no more studies have been done to show that this is in fact the optimal way to do it.

Chapter 5

Conclusion

This paper details how to use multiply imputed data to answer survival analysis and causal analysis questions. The motivating example was cancer data, and the methods are tailored towards that. Along the way, we discover new visualization and pooling tools to aid in analysis of multiply imputed data. We test the methods out on a large cancer dataset, trying to draw meaningful inference from a dataset with substantial missingness.

Appendix A

Appendix

adfsdfsdf

Bibliography

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

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

[3] 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.

[4] 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.

[5] 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.